Abstract Title: A visual analytics antibiogram dashboard as part of a comprehensive approach to perioperative antibiotic administration

Authors: Luis M. Ahumada, MSCS,[a] Allan F. Simpao, MD,[b] Jorge A. Galvez, MD,[b] Mohamed A. Rehman, MD,[b] Jeffrey Gerber, MD, PhD,[c] John Martin, MBA,[a] Beatriz Larru, MD,[c] Kaede Ota, MD, MSc,[c] Telene Metjian, PharmD,[d] Bimal Desai, MD, MBI[c]

[a] The Children's Hospital of Philadelphia (CHOP) Data Analytics and Enterprise Reporting Group, Philadelphia, PA; [b] CHOP Department of Anesthesiology and Critical Care; [c] CHOP Department of Pediatrics; [d] CHOP Department of Pharmacy Services

Introduction: Many hospitals routinely perform antimicrobial susceptibility testing for bacterial pathogens. The results are often organized into a summary table, or antibiogram, that clinicians use as a reference guide for antimicrobial resistance patterns. Antibiograms display information that can be used to raise awareness of resistance problems, support the use of optimal empiric therapy, and identify opportunities to reduce inappropriate antibiotic usage.[1] Most hospitals distribute an institution-specific antibiogram as a static document that is generated from laboratory data on an annual basis.

Anesthesiologists and surgeons must select and administer perioperative antibiotics either for prophylaxis or to treat an active systemic infection. To assist in this decision-making process, we developed a secure, Web-based, institution-specific, user-friendly visual analytics antibiogram dashboard using electronic health record (EHR) data in near real-time that can be accessed in the operating room setting using the anesthesia information management system computer workstation.

Methods: We created a visual analytics antibiogram dashboard using both Structured Query Languge queries of our EHR database and enterprise visual analytical software to track bacterial pathogens and their antimicrobial sensitivity at CHOP. (Figure 1) The antibiogram dashboard provides a user interface to explore our hospital's laboratory EHR data in near-real time and facilitates the rapid assessment of susceptibilities and resistances of microorganisms to various antibiotics.

Results: A visual analytics antibiogram dashboard specific to our institution was designed and implemented as described in the methods. The dashboard allows the user to display up-to-date, hospital-specific antibiotic sensitivity data for a particular organism using a variety of filters, groupers, and drop down menus.

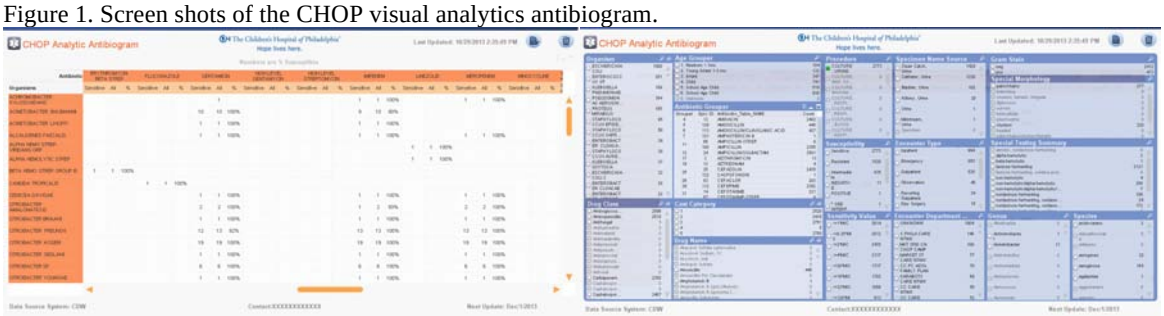
Discussion: Pediatric anesthesiologists and surgeons must consider many factors when selecting and administering perioperative antibiotics. While infectious disease specialists usually guide the antibiotic choices and dosages, there remains a dearth of information at the time of antibiotic administration in the operating room regarding the susceptibility of organisms to the chosen antibiotic. We created a visual analytics antibiogram dashboard that incorporates near-real time laboratory and pharmacy data regarding organism speciation, antibiotic susceptibility, and the cost of antibiotic options. This dashboard will be an integral part of a project to optimize perioperative antibiotic treatment based on our hospital's EHR data.

References

1. Hindler JF, Stelling J. Analysis and presentation of cumulative antibiograms: a new consensus guideline from the Clinical and Laboratory Standards Institute. Clin Inf Dis 2007; 44(6):867-73.

Figures

Figure 1. Screen shots of the CHOP visual analytics antibiogram.



Associations between Use of Electronic Health Record Features and Health Care Quality in Ambulatory Care

Jessica S. Ancker, MPH, PhD,[1] Lisa M Kern, MD, MPH,[1] Alison Edwards, MStat,[1]
Michael Silver, MS,[1] Sarah Nosal, MD,[2] Daniel M Stein, MD, PhD,[1]
Diane Hauser, MPA,[2] Rainu Kaushal, MD, MPH,[1]

1. Weill Cornell Medical College, Department of Healthcare Policy and Research, New York, NY; 2. Institute for Family Health, New York, NY

Abstract: *Characterizing individual physician variations in use of EHR features may be informative as a predictor of the effects of EHRs. The current study is designed to assess the relationship between use of individual EHR features and "meaningful use" healthcare quality measures at the physician level. Use of EHR features and 17 "meaningful use" clinical quality measures were evaluated on the basis of encounters from January 2012 through June 2013. Use of condition-specific best-practice alerts (BPAs) and order sets was associated with better scores on clinical quality measures that captured processes relevant to diabetes, cancer screening, tobacco cessation, and pneumonia vaccination. In the early stages of the "meaningful use" program, use of specific EHR features was associated with modest improvements in certain quality measures focusing on healthcare processes, rather than healthcare outcomes. Longer-term follow-up may be required to link improved processes to outcome performance.*

Introduction: Studies of the effects of electronic health records (EHRs) on healthcare quality have produced mixed results. As more physicians adopt EHRs in the era of "meaningful use," characterizing individual physician use of EHR features, rather than simply assessing the availability of an EHR, may create more informative predictors of the effects of EHRs (1, 2). The current study is designed to assess the relationship between use of individual EHR features and "meaningful use" healthcare quality measures.

Methods: In this cross-sectional study, we included all encounters by primary care providers eligible for meaningful use at a federally qualified health center (FQHCs) network. Use of EHR features and 17 "meaningful use" clinical quality measures were evaluated on the basis of encounters over 18 months. Negative binomial regression models were run to assess the association between feature use and quality on the level of the provider.

Results: Sixty-five providers were included. Use of condition-specific best-practice alerts (BPAs) and order sets was associated with better scores on clinical quality measures that captured processes relevant to diabetes, cancer screening, tobacco cessation, and pneumonia vaccination. For example, responsiveness to breast cancer screening alerts was associated with higher rates of breast cancer screening (incidence rate ratio [IRR] 1.004; $p < .001$) and responsiveness to diabetes-relevant alerts was associated with better performance on measures of diabetes foot exam performance (IRR 1.003; $p = .004$) and LDL testing (IRR 1.006; $p < 0.001$). By contrast, use of the same EHR features was not positively associated with performance on most quality metrics that focused on healthcare outcomes (such as number of diabetes patients with LDL < 100 mg/dL).

Discussion: In the early stages of the "meaningful use" program, use of specific EHR features was associated with modest improvements in certain health care quality process metrics assessed through EHR data, but not with most outcomes-focused quality metrics. Longer-term follow-up may be required to determine whether improved processes will lead to improved outcome performance. The EHR features associated with improved processes may be instrumental in improving the quality of healthcare delivered by physicians who use EHRs. The results may also provide guidance for continued improvements in EHR design as well as end-user training.

References

1. Ancker JS, Kern LM, Abramson E, Kaushal R. The triangle model for evaluating the effect of health information technology on healthcare quality and safety. JAMIA. 2011;18:749-53. Epub August 23, 2011.
2. Lanham HJ, Sittig DF, Leykum LK, Parchman ML, Pugh JA, McDaniel RR. Understanding differences in electronic health record (EHR) use: linking individual physicians' perceptions of uncertainty and EHR use patterns in ambulatory care. Journal of the American Medical Informatics Association. 2014;21(1):73-81.

Author: Onur Asan, PhD

Center for Patient Care and Outcomes Research, Division of General Internal Medicine, Medical College of Wisconsin, Milwaukee, WI, USA

Title: Physicians' Use of Electronic Health Records (EHRs) as a Communication Tool in Primary Care Visits.

Introduction: Some recent studies advocate that EHRs can be used as a tool to communicate and share information through screen with patients during the clinical encounter; therefore, EHRs in the exam room can contribute to patient-centered goals such as making patients more empowered and activated. Our previous studies indicated that, not only do different physicians tend to take very different approaches regarding interacting with EHR, but a single physician might also have different screen sharing behaviors from visit to visit. The goal of this study is to explore work system factors influencing physicians' information (screen) sharing behaviors in the visit.

Methods: Purposive convenience sampling method was followed to recruit physicians from the clinics. We conducted semi-structured interviews with 14 physicians in total. The interview guide and questions were developed based on the conceptual framework of the study: the work system model. Qualitative content analysis was used to analyze the interview data to identify factors influencing physicians' screen (information) sharing styles during primary care encounters.

Results: The qualitative interview identified factors influencing physicians' decisions to share the screen (20 factors), not to share (14 factors) the screen, and identified 21 facilitators which makes it easier to share the screen and 40 barriers which makes it harder to share the screen. These factors are all grouped under broader categories (educating patients, reviewing results/records, patient interest, physician-patient interaction, decision making, patient demographics, time, policies, physical/mental health, sensitive information, confidentiality, information display and density, patient interest, physician computer skills, patient demographics) and linked to appropriate work system elements.

Discussion: This study showed that there are several significant factors influencing physicians' decisions and behaviors about screen sharing, such as patient related and organizational factors. It was interesting to see that sometimes a single item was reported as a factor influencing a physician's decision not to share the screen by one physician, and a barrier which makes it harder to share screen during the visit by another physician. Physicians mostly reported that screen sharing is beneficial for patients and helps patients to better understand the information provided and become more involved in their own healthcare, although they also noted that current EHRs are not designed as a communication tool. Physicians also reported several advantages of screen sharing: it is a good way to educate the patient, it improves new patient trust in the physician, it helps in shared decision making, and it helps patients to visualize certain things that are hard to conceptualize without a visual aid. Physicians stated that these advantages cause screen sharing to have the potential to foster collaboration between physician and patient, leading to better outcomes in the end, better patient education, better relationships with the patient since the physician in engaging them in their healthcare by sharing the screen, better team work between the patient and physician to improve the patient's health, improved patient understanding of health plans and entered records, and making data entry more transparent so patients have a better trust in physician. Furthermore, two main reasons for avoiding screen sharing were also supported by other studies: 1-Documenting or reviewing sensitive information or legal issues, fake pain, obesity, child abuse, suicide. 2- Lack of time and time pressure. This study is vital to understand barriers and facilitators to HIT information sharing in order to inform better EHR designs.

Multiple Perspectives on Clinical Decision Support: A Qualitative Study of Fifteen Clinical and Vendor Organizations

Joan S. Ash, Ph.D.[1], Dean F. Sittig, Ph.D.[2], Carmit K. McMullen, Ph.D.[3], Adam Wright, Ph.D.[4,5,6], Arwen Bunce, M.A.[3], Vishnu Mohan, M.D.[1], Deborah J. Cohen, Ph.D.[1], Blackford Middleton, M.D., M.P.H., M.Sc.[7]

[1] Oregon Health & Science University, Portland, OR, USA
[2] University of Texas School of Biomedical Informatics, Houston, TX, USA
[3] Kaiser Permanente Center for Health Research, Portland, OR, USA
[4] Brigham and Women's Hospital, Boston, MA, USA,[5]Harvard Medical School, Boston, MA, USA[6] Partners HealthCare, Boston, MA, USA
[7] Vanderbilt University, Nashville, TN, USA

Introduction

Primarily due to current U.S. government incentives through the American Reinvestment and Recovery Act (ARRA),[1] hospitals and ambulatory care organizations are increasingly purchasing commercial electronic health record (EHR) systems with computerized clinical decision support (CDS), and/or they are buying CDS directly from content development vendors. Challenges to CDS development, management, and use are sociotechnical in nature, involving people, processes, and technology.[2]

Prior studies concerning the sociotechnical aspects of CDS have focused on the perspectives of individual healthcare organizations [3-6] or vendors [7] in isolation, with no studies comparing the differing perspectives of each stakeholder group. Moreover, the perspectives of commercial organizations have largely been overlooked.

To capture a picture of the entire CDS landscape including views of multiple stakeholders within and outside healthcare organizations, we pose the following research question: How are the views of clinical stakeholders, CDS content vendors, and EHR vendors alike or different with respect to challenges in the development, management, and use of CDS?

Methods

We used the Rapid Assessment Process (RAP), as previously described, [8] for studying 15 organizations, though we adapted it significantly when studying vendor sites.[7] Institutional review boards (IRBs) at the investigators' institutions and each clinical site with an IRB approved the study. For clinical sites, five inpatient and five outpatient sites were selected based on variation in commercial system used, maturity of CDS use, geography, and governance structure. For commercial sites, we purposively [9] selected three different types of content vendors so that we could gain a more comprehensive understanding of the issues. To select EHR vendor sites, we asked a group of experts from healthcare and industry who regularly offer advice about our investigations [10] to help select one primarily ambulatory EHR firm and one primarily inpatient EHR vendor with strong CDS products from among those used at one or more of our clinical study sites. This would allow us to directly compare what users and vendors said about the same CDS products. At clinical sites, we selected subjects for interviewing and observing based on their CDS-related roles. We made an effort to seek out clinical champions, normal users, and skeptical users in addition to CDS experts. During content vendor visits, we interviewed individuals in particular roles, including the CEO, vice presidents, content development and management staff, technical/interoperability staff, and informaticians. For the EHR vendors, we targeted staff members who were most involved with CDS. Data collection primarily consisted of semi-structured interviews and, at clinical sites, observations. Broad areas explored during interviews and observations at all sites included 1) the meaning of CDS, 2) the culture and history of the organization, 3) knowledge management practices, 4) CDS roles, and 5) challenges and the future of CDS. At clinical sites, we also asked about governance of CDS and at commercial sites we discussed their products, customer use of the products, and the marketplace. A grounded theory content analysis approach was used for analyzing data. [11]

Results

We conducted 206 formal interviews and 268 hours of observation. Figure 1 graphically displays the multiple perspectives of the three groups and the overlapping of themes among them. The groups share views on the importance of appropriate manpower, careful knowledge management, CDS that fits user workflow, the need for

communication among the groups, and for mutual strategizing about the future of CDS. However, views of usability, training, metrics, interoperability, product use, and legal issues differed.

Figure 1. Themes shared by clinical and vendor organization representatives.

Discussion and Conclusion

Recommendations for improvement include increased collaboration, legal protections, manpower training, and research about CDS sharing. A national impetus to improve the value and safety of CDS that would include legal protections and incentives for sharing content would be beneficial.

References

1. H.R. 1 [111th]: American Recovery and Reinvestment Act of 2009 (Gov-Track.us) [Internet]. Available from: http://www.govtrack.us/congress/bill.xpd?bill=h111-1. Accessed March 7, 2014.
2. Ahmad F, Norman C, O'Campo P. What is needed to implement a computer-assisted health risk assessment tool? An exploratory concept mapping study. *BMC Med Inform Dec Mak* 2012;**12**:149
3. Bates DW, Kuperman GJ, Wang S, et al. Ten commandments for effective clinical decision support: Making the practice of evidence-based medicine a reality. *J Am Med Inform Assoc* 2003;**10**(6):523-30.
4. Lesselroth B, Yang J, McConnachie J, Brenk T, Winterbottom L. Addressing the sociotechnical drivers of quality improvement: A case study of post-operative DVT prophylaxis computerized decision support. *BMJ Qual Safety* 2011;**20**(5):381-9.
5. Campion TR Jr., Waitman LR, Lorenzi NM, May AK, Gadd CS. Barriers and facilitators to the use of computer-based intensive insulin therapy. *Int J Med Inform* 2011;**80**(12):863-71.
6. Goldstein MK, Coleman RW, Tu SW, Shankar RD, O'Connor MJ, Musen MA, Martins SB, Lavori PW, Shlipak MG, Oddone E, Advani AA, Gholami P, Hoffman BB. Translating research into practice: Organizational issues in implementing automated decision support for hypertension in three medical centers. *J Am Med Inform Assoc* 2004;**11**(5):368-76.
7. Ash JS, Sittig DF, McMullen CK, McCormack J, Wright A, Bunce A, Wasserman J, Mohan V, Cohen DJ, Shapiro M, Middleton B. Studying the vendor perspective on clinical decision support**.** *AMIA Annu Symp Proc* 2011:26-30.
8. McMullen CK, Ash JS, Sittig DF, Bunce A, Guappone K, Dykstra R, Carpenter J, Richardson J, Wright A. Rapid assessment of clinical information systems in the healthcare setting: An efficient method for time-pressed evaluation. *Meth Inform Med* 2010;**50**(2):299-307.
9. Berg BL, Lune H. *Qualitative Research Methods for the Social Sciences, 8th ed*. Boston, MA: Pearson, 2012.
10. Ash JS, Sittig DF, Guappone KP, Dykstra RH, Richardson J, Wright A, Carpenter J, McMullen C, Shapiro M, Bunce A, Middleton B. Recommended practices for computerized clinical decision support and knowledge management in community settings: a qualitative study. *BMC Med Inform Decis Mak* 2012;12:6.
11. Crabtree BF, Miller WL, eds. *Doing Qualitative Research, 2nd edition*. Thousand Oaks, CA, Sage, 1999.

An Interactive Web-based Interview to Improve Family Medical History Documentation

Adarsha S Bajracharya, M.D., [1] Bradley H Crotty, M.D., M.P.H. [1], Hollis B Kowaloff, B.A., [1]
Warner V Slack, M.D., [1] Charles Safran, M.D., M.S. [1]
1. Division of Clinical Informatics, Department of Medicine, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston MA

Introduction: The family medical history (FMH) is underutilized in primary care for risk stratification. Primary care visits are short, and detailed FMHs are often not obtained due to lack of adequate systems for collection and synthesis.(1) Electronic tools for patients can facilitate the collection of these data so that providers have more time to discuss preventive health care.(2) Beth Israel Deaconess Medical Center has been studying computer-based health care interviews since the 1970s. In the early 1990s we deployed an interview for all hospital employees that identified medical problems modifiable by behavior change.(3) Prior studies have shown that computer-based interview modules are well accepted by patients. Studies have shown that both doctors(4) and patients(5) prefer these methods of data collection. A 2009 NIH State of the Science Conference concluded that studies examining methods for more efficiently collecting FMH in primary care settings were urgently needed.(6) The electronic capture of structured family history for first degree relatives is now a Meaningful Use Stage 2 objective, which makes the development and testing of these types of tools timely.(7)

Objectives: Describe use of a new family medical history interview module after deployment in a hospital patient portal.

Design, setting, and participants: We developed and deployed an interactive web-based FMH interview within PatientSite at Beth Israel Deaconess Medical Center in Boston, MA. BIDMC is an academic medical center, and PatientSite, the patient web portal, has over 60,000 registered users, all of whom have access to the FMH interview. Patient use data were collected prospectively, measuring the frequency of access to the interview as well as to the patient portal together with the age and sex of the participants.

Intervention: The computer-based family medical history was developed based on years of experience with computerized medical histories.(4) Patients respond either by clicking on their answer from a list of choices or by typing their entries. The module begins by asking about a particular condition in the family and if the patient acknowledges having that condition in the family, the interview goes on to ask details of that family member. Additionally, the interview allows patients to stop and then to return in the future to complete the interview, if they wish. It collects information about 39 health conditions for first- and second-degree relatives, and patients can also enter health conditions that run in their families but are not included in the structured part of the interview. Information generated is stored in the patient's electronic medical record for providers to review. The module was made available to all patient portal users without formal announcement.

Outcome: The primary outcome measure was the number of patients who used the family medical history interview module.

Results: 44,910 patients accessed PatientSite between 1/1/2014 and 3/12/2014. In the first 51 days of deployment (1/21/14 through 3/12/14), 1,102 patients accessed the FMH module 1,451 times, and of these, 545 (377 women and 168 men) completed the family history interview. Of those who completed the module, rates varied by age (See fig 1). Patients answered to family medical questions in yes, no, Uncertain, don't understand or I'd rather not answer format and, 28 out of 12,326 responses were noted to be in don't understand category. (See fig 2).

Discussion: Our study is one of the first to report on operational deployment of a family medical history interview tool within a patient portal, truly integrated with the EHR. Our results show that patients of all ages—including those over 75 years—are interested in interested in collaborating with their clinicians to provide their family health information, and well-framed questions may contribute to patient use of such a tool. We also note that the module was deployed without any announcement or invitation and yet had rapid participation. We plan to explore why these specific patients used the tool, how they found out the tool was available, and whether online tools like these improve chart completeness, facilitate patient engagement, and support clinical care.

Figures and tables

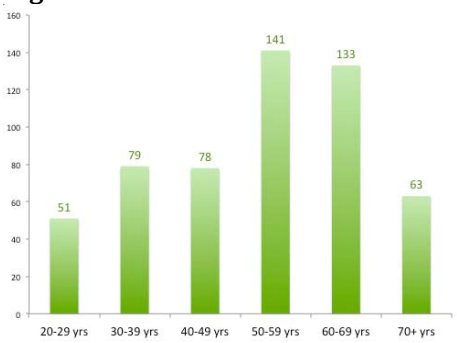


Figure 1. Age distribution of patients who completed interview

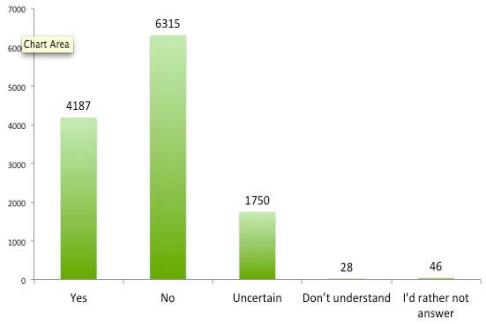
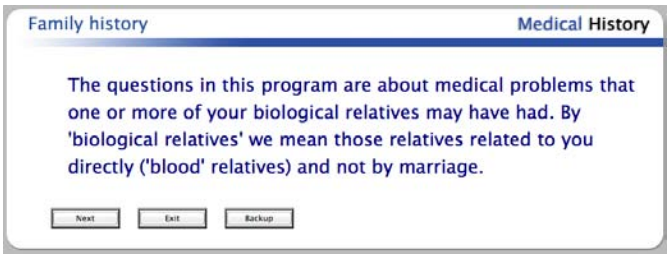
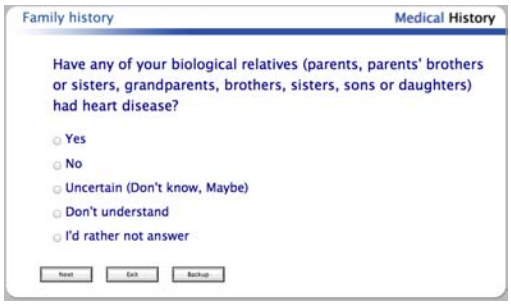


Figure 2. Frequency distribution of patient responses to the family medical questions

Interview Screen shots





The module is available online at:
https://holmes.caregroup.org/scripts/mgwms32.dll/?MGWLPN=MYCROFT&RUN=CMHEMRFamHx&MRN=0901910&token=56a91460-64f1-459a-840a-69bf6a8d512b

References

1. Baer HJ, Schneider LI, Colditz GA, Dart H, Andry A, Williams DH, et al. Use of a web-based risk appraisal tool for assessing family history and lifestyle factors in primary care. J Gen Intern Med. 2013 Jun;28(6):817–24.
2. Berg AO, Baird MA, Botkin JR, Driscoll DA, Fishman PA, Guarino PD, et al. National Institutes of Health State-of-the-Science Conference Statement: Family History and Improving Health. Ann Intern Med. 2009 Dec 15;151(12):872–7.
3. Slack WV, Safran C, Kowaloff HB, Pearce J, Delbanco TL. Be well!: a computer-based health care interview for hospital personnel. Proc Annu Symp Comput Appl Sic Med Care Symp Comput Appl Med Care. 1993;12–6.
4. Slack WV, Kowaloff HB, Davis RB, Delbanco T, Locke SE, Safran C, et al. Evaluation of computer-based medical histories taken by patients at home. J Am Med Inform Assoc JAMIA. 2012 Aug;19(4):545–8.
5. Murray MF, Giovanni MA, Klinger E, George E, Marinacci L, Getty G, et al. Comparing Electronic Health Record Portals to Obtain Patient-Entered Family Health History in Primary Care. J Gen Intern Med. 2013 Dec 1;28(12):1558–64.
6. Feero WG. Connecting the dots between patient-completed family health history and the electronic health record. J Gen Intern Med. 2013 Dec;28(12):1547–8.
7. Office of the National Coordinator for Health Information Technology (ONC), Department of Health and Human Services. Health information technology: standards, implementation specifications, and certification criteria for electronic health record technology, 2014 edition; revisions to the permanent certification program for health information technology. Final rule. Fed Regist. 2012 Sep 4;77(171):54163–292.

Earlier Switching from Intravenous to Oral Antibiotics Due to eReminders

Patrick E. Beeler, MD[1], Stefan P. Kuster, MD[2], Emmanuel Eschmann, MD[1], Rainer Weber, MD[2], Jürg Blaser, PhD[1]

[1]Research Center for Medical Informatics, University Hospital Zurich and University of Zurich, Switzerland

[2]Division of Infectious Diseases and Hospital Epidemiology, University Hospital Zurich and University of Zurich, Switzerland

Abstract

The impact of electronic reminders encouraging early switches from intravenous to oral administration of antibiotics was studied in a hospital-wide, prospective, controlled trial. The reminders were displayed within electronic health records 60h after onset of intravenously administered antibiotic therapies if some patient-specific conditions were met. This intervention fostered early switching from intravenous to oral antibiotics, and thereby significantly reduced the mean duration of intravenous administration by 17%.

Introduction

Paper-based interventions have been shown to encourage switching from intravenous (IV) to oral (PO) administration of antibiotics thereby improving antimicrobial regimens. Important advantages are associated with an early IV-PO switch: Lower risk of catheter-associated infections, reduced nursing workload, and decreased direct and indirect costs. The purpose of this study was to determine whether automated electronic reminders are able to promote early switching.

Methods

In this controlled before-and-after study, an algorithm – starting 60h after onset of an IV antibiotic – automatically checked whether: (i) the therapy was scheduled for an additional 24h or longer, (ii) neutrophil count of the patient exceeded 0.5 G/l, (iii) body temperature was below 38°C, and (iv) patient had the ability to swallow as indicated by active PO orders. If these conditions were met, a non-interruptive red bar was displayed within the top section of the electronic health record (EHR). By clicking the reminder bar a window appeared, offering guidance on whether or not an IV-PO switch was appropriate. The reminder was displayed in the EHR from 60 hours onwards until a physician acknowledged the notification, or the IV therapy triggering the reminder was stopped. However, the reminder was automatically terminated 10 days after its appearance. These reminders were displayed in 12 divisions during the intervention period (year 2012). In contrast, no reminders were visible during the baseline period (year 2011) and in the control group (17 divisions). Comparisons of the durations of therapies were performed using the log-rank test. Levels of $p \leq 0.05$ were considered significant.

Results

A total of 22,863 IV antibiotic therapies were analyzed, and 6,082 (27%) were switched to PO. In the intervention group, 757 IV therapies were administered for a mean duration (±standard deviation) of 5.42 (±8.14) days before switching to PO in the baseline period and 794 courses for 4.47 (±5.49) days in the intervention period ($p=0.0035$), corresponding to a 17% reduction of the IV administration time. No significant decrease of the IV administration time was observed in the control group. The top five IV antibiotics that had been switched to PO were amoxicillin/enzyme inhibitor (EI), representing 45.5% of the switched IV therapies (most often switched to →PO amoxicillin/EI), piperacillin/EI (13.1%; →PO amoxicillin/EI), ceftriaxone (7.1%; →PO amoxicillin/EI), ciprofloxacin (5.7%; →PO ciprofloxacin), and cefuroxime (5.6%; →PO cefuroxime).

Discussion

The electronic reminders significantly reduced the time until switching from IV antibiotic therapies to the oral route of administration by 17%. They were triggered with a delay of 60h after onset of the IV treatment and only if some patient-specific conditions were met, in order to limit alert fatigue. As opposed to previous studies using paper-based interventions, this computer-based approach allows for a hospital-wide implementation with an open-end intervention period since no manpower is required for the continued operation of electronic reminders within the EHR.

Cardiorespiratory physiological data as an indicator of morphine pharmacokinetics and pharmacodynamics in critically ill newborn infants

Nadja Bressan, Ph.D. [1,2], Carolyn McGregor, Ph.D. [1], Andrew James, MBChB, MBI. [2,3]
[1]Faculty of Business and Information Technology, University of Ontario Institute of Technology, Oshawa, ON, Canada; [2]Division of Neonatology, The Hospital for Sick Children and [3]Department of Pediatrics, University of Toronto, Toronto, ON, Canada

Introduction

Morphine is a natural opioid that acts as an agonist at the mu and kappa receptors, which are receptors for analgesia and sedation. It is the commonest drug used for analgesia in newborn infants. Morphine pharmacokinetics and pharmacodynamics (PK/PD) in neonates are not well understood. The objective of this study is to use simulation software for the estimation of drug concentration and for the correlation of morphine concentration with heart rate variability.

Methods

Newborn infants admitted to the Neonatal Intensive Care Unit (NICU) at The Hospital for Sick Children who were enrolled in an institutional Research Ethics Board approved study of neonatal infection and received morphine for analgesia were enrolled in this study. A three compartmental model structure with a biophase was used to implement Krekels' model parameters for morphine[1]. The compartments were described in Krekels' thesis: concentration in a fourth compartment is used for correlation with heart rate variability (HRV) [2]. Four differential equations were modeled by state space approach. The bodyweight and clearances were calculated using an allometric equation with a scaling factor (K) of 1.44. The morphine constant infusion (ml/h) titrated to the premature infants was considered as the input of the PK model. The output of the discrete state space function is used to estimate concentrations every 10 seconds. This study considered gestational age (GA) and postnatal age (PNA) to establish the premature infant weight in the allometric equation to be used in the PK model as shown in Figure 1. PK model output, represented by morphine concentration (um/ml), was synchronized with physiological data acquired through the Artemis platform[3]. The PK algorithm was implemented using IBM's Infosphere Streams software together with Streams Processing Language (SPL) in Artemis as a sub-entity simulator. Artemis is a big data platform operational in the NICU at The Hospital for Sick Children since August 2009[2]. A biharmonic-type interpolation method was performed using MatLab® to solve a 3D scattered data interpolation problem considering morphine concentration, gestational age and HRV. Premature infants and extremely immature preterm were used to design the curve response and to validate the model. The total population was divided into two sub-groups with 12 subjects each, one to design and the second to validate the curve fitting.

Results

Twenty-four subjects, gestational age 24-41 completed weeks (mean ± SD: 33.25 ± 5.42) and birth weight 700-3300 grams (mean ± SD: 2440 ± 910), were studied under morphine concentrations at 100ug/ml, 80ug/ml, 40ug/ml, 20ug/ml. The results demonstrated that the morphine concentration estimated by Krekels' model correlates with heart rate variability for subjects in the gestational age range between 32 and 40 weeks. HRV presented a poor correlation with morphine concentration in subjects between 30 and 24 weeks, as shown in Figure 2. The design group presented a R^2=0.6251; the goodness of validation group presented a Sum of Squares Due to Error (SSE) = 824937.

Discussion

The correlation between drug concentration and heart rate may be considered as a pharmacodynamic marker for morphine in the neonatal population. The modeling of morphine PKPD for the neonatal population must be revisited to consider the immaturity of the newborn infant as indicated by gestational age. The PKPD models for the extremely immature preterm population may be improved through the inclusion of maturation as a covariant.

References

[1] Krekels, Elke Henriëtte Josephina. Size does matter: drug glucuronidation in children. Division of Pharmacology, Leiden/Amsterdam Center for Drug Research (LACDR), Faculty of Science, Leiden University, 2012. [2] McGregor, Carolyn, Catley Christina, James, Andrew. Variability analysis with analytics applied to physiological data streams from the neonatal intensive care unit. Computer-Based Medical Systems (CBMS), 2012 25th International Symposium on. IEEE, 2012.[3] McGregor, Carolyn. Methodologies and applications of continuous variability analysis, Journal of Critical Care, vol. 25(3), p. e6, 2010.

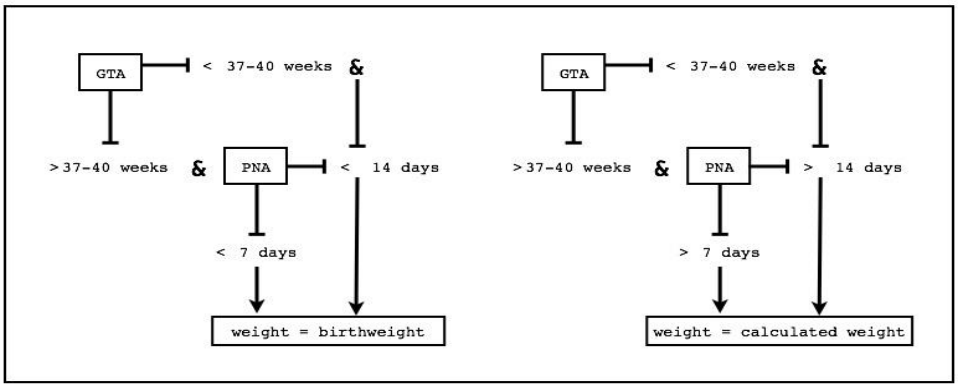


Figure 1. Algorithm for the determination of the actual weight to be used in PKPD modeling.

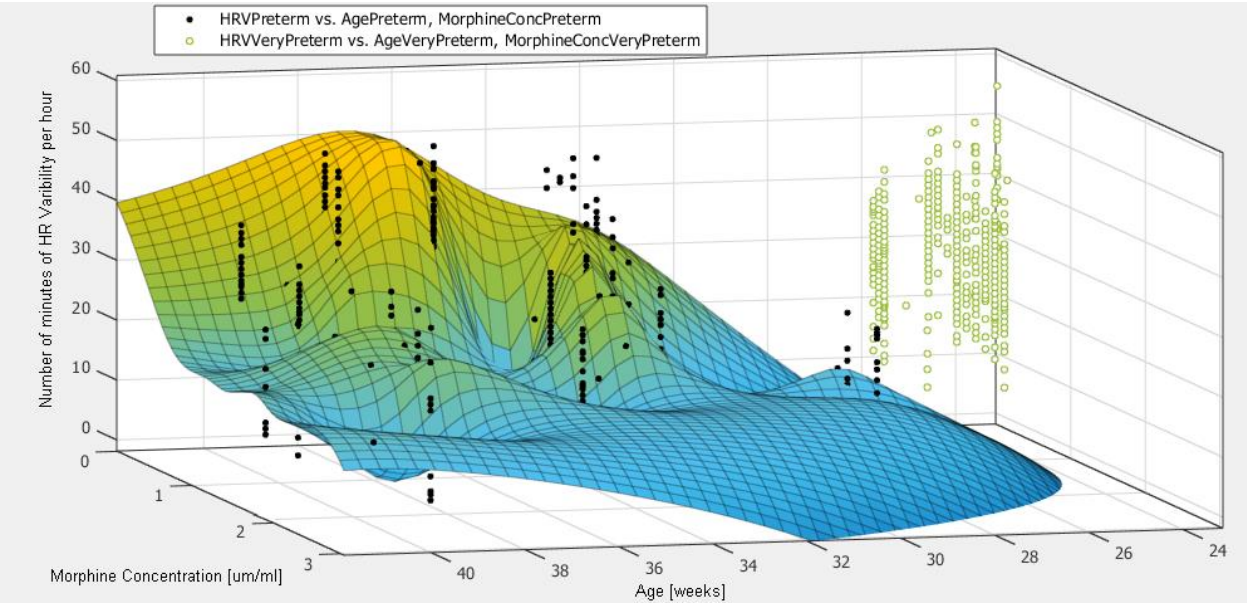


Figure 2. Curve Response Heart Rate Variability, Gestational Age and Morphine Concentration.

Evolving Collaboration Patterns in Medical Research

Jason Cory Brunson[1], Xiaoyan Wang[1,2], and Reinhard Laubenbacher[1]

[1]Center for Quantitative Medicine, University of Connecticut Health Center, Farmington, CT
[2]Department of Family Medicine, University of Connecticut Health Center, Farmington, CT

Abstract Bibliographic databases are a mainstay of scientometrics and social network analysis, yet an essential temporal network analysis of the enormous medical literature remains to be conducted. We employ graph-theoretic tools to capture the evolving structure of the medical research community across the 50 years of MEDLINE's existence. Early analysis reveals that the medical research community has undergone periods both of increasing and of decreasing collaboration and cohesion. This suggests the need to identify real-world influences and to understand the range of evolutionary behaviors and their influencing factors across subject- or institute-specific communities.

Introduction The structure of the medical literature provides an underutilized resource for identifying and evaluating research trends. Changes in this structure reflect changes in the practices, resources, and culture of the medical community. Our goal is to perform a comparative analysis of research initiatives and institutes, in order to identify key structural predictors for network outcomes such as the cohesion of communities and the emergence of specialized fields, and to link these predictors to real-world factors such as the funding of initiatives and the founding of institutes. We begin by investigating the structure and evolution of the aggregate medical literature in an historical context.

Methods Our data consist of all MEDLINE entries from 1964 to the present. We employ graph-theoretic models and diagnostics to study the aggregate medical research network as well as several subnetworks organized around specific research institutes or subject matters.

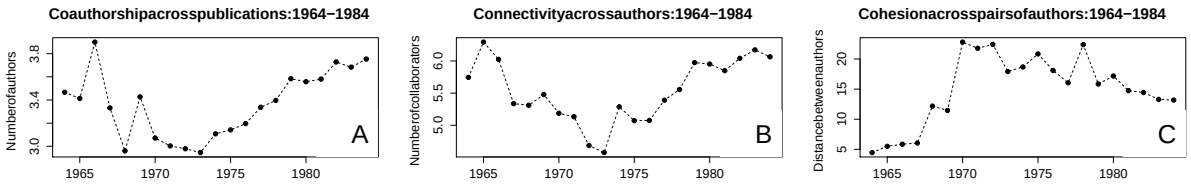


Figure 1: Time series across 1964–1984 for (A) mean coauthors per publication, (B) mean collaborators per author, and (C) mean number of coauthorships required to connect two authors.

Results Preliminary results reveal changing evolutionary behavior in the medical community. At MEDLINE's inception, the community was growing less collaborative, in terms of the typical number of authors on a paper and the number of collaborators of a typical author, and more dispersed, in terms of the 'distance' between authors through coauthorship (see Fig. 1). These trends reversed in the early 1970s, though other common diagnostics of network structure, such as triadic closure and homophily, remained steady. This suggests a profound shift on the way research was conducted between the late 1960s and the late 1970s. A thorough look at subject-specific or institute-specific communities, as well as sensitivity analysis on the whole, will bring this influence into sharper focus.

Discussion Our early results present two directions for further analysis: (1) to detect the effects on network evolution of funding initiatives, the founding of institutes, and other large-scale research-oriented programs, and (2) to understand how research communities organized around specific topics or hosted by specific institutes are distributed across the network diagnostic parameter space. We are currently pursuing both questions.

Assessment of Vital Sign Data Accuracy and Timeliness as Recorded in an EHR

Lorraine R. Buis, PhD,[1] Lisa Gulker, DNP, RN[2] Melissa Plegue, MA[1]

[1] University of Michigan, Department of Family Medicine, Ann Arbor, MI
[2] Applied Clinical Informatics, Tenet Healthcare Inc., Dallas, TX

Introduction: Patient vital signs are one of the most basic and important measurements used within the hospital setting, and vital sign monitoring is an essential component of quality patient care. Vital signs provide a picture of current patient health status, which can aid in treatment protocol decisions, and may lead to early identification for serious conditions such as sepsis, respiratory failure, and other shock states. In order for the algorithms used to predict clinical deterioration to be most effective, we assume that the vital sign data entered into the EHR is both accurate and timely. The goal of this study was to determine the accuracy and timeliness of vital sign data recorded in an EHR in an adult acute care setting.

Methods: We conducted a retrospective records analysis of all vital signs taken over a seven-day period in one acute care unit within a large, eight hospital Midwestern healthcare system. We compared data for seven different vital sign measurements (temperature, systolic blood pressure, diastolic blood pressure, respiratory rate, heart rate, SpO2, and patient self-reported pain scores) that were pulled directly from vital sign monitors each day, to data entered into the EHR. To determine accuracy, we looked for discrepancies between vital sign monitor and EHR data, and to determine timeliness, we compared timestamps from the vital sign monitor and the EHR.

Results: Our sample contained 728 sets of vital signs, from 85 unique patients. Of the seven vital signs captured by the vital sign monitors, for a variety of reasons, only about one-third to one-half of those taken by the vital sign monitor ultimately made it into the EHR. Pain scores were entered least frequently (34% of the time) and blood pressure was entered the most frequently (53% of the time for both systolic and diastolic). Of the measures which were entered into the EHR, error rates between vital sign monitor and EHR were less than 3% for systolic and diastolic blood pressure, as well as temperature. Pain scores had a slightly higher error rate (6%). Respiratory rate had an error rate of 11%, followed by SpO2 (15%). Pulse rate had the highest rate of error (24%), with nearly 1 in 4 pulse rates recorded in the EHR discrepant from what was recorded on the vital sign monitor. Regarding timeliness, on average, it took 49.7 (SD=59.04) minutes from the time vital signs were measured and assessed, to the time they were entered into the EHR. For approximately 15% of vital signs, 90 minutes or more passed from the time of assessment to entry into the EHR. Providers consistently underestimated their lag time from vital sign assessment to EHR entry by an average of 24.9 minutes (SD=60.45).

Discussion: Although the majority of vital signs had low error rates, discrepancies in pulse rate, respiratory rate and SpO2 were unacceptable from a clinical standpoint. Average lag times of nearly 50 minutes from vital sign assessment to EHR entry compromise the ability of healthcare providers to intervene early when clinical deterioration could be detected or predicted. Results from this investigation are essential for building the case for healthcare systems to move toward vital sign integration directly from vital sign monitors to EHR, without the need for human intervention, which may ultimately lead to improved quality of vital sign data and patient care.

Applied Clinical Informatics Best Practices in support of Clinical Next Best Practices: Integrating Knowledge Discovery to Delivery into Workflow

Matthew M. Burton, M.D.[1,2,3,4], David W. Larson, M.D., MBA[2,5], Jenna Lovely, PharmD, BCPS[2,6], Tim Miksch[4], Steve Peters, M.D.[4,7], Tim Larson, M.D.[4,8], John Wald, M.D.[9], Bruce Evans, M.D.[10]

[1]Department of Health Sciences Research, [2]Department of Surgery, [3]Center for the Science of Health Care Delivery, [4]Office of Information and Knowledge Management, [5]Division of Colorectal Surgery, [6]Hospital Pharmacy Services, [7]Division of Pulmonary & Critical Care Medicine, [8]Division of Nephrology & Hypertension, [9]Division of Neuroradiology, [10]Cerebrovascular Neurology, Mayo Clinic, Rochester, MN

Abstract

Consistently delivering Best Practice care to patients in an efficient, effective, and safe manner holds the most promise for transforming not only health care delivery, but the practice of medicine. However, standard use of Best Practice must not only apply to patient care processes, but to the means and mechanisms by which such clinical Best Practices are discovered, disseminated, delivered, and reinforced. Clinical Informatics and Engineering Best Practices further hold much promise to the optimal application of ever more complex and voluminous body of domain knowledge to similarly sophisticated patient information. This study demonstrates the potential for such an approach.

Introduction

Standardized clinical best practices, such as Enhanced Recovery Pathways for the postoperative inpatient setting, serve to ensure that the highest quality care is delivered consistently to achieve the best outcomes (clinical and utilization) in the most efficient and safe manner. Such standardization further lends itself to widely accepted clinical/ health services research as well as informatics, engineering, and management science approaches for continued optimization along all dimensions of quality (safety, timeliness, effectiveness, efficiency, patient-centeredness).

Consistent use of such standards must be integrated into clinician workflows and mental models such that they are easy and intuitive and that deviation from such a standard is obvious, assessable, and actionable. Furthermore, the discovery and rapid integration of newly recognized best practices must be similarly integrated into optimal workflows facilitated through best practice means (processes) and mechanisms (information and knowledge management tools) often co-redesigned. Such an endeavor of continually delivering the very 'next' best practice through effective, efficient, and sustainable means necessitates adherence to likewise advanced best practice health system engineering and applied clinical informatics principles, approaches, and tools.

Methods

For this study, we first employed a mixed-method approach to capturing clinical workflow including mental models and information flow to define the information and knowledge management needs of an academic surgical practice. These methods include semi-structured interviews, flow mapping, direct observation, screen capture, behavioral and cognitive task analysis, and electronic information mining (event logs, report generation). Next, we conducted iterative prototyping and in-situ testing using user experience design methods to design analytics, practice management dashboards, and mobile point of care tools. Lastly, we implemented these designs through Agile software development using various software architecture and medical informatics best practices (SOA, stateful knowledge representation, and standardized terminologies and information models). Study design was a nested, interrupted time-series.

Results

Key clinical workflows and clinician mental models were identified and supporting processes and tools were co-redesigned for optimal performance. Surgical Complications, Utilization (LOS), Time-on-task, and Mental Workload were all significantly reduced while Pathway Compliance was sustained.

Conclusion

We have demonstrated a sustainable, transferable, and diffusible health systems engineering and applied clinical informatics approach to optimally delivering 'next' best practice clinical pathways. We have further extended the principle of standardization on best practice for clinical pathways to the supporting or enabling engineering, informatics, and redesign principles, methods, and tools. We have leveraged informatics and software development best practices for scalable, extensible infrastructure such that these informatics interventions can be sustained, transferred to additional clinical domains, and more broadly disseminated. We have broadly applied process reengineering best practices to eliminate waste in the practice of academic medicine (surgery) in support of knowledge work- from discovery to delivery.

References

1. Lovely, JK. Maxson, PM. Jacob, AK. Cima, RR. Horlocker, TT. Hebl, JR. Harmsen, WS. Huebner, M. Larson, DW. Case-matched series of enhanced versus standard recovery pathway in minimally invasive colorectal surgery. British Journal of Surgery. 2012; 99: 120-6

2. Stead WW, Lin HS, (eds.). Computational Technology for Effective Health Care: Immediate Steps and Strategic Directions. Committee on Engaging the Computer Science Research Community in health Care Informatics. Washington, DC: The National Academies Press, 2009.

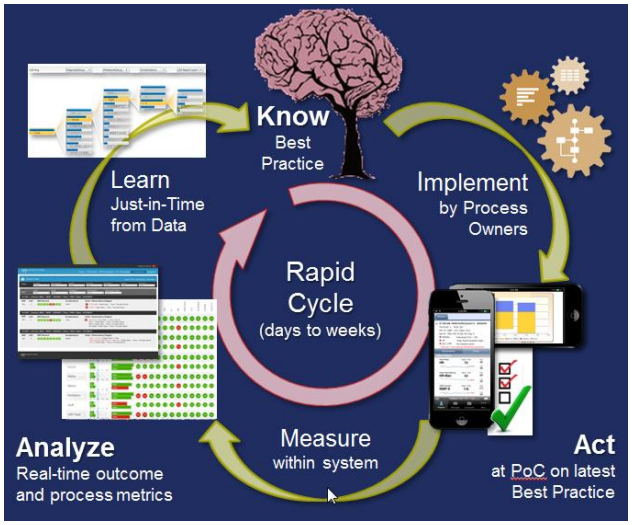


Figure 1. Knowledge-to-Delivery Lifecycle.

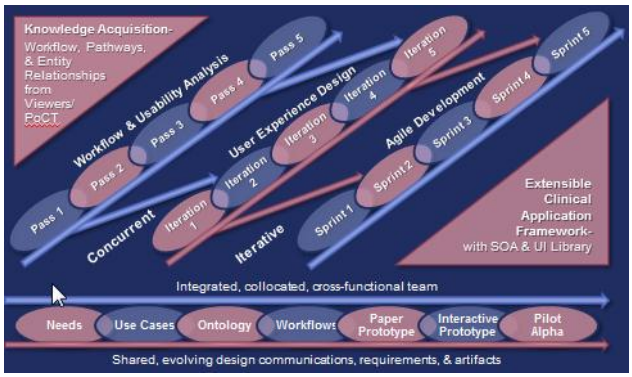


Figure 2. Redesign Best Practices for Informatics Interventions.

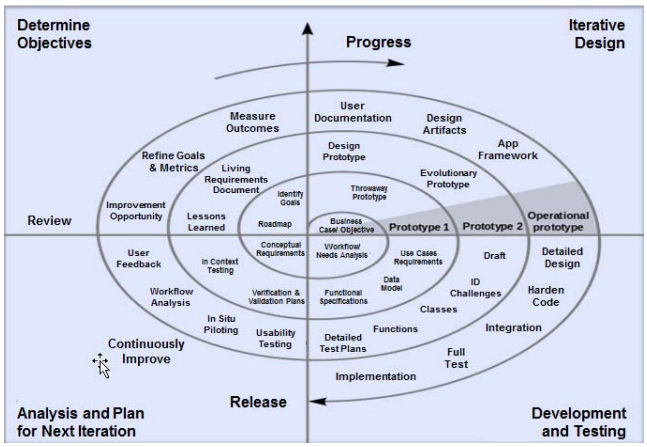


Figure 3. Spiral Model for Rapid-cycle Practice Redesign.

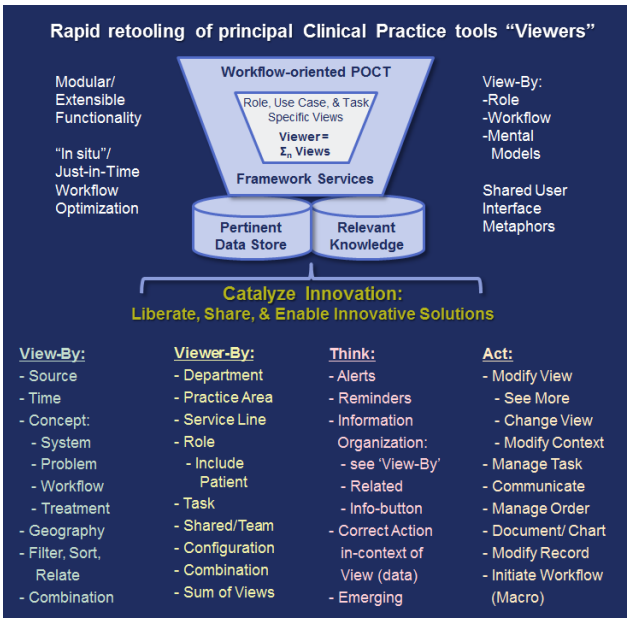


Figure 4. Extensible Clinical Application Framework-Role-oriented Views for Knowledge Discovery to Delivery.

Provider Workflow/ Effort	Current Workflow of Provider (Caring for 9 Patient)	New PoC Tool (mobile) Workflow of Provider
Information Systems	11+	1
Use of Paper Intermediates	5+	0
Manual Pathway/ Complication Calculations	>36	0
Screen Transitions (Inter-application)	237 (43)	25 (0)
Mouse Clicks	619	25
Estimated Cognitive Load Index	1,623	75 (<5% of current)
Time (minutes)	30:14 (95% on navigation)	< 4:30 (95% on Clinical)

Table 1. Pre and Post Workflow and Cognitive Workload Analysis.

PheWAS and Genetics Define Subphenotypes in Drug Response

Robert J. Carroll, MS[1], Jeremy L. Warner, MD, MS[1], Anne E. Eyler, MD[2], Charles F. Moore, MD[1], Jayanth R. Doss, MD[1], Katherine P. Liao, MD[3], Robert M. Plenge, MD, PhD[4], Joshua C. Denny, MD, MS[1]

[1]Vanderbilt University School of Medicine, Nashville, TN;[2] NorthCrest Medical Center, Springfield, TN;[3]Brigham and Women's Hospital, Boston, MA;[4]Merck, Boston, MA;

Introduction

Rheumatoid arthritis (RA) is a chronic, debilitating disease with significant morbidity and mortality. The American College of Rheumatology (ACR) guidelines for treatment of RA include the early use of anti-TNF medications in cases that do not respond to methotrexate or other non-biologic combination regimens[1]. Anti-TNF treatment is expensive, is contraindicated in cases of infections due to the immune modification, and is not always efficacious. Determining the genetic predictors of response and helping elucidate the biology of response are factors important to improving treatment options.

Methods

We identified 141 individuals with a confirmed clinical diagnosis of RA and anti-TNF medication mentions in their clinical documentation from Vanderbilt's Synthetic Derivative, a de-identified image of the EHR linked to the BioVU DNA repository[2]. Clinician review confirmed 103 individuals were administered anti-TNFs. Twenty confirmed individuals were randomly selected for review by a second clinician. The kappa was 0.56 (across response and non-response for three drugs); the disagreements were resolved and improved future review guidelines. To select more charts for review, we identified a threshold of etanercept (Enbrel) mentions from our initial reviews to yield a PPV>98% for true drug exposure. We applied a previously published algorithm for RA case identification and the etanercept mention threshold to select more charts for review. A subset of records was then reviewed by three rheumatologists to determine if they were responders, non-responders, or of indeterminate response to anti-TNF therapy. Individuals were considered responders if there was evidence for response at 6 months after initial treatment. However, if the individual appeared to no longer respond at 12 months, they were labeled a non-responder. A change in patient status indicating non-response after a 12 month period was considered a responder, but flagged with a "fade" effect. In addition, we included those individuals whose response was unsure as "unsure". Kappa was 0.53 over 20 charts shared between two reviewers. We applied a Support Vector Machine (SVM) that used n-grams (up to 5-grams) in a window up to 20 words before and after the drug mentions to predict response. We evaluated the algorithm using 5-fold cross validation. We further expanded our case and control set for the genetic and PheWAS by applying the predictive model to those individuals not reviewed to expand our case and control set. After identifying responders and non-responders, we applied the phenome-wide association study (PheWAS) method to identify clinical phenotypes associated with drug non-response[3]. The results of the PheWAS were used to inform a targeted genetic analysis.

Results

358 total etanercept treated individuals were reviewed in total with 231 responders, 68 non-responders, and 59 of uncertain response. The SVM had an area under the receiver operator characteristic curve of 0.88. The PheWAS identified four associations at a nominal p of 0.01, including degenerative disc disorders, which were disproportionately represented in the non-responding individuals. As axial skeleton involvement is more common in diseases such as ankylosing spondylitis and psoriatic arthritis, which are more commonly associated with *HLA-B27* positivity, we investigated genetic loci on chromosome 6 near the *HLA* and *TNF* genes[4]. We found a number of SNPs in this region associated, including rs13202464 (odds ratio=3.08, p=0.0008). This SNP is known to tag *HLA-B27*. See Figure 1 and Tables 1 and 2 for more results.

Discussion

We developed and coupled phenotype algorithms to identify a pharmacogenetic trait of drug efficacy: specifically, RA patients that responded and failed to respond to anti-TNF therapy. Using PheWAS, we identified clinical associations differentiating these two groups of RA patients, and genetic analysis suggests that non-responders are much more likely to have risk HLA-B27 variants by virtue of this tag SNP. Thus, these results suggest that underlying RA subphenotypes may be responsible for difference in drug response rather than factors more directly related to drug metabolism. Use of PheWAS to identify comorbidities associated with drug efficacy may assist in redefining disease toward a vision of precision medicine.

Figure 1. PheWAS Manhattan plot of response status

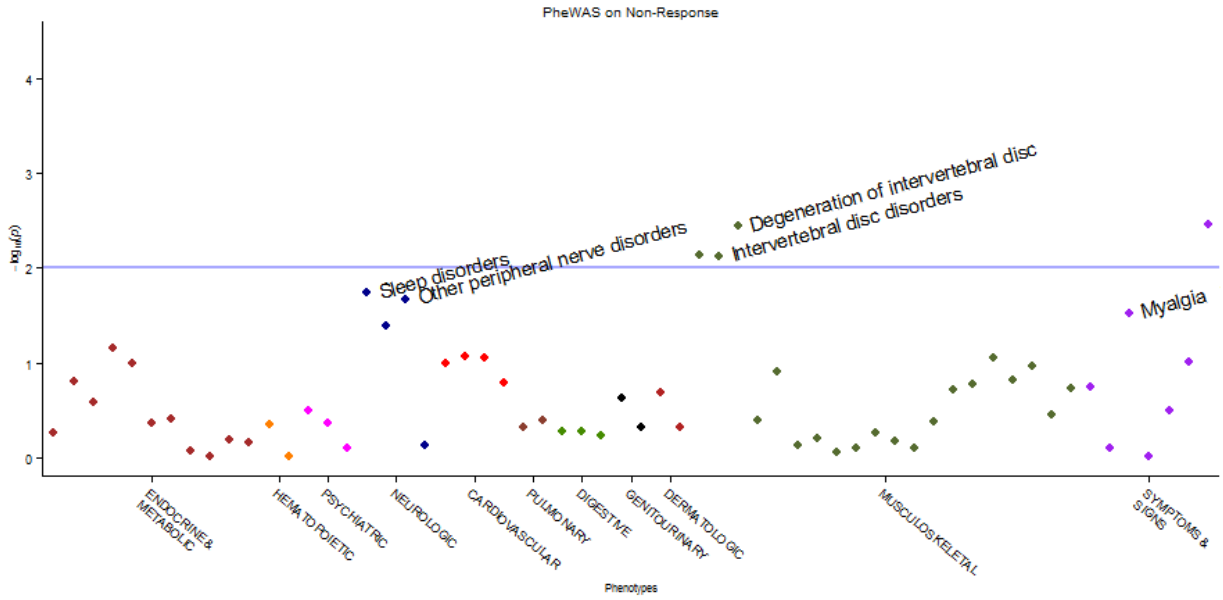


Table 1a. PheWAS results (p<0.01)

PheWAS Phenotype	OR	p
Malaise and fatigue	2.54	0.00346
Degeneration of intervertebral disc	3.64	0.00362
Spinal stenosis	3.64	0.00719
Intervertebral disc disorders	3.05	0.00754

Table 2. Genetic analysis results (p<0.01)

SNP	Variant	OR	P
rs13202464	G	3.08	0.000836
rs9266395	T	0.35	0.00343
rs2523586	C	1.96	0.00471
rs3819299	C	2.69	0.00936
rs9266329	A	0.42	0.00995

References

1 Singh JA, Furst DE, Bharat A, *et al.* 2012 update of the 2008 American College of Rheumatology recommendations for the use of disease-modifying antirheumatic drugs and biologic agents in the treatment of rheumatoid arthritis. *Arthritis Care Res* 2012;**64**:625–39. doi:10.1002/acr.21641

2 Roden DM, Pulley JM, Basford MA, *et al.* Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin Pharmacol Ther* 2008;**84**:362–9. doi:10.1038/clpt.2008.89

3 Denny JC, Ritchie MD, Basford MA, *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics* 2010;**26**:1205–10. doi:10.1093/bioinformatics/btq126

4 A. Brown, Matthew, Behalf of the Wellcome Trust Case-Control Consortium 2, on, Spondyloarthritis Consortium, Australo-Anglo-American. Genome-Wide Association Study Of Ankylosing Spondylitis Identifies New Loci, A Tag SNP For HLA-B27, And An Interaction Between HLA-B27 And Variants In ERAP1. ;Arthritis Rheum 2010;62 Suppl 10 :1355 DOI: 10.1002/art.29121.

Hispanic Patients' Role Preferences in Primary Care Treatment Decision Making

Kenrick Cato, RN, PhD [1], Suzanne Bakken, RN, PhD[1,2]

[1]School of Nursing and [2]Department of Biomedical Informatics, Columbia University, New York, NY

Introduction

Shared decision making (SDM) is considered to be a crucial component of high quality and safe patient-centered primary care treatment[1]. Hispanics are the fastest growing minority group in the United States[2] and they experience substantial health disparities[3]. The aim of this study was to examine the factors that correlate with Hispanics' decision role preferences for participation in treatment decision making with their primary care clinician.

Methods

Hispanic patients (n=772) were recruited from five zip codes in the Washington Heights/Inwood community of New York City and survey data were collected via interview by bilingual community health workers in four New York-Presbyterian Ambulatory Care Network clinics. Data were analyzed using multinomial logistic regression to investigate the association between sociodemographic and health factors and role preference (Control Preference Score) [4] in primary care treatment decision making (passive, shared, active); passive role as the reference range.

Results

Most survey respondents preferred to participate in medical treatment decisions in a shared or active role (90%) and also had inadequate health literacy (95%). The odds of wanting to participate in decision making in a shared role with a primary care provider significantly increased with younger age (OR=0.98, 95% CI [0.96- 0.99], p =0.01), less than 21 years living in the United States (OR=0.48, 95% CI [0.27- 0.88], p =0.02), more adequate health literacy (Newest Vital Sign) (OR=.46, 95% CI [0.25- 0.83], p =0.01), better ability to understand health instructions, pamphlets or written health materials (OR=0.55, 95% CI [0.31- 0.99], p =0.05), and higher social role performance (OR=0.97, 95% CI [0.94- 0.99], p =0.04). Statistically significant odds for preference for an active role were higher education (OR=3.11, 95% CI [1.20- 8.04], p =.02), less than 21 years living in the United States (OR=0.37, 95% CI [0.19- 0.73], p =0.004), and younger age (OR=0.98, 95% CI [0.95- 0.99], p =0.02). However, the overall models demonstrated poor fit with study data explaining 10% -14% of the variation of the dependent variable.

Conclusion

Our analysis suggested a number of patient specific factors that should inform future informatics interventions related to Hispanic patients' SDM with their primary care clinician. This study confirmed the influence of age, depression, years lived in the United States and education on treatment decision control preference for Hispanics. Furthermore, the relationship between variables investigated in our study and their relationship to desired role in SDM for Hispanic primary care patients can be utilized by informatics tools in a number of ways: 1) to customize the user experience. 2) To target sub-populations of patients. For example, since while taking all other variables into consideration, age and education level are strongly correlated SDM, one might design a tool that solicits involvement in treatment decisions in a different way than you might from older, less educated patients. 3) To screen and validate patient SDM role preference. 4) To promote SDM, by providing tailored treatment decision aids to Hispanic primary care patients. In other words, utilizing patient level variables to customize the content of decision aids.

Acknowledgements: This work was supported by the following grants R01HS019853, T32NR007969.

References

1. Barry, M. J. & Edgman-Levitan, S. (2012). Shared decision making — The pinnacle of patient-centered care. The New England Journal of Medicine, 366(9), 780-781.
2. Ennis, S. R. (2011). The Hispanic Population: 2010: 2010 Census Briefs. Washington, DC: Bureau of the Census (DOC).
3. Adler, N. E. & Rehkopf, D. H. (2008). U.S. disparities in health: Descriptions, causes, and mechanisms. Annual Review of Public Health, 29, 235-252.
4. Degner, L. F. (1997). The Control Preferences Scale. Canadian Journal of Nursing Research, 29(3), 21-43.

Next-Generation Terminology Requirements for Interprofessional Care Planning

Sarah Collins, RN, PhD[1,2,3], Kira Tsivkin[1], Stephanie Klinkenberg-Ramirez[1], Dina Iskhakova[1], Hari Nandigam MD, MSHI[1], Perry L. Mar, PhD[1,2,3], Roberto A. Rocha MD, PhD[1,2,3]

[1]Partners HealthCare System; [2]Brigham and Women's Hospital; [3]Harvard Medical School

Abstract

Team-based interprofessional care planning is associated with better care coordination and patient outcomes; yet, known terminology management infrastructures within electronic health records (EHRs) do not support sociotechnical requirements such as linking and sharing documented concepts among various professionals. This study, which will be complete in August 2014, is conceptualizing, developing, and evaluating next-generation approaches to terminology management for interprofessional care planning based on clinician-validated care planning scenarios. We will present our derived set of terminology-specific requirements and standards-based terminology model for coordinated team-based care, knowledge development of process-outcome associations, clinical decision support, and a learning health system based on validated clinical care planning scenarios.

Introduction

EHR terminology infrastructures are typically based on requirements of one profession, using one terminology, with few linkages between concepts. The requirements for interprofessional care planning will expand exponentially for EHRs shared by teams comprised of multiple health professions with overlapping terminology needs. The collaborative and cooperative activity of documenting on a shared care plan introduces interesting dynamics and opportunity and will require "smarter systems" that enable reconciliation of multiple plans of care and treatment plans into a patient centered and longitudinal care plan.[1,2] This study defined requirements and developed a proposed next-generation terminology model for patient-centered interprofessional care planning.

Methods

We developed interprofessional care planning scenarios for validation by an interprofessional panel of subject matter experts (SMEs) from primary, long-term, acute, and critical care. Six clinical scenarios were related to primary care for diabetes, depression and renal failure; episodic care for acute asthma; and critical care for oncology. We refined our conceptual model[3], defined requirements and analyzed terminology management approaches. We developed an information model using Object-Role Modeling (ORM) language that was aligned with HL7 Care Plan Domain Analysis Model. Next we defined a terminology management model that leveraged SNOMED CT and other standard terminologies. The model is being validated against a second set of scenarios and will be presented.

Results

Terminology model requirements with examples derived from clinical scenarios validated by SMEs representing medicine, nursing, pharmacy, social work, care coordination and specialty areas will be described. Requirements include: 1) high level categorization scheme of concerns, goals, and interventions (see Figure 2); 2) sets of profession-specific concerns and interventions (e.g., Primary Care Provider concern of 'uncontrolled type I diabetes mellitus' and Care Coordinator concern of 'patient has deficient knowledge related to glycemic control'); 3) sets of shared patient goals and sub-goals (e.g., 'gain glycemic control' -> 'eat a more nutritious diet' -> 'post-prandial blood glucose of less than 180'); 4) modifiers for goals to capture evaluation toward goal achievement (e.g., 'met', 'not met', 'in progress'); 5) semantic linkages between categories and across clinical concepts and professions (see Figure 2), 6) modifiers for concerns and interventions to communicate decision rationales and anticipatory guidance among the care team in the context of shared care plan documentation (e.g. 'referral to pharmacist to understand any reasons for lack of medication adherence and for medication education').

Conclusion

We identified a set of requirements and a terminology model for interprofessional care planning. The linkages of concepts to meet clinician validated care planning scenarios confirm a need for a "next-generation" terminology modeling approach. Future work should focus on extending and validating our terminology model.

References

1. Patient Care Workgroup. Care Plan Domain Analysis Model, Release 1, September 2013 HL7 Informative Ballot; 2013:77.
2. Zhou X, Zheng K, Ackerman MS, Hanauer D. Cooperative Documentation : The Patient Problem List as a Nexus in Electronic Health Records. In: CSCW 2012 ACM Conference February 11-15, 2012, Seattle, WA; 2012:911–920.
3. Tsivkin K, Collins S. Knowledge Management Terminology Infrastructure to Support Interdisciplinary Plans of Care. In: American Medical Informatics Association; 2013:1376.

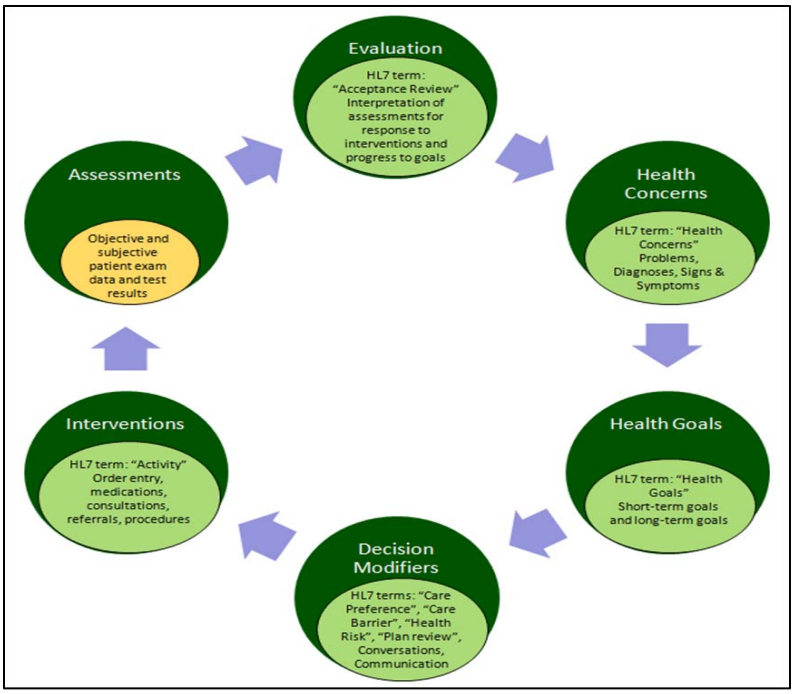


Figure 1. Conceptual Model of Interprofessional Care Planning

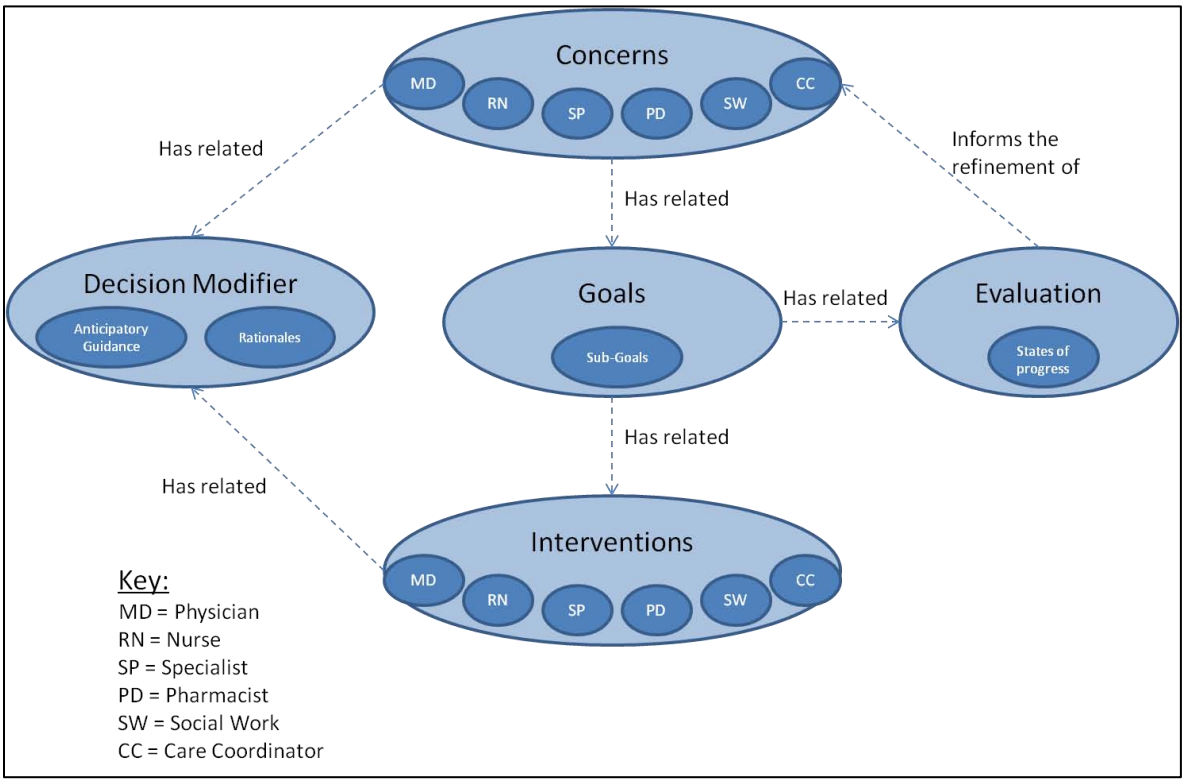


Figure 2. Version 1 Terminology Model Diagram Based on Initial set of Requirements

Acknowledgements: Project funded by a Partners-Siemens Research Council Grant (title: "Knowledge Management Terminology Infrastructure to Support Interprofessional Plans of Care").

Not Just For The Millennials: A Cross-Sectional Analysis Of Secure Messaging Use Among Elders, Families, And Physicians

BH Crotty MD MPH[1], J O'Brien BS[1], M Dierks MD[1], X Lu BS[1], H Feldman MD[1], C Safran MD MS[1]
Division of Clinical Informatics, Beth Israel Deaconess Medical Center

Learning Objectives

1. To determine how elders (≥75 years of age) and families use secure messaging to communicate with their physician

Introduction

As our population ages, patients, families, and caregivers will need an increasingly diverse set of tools for collaboration and coordination of geriatric care. In recent years, secure messaging through patient web portals has become available, but little is known about how elderly patients and their families use the tool.

In this study, we sought to assess trends in the use of secure messaging with providers via a patient portal among patients aged 75 years and over. Additionally, we performed a content analysis to characterize the reasons why patients might be using secure messaging, and to determine if family members also used the portal to communicate with providers.

Methods

This was a retrospective review of secure messaging data sent via the Beth Israel Deaconess Medical Center's patient portal during the period from 2005 to 2010, with chart review of a random sample of messages from 2010. In the first analysis, we obtained and analyzed utilization data from the secure messaging feature of the patient portal on all patients who were registered users of the patient portal, stratifying by age. Each secure message was considered a distinct evaluable event. In the second analysis, we randomly selected 60 messages from unique patients who were users of the messaging feature during 2010. We reviewed the content of each message, categorizing it into one of ten non-mutually exclusive categories, and ascertained whether the sender was the patient or a family member. Concurrently, we reviewed the clinical record determine if the message coincided with any recent ambulatory visits.

Results

In 2010, patients 75+ accounted for 6.1% of all registered patients. Of messages sent by patients to physicians, the percent sent by or on behalf of patients over the age of 75 increased from 2% in 2005 to 5% in 2010, where 541 unique patients sent a total of 4064 messages. Nearly all messages (98%) were sent to primary care providers. Patients sent 85% of messages, while individuals who identified themselves in the message as an informal caregiver of the patient sent 15% of messages from a given patient's account (of these, adult children of the registered patient accounted for 78%). One-third of messages were within 15 days of a visit to the physician receiving the message, and 12% were within 15 days of an emergency room visit. Messages were mostly for medication management (27%), results management (25%), to discuss a clinical concern (23%), or for coordination of care (18%). We found no significant differences comparing messages sent by patients and by family members.

Conclusions

Elderly patients are increasingly using secure messaging, though they still account for a relatively low percentage of total messages. Caregivers, mostly adult children, accounted for 15% of all messages, and appear to use the secure messaging for care coordination and visit follow-up. This suggests a need to develop more robust methods for proxy access for patient caregivers. Notably, questions about coordination of care accounted for nearly one in five messages. As elders age and families take on more caregiving and coordination, asynchronous messaging may play an important role in meeting communication and information needs.

Table: Message content according to the sender in a sample of 60 randomly selected messages. Messages were assigned non-mutually exclusive categories. All messages came from the patient's account.

Message Content	Sent By Patient (n=51)	Sent By Family (n=9)	Overall (n=60)
General Health Question	10%	11%	10%
Coordination of Care	16%	33%	18%
Clinical Concern	24%	22%	23%
Social Concern	2%	0%	2%
Psychiatric Concern	11%	2%	3%
Clinical Update	12%	22%	13%
Visit Follow-Up	2%	22%	5%
Results Management	24%	33%	25%
Healthcare Finances	6%	0%	5%
Medication Management	29%	11%	27%

References

Anand, S.G. et al., 2005. A content analysis of e-mail communication between primary care providers and parents. Pediatrics, 115(5), pp.1283–1288.

Byrne, J.M., Elliott, S. & Firek, A., 2009. Initial experience with patient-clinician secure messaging at a VA medical center. Journal of the American Medical Informatics Association : JAMIA, 16(2), pp.267–270.

White, C.B. et al., 2004. A content analysis of e-mail communication between patients and their providers: patients get the message. Journal of the American Medical Informatics Association : JAMIA, 11(4), pp. 260–267.

Engaging Patients, Providers, and Institutional Stakeholders in Developing a Patient-centered Microblog

Anuj K Dalal, MD[1,2] Patricia Dykes, RN, DNSc,[1,2] Kelly McNally,[1] Diana Stade,[1] Kumiko Ohashi, PhD, RN,[1] Sarah Collins, PhD, RN,[1,2] David W Bates, MD, MSc,[1,2] Jeff Schnipper, MD, MPH[1,2]

[1]Brigham and Women's Hospital, Boston, MA, [2]Harvard Medical School, Boston, MA

Introduction: Care team communication in acute care settings is fragmented, inefficient, and frustrating.[1, 2] Providers share clinical impressions, opinions, and knowledge informally. Patients clamor for seamless communication with their providers.[3] More efficient management of the informal dialog among patients, inpatient and ambulatory providers during hospitalization is necessary to develop a collaborative and meaningful, patient-centered plan of care. Hospitals require a unified approach to reconciling the informal conversations that occur in text pages, emails, and in some instances, unsecure text messages, to coordinate care more effectively. Web-based tools such as microblogs could facilitate seamless communication and collaboration among patients and providers, but introducing a potentially disruptive technology is not without challenges. Often developers do not engage stakeholders and potential users during the development process – this may limit adoption.[4] The purpose of this study is to describe our experience at engaging patients, providers, and institutional stakeholders in the development of a patient-centered microblog, a single platform where patients and providers can communicate and collaborate.

Methods: This study took place at a large academic medical center in Boston, Massachusetts. We adapted AHRQ's Guide to Patient and Family Engagement in Safety and Quality to serve as a conceptual model to involve patients, providers, and institutional stakeholders in the design and development of our patient-centered microblog (Figure 1). We conducted a total of 10 patient interviews (2 patient advocates, 8 patients), 3 focus groups of 4-6 providers each (physician assistants, interns, residents, attendings, nurses), and 8 interviews of *institutional stakeholders* (2 clinical directors, 2 nursing unit directors, 2 clinical informatics leaders, director of our Center for Patient and Family Relations, and general medicine residency program director) lasting up to 1 hour. The following categories were addressed: potential uses of the microblog, care team member identification/role assignment, clinical workflow, alert fatigue, attitudes and emotions, and medical-legal ramifications. Participants were encouraged to provide feedback regarding the engagement process. We used content analysis methods to interpret descriptive data from focus groups and interviews, and a 2-person consensus approach to identify major themes.

Results: Descriptive feedback and derived themes by type of forum are presented in Table 1. In general, we found a high degree of enthusiasm and support among patients, providers, and institutional stakeholders for using the patient-centered microblog as a platform to get patients and providers "on-the-same-page" with regard to the patient's plan of care but this was tempered by patient, provider, and institutional stakeholder concerns. Specifically, these include providing access to designated caregivers, managing notification settings, integrating into clinical workflow access points, responsibility for managing microblog conversations, providing appropriate context for use, managing patient and provider expectations regarding timing of responses, and addressing medico-legal considerations with regard to patient-provider messaging. Finally, we received positive feedback with regard to our engagement process. All provider participants expressed willingness to participate in our planned pilot (Box 1).

Conclusion: Overall, our experience at engaging patients, providers, and institutional stakeholders during the design and development of a potentially disruptive technology was positive. Although providers and stakeholders were enthusiastic about the idea of using the patient-centered microblog in the acute care setting, they expressed a moderate level of concern with regard to actually using it. Our next steps are to pilot test, iteratively refine, implement, and formally evaluate the patient-centered microblog in the acute care setting.

Acknowledgements: The Brigham and Women's Hospital PROSPECT project is part of the Libretto Consortium supported by the Gordon and Betty Moore Foundation.

References:

1. Coiera E. When conversation is better than computation. J Am Med Inform Assoc. 2000; 7:277-286.
2. Dayton E, Henriksen K. Communication failure: Basic components, contributing factors, and the call for structure. Jt Comm J Qual Patient Saf. 2007; 33:34-47.
3. Dykes PC, Carroll DL, Hurley AC, et al. Building and testing a patient-centric electronic bedside communication center. J Gerontol Nurs. 2013; 39:15-19.
4. Yu P, Gandhidasan S, Miller AA. Different usage of the same oncology information system in two hospitals in Sydney--lessons go beyond the initial introduction. Int J Med Inform. 2010 Jun;79(6):422-9.

Figure 1. Patient-centered microblog. Patient can view the microblog conversation via a tablet device (left). Providers can view the patient-provider and provider-only microblog conversation via a clinical workstation (right).

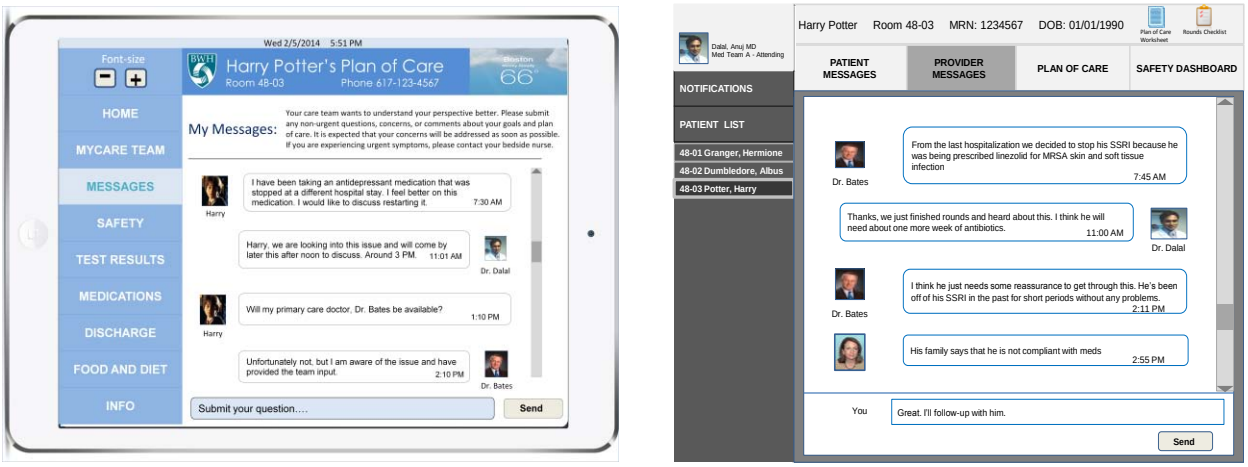


Table 1. Design and Development of a Patient-Centered Microblog: Major Themes from Focus Groups and Individual Interviews by Type of Forum

Type of Forum	Feedback from Focus Groups & Interviews	Themes
Patients[*]	• Useful for asking questions related to the plan of care (goals, problems, schedule) that arise after rounds or which they forget to ask during rounds • Providing access to more "tech-savvy" caregivers, especially for elderly, disabled, less activated patients; engaging caregiver in general • Requires a method of easily identifying name and role of each provider • Useful for scheduling and coordinating patient/caregiver/provider meetings • Expressed apprehension about overburdening physicians with messages	High degree of patient/caregiver enthusiasm for messaging directly with providers outside of rounds and face-to-face visits. For less "tech-savvy" patients, designating family caregiver access would be useful
Providers[†]	• Useful for coordinating care among non-localized, ambulatory and sub-specialty providers • Message "read receipt" and acknowledgment functionality would be useful • Access to tool should not be a separate workflow • Uniformly concerned regarding notification overload → requested user-configurable notifications settings (e.g., email, mobile "app" push notification) • Uniformly concerned that use of instant messaging tools by patients would encourage the expectation of a "real-time" response; expressed concern regarding potential abuse of tool (e.g., overburdening providers regarding "trivial" concerns) • Expressed apprehension regarding notifications from microblog, email, pagers, etc.; concern with regard to missing important patient messages • No one wants to take responsibility for "owning" the microblog conversations • Expressed interest in reviewing patient messages before rounds as it would help guide face-to-face discussions with patients/caregivers (especially for questions that come up after rounds)	High degree of enthusiasm for using the microblog to facilitate transparency of informal patient-specific dialog among providers was tempered with concerns related to notifications, workflow integration, responsibility for managing the microblog, and perception of being required to respond to all patient messages in real-time
Institutional Stakeholders[‡]	• Useful to coordinate care for patient populations at high risk for readmissions • Limit scope to patient-specific, clinical messaging • Clearly distinguishing patient from provider conversation threads → patient-provider messaging should be treated similar to current patient-portal messaging but provider-only conversations should remain informal • Limit context to plan of care and goals of care discussions • Access for family caregivers crucial when patients do not have capacity, but this may be challenging • Responsibility for managing care team role assignment and microblog conversations is murky • Uncertainty with regard to keep it separate from medical record • Managing expectations for patient and provider messaging is needed	High degree of enthusiasm for coordinating care across care settings was tempered by need to manage patient and provider expectations for use, providing appropriate context and scope for use, delineating responsibility for managing microblog conversations, challenge of providing access to family caregivers, and mitigating legal concerns regarding informal and formal microblog conversations

[*]2 patient advocates; 8 patients; [†]physician assistants, interns, residents, attendings, nurses
[‡]2 clinical directors, 2 nursing unit directors, 2 clinical informatics leaders, director of the Center for Patient and Family Relations, and general medicine residency program director

Box 1. Comments from focus group participants

> "We are so appreciative to be included in a proactive manner and to feel heard. Thank YOU for working on quality improvement projects that will hopefully improve our systems and patient care/satisfaction."
>
> "Though our feedback may sound pessimistic, I think we are all in support of the overall goal/big picture and just want to ensure we bring up concerns early so that any that can be addressed/modified will be to help ensure later success."

A smart suite for the evaluation of data generated by a smartphone application for nutritional triage in oncological outpatients

Jeroen S. de Bruin, PhD[1], Christian Schuh, PhD[1], Eva Luger, MSc,[2] Michaela Gall, BSc[3], and Karin Schindler, PhD[3]

[1] Section for Medical Expert and Knowledge-Based Systems, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Vienna, Austria;

[2] Institute for Social Medicine, Center for Public Health, Medical University of Vienna, Vienna, Austria;

[3] Department of Internal Medicine III, Vienna General Hospital, Vienna, Austria;

Abstract

Backgrounds and objectives: *Weight loss in cancer patients is common because of the disease itself and the anti-cancer treatment [1], and is significantly related to various negative effects on patient health and recovery, e.g., worse clinical outcome, poor tolerance of anti-cancer treatment, reduced quality of life and increased mortality [2-4]. Unfortunately, nutritional deterioration and weight loss often go unrecognized, and when recognized are often accepted by patients and caregivers as unmodifiable, while nutrition therapy could be used to improve nutritional status and slow down the process of nutritional deterioration. Our objective was to create a clinical decision support system that uses data generated by a smartphone application for nutritional triage in oncological outpatients to detect nutritional deterioration and enable interventions by physicians and nutritional experts.*

Methods: *To gather feedback on a patient's perceived quality of life and nutritional status, we created an android smartphone application that presents a set of questions to a patient on a daily basis, and an interface in the hospital information system to present those data to the physicians. To estimate the perceived quality of life, we incorporated the European Organization for Research and Treatment of Cancer (EORTC) QLQ-C30 survey, basic model [5]; the scored Patient-Generated Subjective Global Assessment (PG-SGA) was used to detect malnutrition [6]. We used a sample of ten patients from an ongoing clinical study to evaluate the usability and additional medical value of the application. Patients included in this study had to be oncological patients with solid tumors (lung, ear-nose-throat, or gastrointestinal tumors), age 18 or older, receive ambulatory anti-cancer treatment as outpatient for at least a month, and had to have access to a smartphone and wireless internet connection, as well as weighing scales. Patients with a life-expectancy of less than six months were excluded.*

Results: *Questions from both questionnaires were combined into a single questionnaire comprising 34 questions. These questions were divided by theme into eleven categories i.e., Weight, Food 1-2, Function 1-3, Daily life 1-3, Quality of Life and Medical visits. The collected data were forwarded to the hospital information system where they were available to physicians; The data displayed included the native scoring mechanisms of QLQ-C30 and PG-SGA were used to indicate a patient's health status or performance in each of these categories. Based on the combined PG-SGA and QLQ-C30 scores, the physicians are able to make more accurate nutritional triage recommendations. Furthermore, the usability study showed that 90% of all patients found the application easy to use and handle, and were of the opinion that it had a good additional medical value; 40% felt it influenced their daily lives and awareness.*

Discussion: *We presented a smartphone application and the integration of generated data in medical routine for oncological outpatients to keep an accurate record of their nutritional status and perceived quality of life. At present, the smartphone application is used in a clinical trial involving 30 patients, and the routine integration is still being extended as more data comes in. Patients indicated that they were generally satisfied with the smartphone application, and using smartphones as a method to monitor their health. Currently we are also working with nutritional experts to develop a rule-based system for the generation of nutritional alerts and interventions, and to add an epidemiological module in order to analyze the entire patient population.*

[1] L. Mariani, S. Lo Vullo, and F. Bozzetti, Weight loss in cancer patients: a plea for a better awareness of the issue, *Supportive care in cancer : official journal of the Multinational Association of Supportive Care in Cancer 20(2)* (2012) 301-9.

[2] J. Arends, G. Bodoky, F. Bozzetti, K. Fearon, M. Muscaritoli, G. Selga, et al., ESPEN Guidelines on Enteral Nutrition: Non-surgical oncology, *Clin Nutr 25(2)* (2006) 245-59.
[3] K. Fearon, F. Strasser, S.D. Anker, I. Bosaeus, E. Bruera, R.L. Fainsinger, et al., Definition and classification of cancer cachexia: an international consensus, *Lancet Oncol 12(5)* (2011) 489-95.
[4] M.B. Huhmann, and R.S. Cunningham, Importance of nutritional screening in treatment of cancer-related weight loss, *Lancet Oncol 6(5)* (2005) 334-43.
[5] N.K. Aaronson, S. Ahmedzai, B. Bergman, M. Bullinger, A. Cull, N.J. Duez, et al., The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology, *Journal of the National Cancer Institute 85(5)* (1993) 365-76.
[6] J. Bauer, S. Capra, and M. Ferguson, Use of the scored Patient-Generated Subjective Global Assessment (PG-SGA) as a nutrition assessment tool in patients with cancer, *European journal of clinical nutrition 56(8)* (2002) 779-85.

HARVEST, a Holistic Patient Record Summarizer at the Point of Care

Noémie Elhadad, PhD[1], Sharon Lipsky Gorman, MA[1], Jamie S. Hirsch, MD, MS[1], Connie Liu, MS[1], David K. Vawdrey, PhD[1], Marc Sturm[2]
[1]Columbia University, New York, NY; [2]NewYork-Presbyterian Hospital, New York, NY

Introduction

In many care settings, clinicians are faced with an overwhelming amount of complex information about their patients, with little time for chart review [1]. Failure to digest patient data may result in errors in diagnosis and delayed care [2,3]. The need for better health information management and visualization tools has long been recognized [4,5], yet current electronic health records (EHRs) do not yet provide the cognitive support necessary for effectively and efficiently reviewing patient data. EHR displays are plagued by low information content, do not honor established interface design principles, and cannot be readily customized without imposing a considerable burden on clinicians and information technology professionals [6,7]. Important research has been undertaken on visualization of patient histories [8–10] and domain-specific summarizers [11], motivating our hypothesis that a holistic patient-record summarizer can impact care in a beneficial fashion. In this abstract, we describe the design, implementation, and deployment of HARVEST, a longitudinal patient record summarization system.

Methods

HARVEST is an interactive, problem-oriented patient record summarization system (see Figure below) [12]. The summarizer differs from previous work in three critical ways: (i) it extracts content from the patient notes, where key clinical information resides; (ii) it aggregates and presents information from multiple care settings, including inpatient, ambulatory, and emergency department encounters; and (iii) it is integrated into two commercial EHR systems, and is available for all patients in our institution, not just a curated dataset or for specific patient cohorts.

The natural language processing (NLP) of clinical notes is carried out through a named-entity recognition system that indexes concept mentions. Because HARVEST aggregates problems, NLP was constrained to extracting concepts from the UMLS semantic group "Disorder" only, restricted to the SNOMED-CT Core Problem List. Concept salience weights are computed dynamically to reflect both the frequency of the concept in the patient notes in a given time slice of the record, and the prevalence of the concept across all patients in the institution. To enable parsing and salience computation at scale, we created a distributed computing infrastructure (using Apache Hadoop) and implemented a map-reduce version of our NLP system to parse the notes from a variety of HL7 interface feeds.

Results

The Hadoop infrastructure enabled us to accommodate the large volume of clinical documents generated in our institution (650,000 notes per month): a small four-node cluster processed 20,000 notes/second compared to 500 notes/second in a non-distributed computing environment. The infrastructure also permitted us to experiment with different parsing and salience computation strategies through short development cycles.

HARVEST was deployed within NewYork-Presbyterian's iNYP clinical information review system as a beta release in September 2013. As a first phase of deployment and to study the impact of HARVEST on clinical care, access was limited to internal medicine residents and residents and attending physicians in the emergency department (ED). Three primary use cases for the summarizer have emerged thus far: (i) in the ED, as a way to capture the essential knowledge about a patient's history in an efficient fashion, including previous emergency visits and patterns of visits; (ii) in internal medicine, as a support tool for performing an in-depth patient chart review before admitting a patient to the hospital; and (iii) in the primary care clinic, for chief residents as an education tool during precepting hours with residents.

Discussion

HARVEST addresses an unmet need for clinicians at the point of care, facilitating effective and efficient review of essential patient information. The deployment of HARVEST in our institution allows us to study patient record summarization as an informatics intervention in a real-world setting. It also provides an opportunity to learn how clinicians use the summarizer, enabling informed interface and content iteration and optimization to improve patient care. Future work includes customizing the content selected by HARVEST for different types of clinicians.

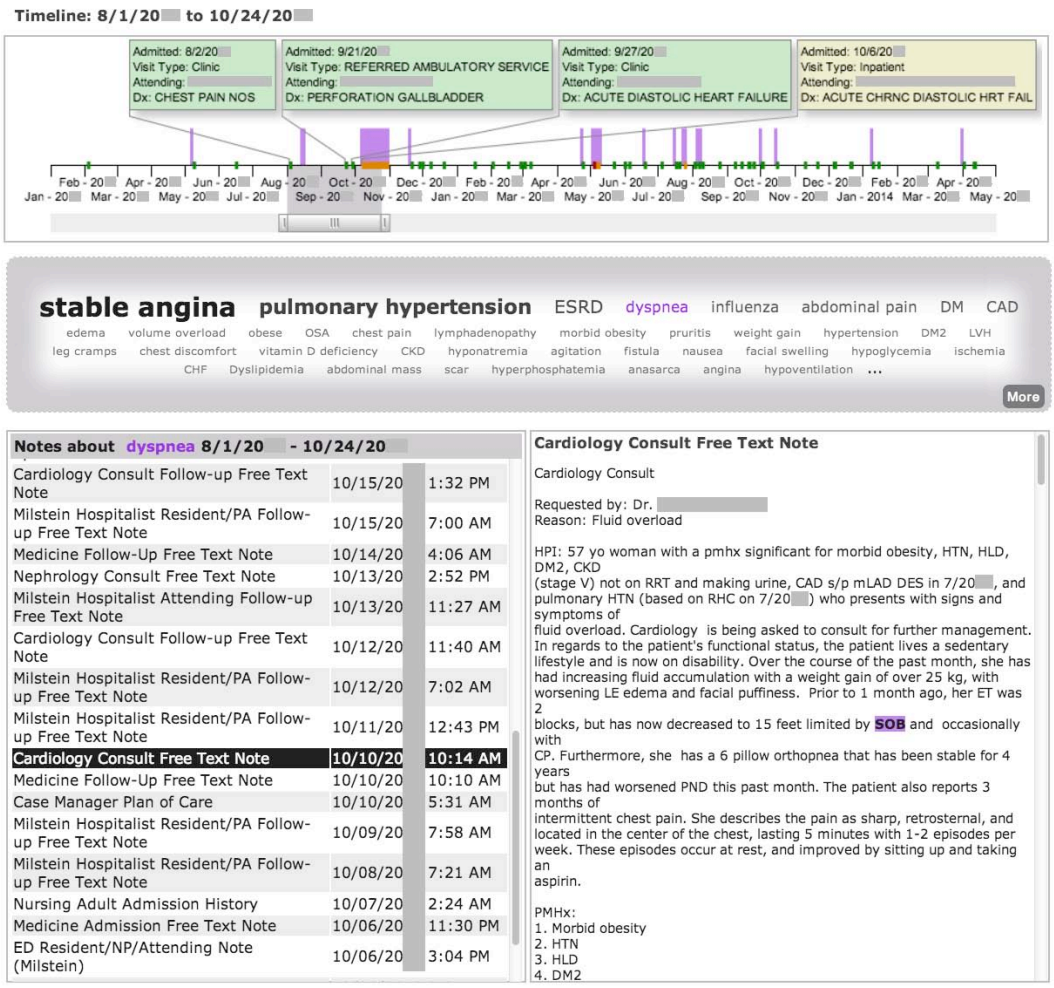


Figure 1. De-identified HARVEST screenshot for a sample patient, part of iNYP. For the selected time range (3 months), stable angina, pulmonary hypertension, end-stage renal disease, and dyspnea are the most prominently documented problems. HARVEST also identified diabetes mellitus, hypertension, and dyslipidemia as important problems. The problem "dyspnea" was selected. The Notes panel lists all notes in the selected time range that mention this problem. A cardiology consult note is selected and displayed in the lower right panel, with all mentions of dyspnea (and synonyms) highlighted. On the timeline, documentation of dyspnea is highlighted by purple bars, indicating that dyspnea was a particularly salient issue at that time, as well as 6 months later.

References

1 Reichert D, Kaufman D, Bloxham B, *et al.* Cognitive analysis of the summarization of longitudinal patient records. In: *Proceedings of the Annual American Medical Informatics Association Fall Symposium (AMIA)*. 2010. 667–71.

2 Christensen T, Grimsmo A. Instant availability of patient records, but diminished availability of patient information: a multi-method study of GP's use of electronic patient records. *BMC Med Inform Decis Mak* 2008;8:12.

3 Singh H, Spitzmueller C, Petersen NJ, *et al.* Information overload and missed test results in electronic health record–based settings. *JAMA Intern Med* 2013;173:702–4.

4 Powsner SM, Tufte ER. Graphical summary of patient status. *Lancet* 1994;344:386–9.

5 Payne TH. Computer decision support systems. *Chest* 2000;118:47S–52S.

6 Blumenthal D, Tavenner M. The 'meaningful use' regulation for electronic health records. *N Engl J Med* 2010;363:501–4.

7 Laxmisan A, McCoy AB, Wright A, *et al.* Clinical summarization capabilities of commercially-available and internally-developed electronic health records. *Appl Clin Inform* 2012;3:80–93.

8 Plaisant C, Mushlin R, Snyder A, *et al.* LifeLines: using visualization to enhance navigation and analysis of patient records. In: *Proceedings of the Annual American Medical Informatics Association Fall Symposium (AMIA)*. 1998. 76–80.

9 Shahar Y, Boaz D, Tahan G, *et al.* Interactive Visualization and Exploration of Time-oriented Clinical Data Using a Distributed Temporal-Abstraction Architecture. In: *Proceedings of the Annual American Medical Informatics Association Fall Symposium (AMIA)*. 2003. 1004.

10 Tao C, Wongsuphasawat K, Clark K, *et al.* Towards event sequence representation, reasoning and visualization for EHR data. ACM Press 2012. 801.

11 Wilcox A, Jones SS, Dorr DA, *et al.* Use and Impact of a Computer-Generated Patient Summary Worksheet for Primary Care. In: *Proceedings of the Annual American Medical Informatics Association Fall Symposium (AMIA)*. 2005. 824–8.

12 Weed LL. Medical records that guide and teach. *N Engl J Med* 1968;278:652–657 concl.

Designing Specific Alerts for Potassium-Increasing Drug-Drug Interactions

Emmanuel Eschmann, MD, Patrick E. Beeler, MD, Jürg Blaser, PhD
Research Center for Medical Informatics, University Hospital, Zurich, Switzerland

Abstract: *Electronic warnings against hyperkalemia during potassium-increasing drug-drug interactions (DDIs) are often overridden due to their low specificity. The treatments of 76,467 inpatients were analyzed to design more specific alerts. Unnecessary alerts may be suppressed at onset of DDI by considering the most recent serum potassium level or additional patient parameters. However, short-term forecasts based on the periodically monitored serum potassium level achieve the highest specificity compared to an initial prediction for the entire DDI period.*

Introduction: The warnings against hyperkalemia (serum potassium [K^+] ≥5.5mEq/l) triggered during K^+-increasing drug-drug interactions (DDIs) are often ignored due to their low specificity and might therefore induce alert fatigue. This retrospective analysis aimed at the development of highly specific alerts. The impact of the novel alerting concept will be assessed in a prospective clinical trial.

Methods: The data of all inpatients admitted to the University Hospital Zurich during a 25 months period were included, except for ICU stays. The K^+-increasing DDIs were identified using the knowledge base hospINDEX (distributed by e-mediat AG, Switzerland). Only the DDIs with the highest severity levels "contraindicated" and "monitoring or adaptation required" were considered. Generalized additive models were used to identify important patient parameters. The efficacy of the models to detect those DDIs likely to induce hyperkalemia was retrospectively validated. The alert thresholds were defined by applying Youden's J statistic on ROC curves, using 50% of the data as the calibration set and 50% as the validation set. ROC curves were compared with Delong's test.

Results: We analyzed 1.5 million drug orders of 76,467 patients and identified 8,431 K^+-increasing DDIs (mean duration 6.4 days) which induced 151 hyperkalemias. Tab. 1 compares common (A_{none}, A_{K+}) and novel alerts (A_{opt}, B_{K+}, B_{opt}). The A-alert rules (A_{none}, A_{K+}, A_{opt}) were applied at onset of each K^+-increasing DDI: A_{none} considered no patient parameters, A_{K+} took into account the serum K^+ level at onset of DDI, and A_{opt} included 8 additional patient parameters. In contrast, the B-alert rules (B_{K+}, B_{opt}) were applied not only at onset of the DDI, but again for each serum K^+ level obtained during the DDI: B_{K+} considered merely the current serum K^+ level, whereas B_{opt} additionally included 10 patient parameters.

Table 1. Comparison of 5 alerting concepts.

Triggering event	Warnings against	Label	Parameters considered	Sensitivity	Specificity
start of DDI	risk for hyperkalemia during the entire period of DDI	A_{none}	none	100.0%	0.0%
		A_{K+}	last serum K^+ before onset of DDI ≥4.2mEq/l	54.8%	68.6%
		A_{opt}	9 patient parameters	59.4%	71.6%
start of DDI and each serum K^+ measurement during DDI	risk for hyperkalemia within the next 48h	B_{K+}	current serum K^+ ≥4.4mEq/l	70.5%	76.5%
		B_{opt}	11 patient parameters	75.6%	72.4%

Compared to the systematic warnings at onset of each K^+-increasing DDI (A_{none}), the concept A_{K+} increased the specificity but decreased the sensitivity by restricting the alerts to K^+ levels ≥4.2 mEq/l. The proposed novel alerts A_{opt}, B_{K+} and B_{opt} provided enhanced sensitivity and specificity compared to A_{K+}: Including 8 patient parameters in addition to the most recent serum K^+ level significantly improved the alert effectiveness (A_{opt} vs. A_{K+}; Delong's test: $p=0.024$). Focusing on short-term predictions of hyperkalemia occurring within the next 48h further increased both sensitivity and specificity (B_{K+} vs. any A-alert; $p<0.05$). However, including additional patient parameters did not significantly improve sensitivity and specifity (B_{opt} vs. B_{K+}; $p=0.056$).

Discussion: The effectiveness of alerts could be improved stepwise, (i) by considering the most recent serum K^+ level at onset of DDI, (ii) by including additional patient parameters, and (iii) by considering 48h forecasts instead of predictions for the entire DDI-period. Thus, focusing on short-term predictions of the risk for hyperkalemia based on serum K^+ monitoring improves the alert specificity. An ongoing randomized clinical trial will assess whether this novel concept reduces alert fatigue and increases patient safety.

Automated Detection of Early Physiological Deterioration in Hospitalized Patients

R. Scott Evans MS, PhD, FACMI[1,2], Kathryn G. Kuttler PhD[1,3], Kathy J. Simpson RN, BSN[4], Stephen Howe MA, MS[1], Peter F. Crossno MD, FACP[3], Kyle V. Johnson BS[1], Misty N. Schreiner RN, BSN, CCRN[4], James F. Lloyd BS[1], William H. Tettelbach MD, FACP[5,6], Roger K. Keddington APRN MSN CEN[7], Alden Tanner RN, BSN, CCRN[4], Chelbi Wilde RN, BSN[4], Terry P. Clemmer MD[8,9]

[1]Homer Warner Center for Informatics Research, Intermountain Healthcare, Salt Lake City, UT; [2]Biomedical Informatics, University of Utah School of Medicine Salt Lake City, UT; [3]Pulmonary and Critical Care, Intermountain Medical Center, Murray, UT [4]Shock Trauma Intensive Care, Intermountain Medical Center, Murray, UT; [5]Hyperbaric Medicine, Wound Care & Infectious Diseases, Intermountain Healthcare, Salt Lake City, UT; [6] Department of Anesthesiology, Duke University School of Medicine, Durham, NC; [7]Intensive Medicine/Emergency Services, Intermountain Healthcare, Salt Lake City, UT; [8]Critical Care Medicine, LDS Hospital, Salt Lake City, UT; [9]Department of Medicine, University of Utah School of Medicine, Salt Lake City, UT

Introduction: The average age of the US population is increasing as is the complexity of their medical care. As a result, up to 5% of patients experience physiologic deterioration (PD) during their hospital stay resulting in admission to the intensive care unit (ICU) or death[1, 2]. Studies reveal that many of these adverse events are preceded by indicators of PD[3-11]. Medical emergency teams (MET) were developed to prevent patient crisis before a cardiopulmonary arrest [12]. However, delayed MET calls are common and patients who are attended to within 30-60 minutes of PD have significantly lower mortality rates [8, 13, 14]. Therefore, to be effective, MET must have an afferent limb (case detection and response triggering) in addition to an efferent limb (medical response)[15].

Methods: We created a MET Risk committee comprised of critical care nurses from the MET and its nursing and medical directors, intensive care physicians, an infectious disease physician and medical informaticists. A decision support tool was developed which monitored hospitalized patients every 5 minutes and used vital sign and neurologic data in our electronic medical record (EMR) to identify patients with early PD. Once the positive predictive value (PPV) of our PD alert model was acceptable, we went live on a 33-bed medical and oncology floor (A) and a 33-bed non-ICU surgical trauma floor (B). During the intervention year, pager alerts of early PD were sent automatically to charge nurses along with access to a graphical point-of-care web page to help facilitate patient evaluation. Nurses were requested to fill out a form describing the validity of and their response to the PD alerts.

Results: Patients on unit A were significantly older and had significantly more comorbidities than unit B. During the intervention year, unit A patients had a significant increase in length of stay and total hospital cost, an increase in ICU transfers ICU 163 (5.1%) of 3189 compared to 146 (4.3%) of 3423 ($p = 0.1163$), significantly more MET calls (60 vs 29, $p = 0.0004$) and significantly fewer patients died (84 (2.6%) vs 125 (3.7%), $p = 0.022$) compared to the pre-intervention year. No significant differences in outcome were found on unit B and no differences between pre-intervention and intervention patient populations were found in either unit. Nurses called a physician based on 51% and 44% of the PD alerts in units A and B and interventions were initiated based on 59% and 52% of the PD alerts.

Discussion: There is variability in the ability of healthcare to detect patients who experience PD[16]. We report the results of a four year effort to use our EMR to develop, implement and evaluate an automated case detection and response triggering system for PD that meets nursing workflow and endorsement. Nurses on both study units reported an appreciated difference in their workflow based on the early identification of patients with PD. This study found use of computerized decision support provided a way to constantly monitor patients and notify nursing of early PD which resulted in a significant increase in appropriate MET calls and a significant decrease in mortality in the nursing unit containing older patients with multiple comorbidities. Moreover, with the patient trending information contained in the graphical alerts, nurses reported they had the information they needed to evaluate the patient status, felt more confident about their assessment, and were more at ease about requesting additional help.

1. Bell MB, Konrad D, Granath F, Ekbom A, Martling CR. Prevalence and sensitivity of MET-criteria in a Scandinavian University Hospital. Resuscitation. 2006 Jul;70(1):66-73. PubMed PMID: 16757089.
2. McFarlan SJ, Hensley S. Implementation and outcomes of a rapid response team. Journal of nursing care quality. 2007 Oct-Dec;22(4):307-13, quiz 14-5. PubMed PMID: 17873726.
3. Goldhill DR, White SA, Sumner A. Physiological values and procedures in the 24 h before ICU admission from the ward. Anaesthesia. 1999 Jun;54(6):529-34. PubMed PMID: 10403864.
4. Hillman KM, Bristow PJ, Chey T, Daffurn K, Jacques T, Norman SL, et al. Antecedents to hospital deaths. Internal medicine journal. 2001 Aug;31(6):343-8. PubMed PMID: 11529588.
5. Hillman KM, Bristow PJ, Chey T, Daffurn K, Jacques T, Norman SL, et al. Duration of life-threatening antecedents prior to intensive care admission. Intensive care medicine. 2002 Nov;28(11):1629-34. PubMed PMID: 12415452.
6. Schein RM, Hazday N, Pena M, Ruben BH, Sprung CL. Clinical antecedents to in-hospital cardiopulmonary arrest. Chest. 1990 Dec;98(6):1388-92. PubMed PMID: 2245680.
7. Buist MD, Jarmolowski E, Burton PR, Bernard SA, Waxman BP, Anderson J. Recognising clinical instability in hospital patients before cardiac arrest or unplanned admission to intensive care. A pilot study in a tertiary-care hospital. The Medical journal of Australia. 1999 Jul 5;171(1):22-5. PubMed PMID: 10451667.
8. Buist M. The rapid response team paradox: why doesn't anyone call for help? Critical care medicine. 2008 Feb;36(2):634-6. PubMed PMID: 18216622.
9. Kause J, Smith G, Prytherch D, Parr M, Flabouris A, Hillman K, et al. A comparison of antecedents to cardiac arrests, deaths and emergency intensive care admissions in Australia and New Zealand, and the United Kingdom--the ACADEMIA study. Resuscitation. 2004 Sep;62(3):275-82. PubMed PMID: 15325446.
10. Cioffi J. Recognition of patients who require emergency assistance: a descriptive study. Heart & lung : the journal of critical care. 2000 Jul-Aug;29(4):262-8. PubMed PMID: 10900063.
11. Ott LK, Pinsky MR, Hoffman LA, Clarke SP, Clark S, Ren D, et al. Medical emergency team calls in the radiology department: patient characteristics and outcomes. BMJ quality & safety. 2012 Jun;21(6):509-18. PubMed PMID: 22389020. Pubmed Central PMCID: 3630458.
12. Bruckel J. Evidence-based medicine and rapid response team implementation. McGill journal of medicine : MJM : an international forum for the advancement of medical sciences by students. 2006 Jan;9(1):5-7. PubMed PMID: 19529801. Pubmed Central PMCID: 2687898.
13. Kumar A, Roberts D, Wood KE, Light B, Parrillo JE, Sharma S, et al. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. Critical care medicine. 2006 Jun;34(6):1589-96. PubMed PMID: 16625125.
14. Chan PS, Krumholz HM, Nichol G, Nallamothu BK, American Heart Association National Registry of Cardiopulmonary Resuscitation I. Delayed time to defibrillation after in-hospital cardiac arrest. The New England journal of medicine. 2008 Jan 3;358(1):9-17. PubMed PMID: 18172170.
15. Devita MA, Bellomo R, Hillman K, Kellum J, Rotondi A, Teres D, et al. Findings of the first consensus conference on medical emergency teams. Critical care medicine. 2006 Sep;34(9):2463-78. PubMed PMID: 16878033.
16. Galhotra S, DeVita MA, Simmons RL, Schmid A, members of the Medical Emergency Response Improvement Team C. Impact of patient monitoring on the diurnal pattern of medical emergency team activation. Critical care medicine. 2006 Jun;34(6):1700-6. PubMed PMID: 16625132.

RexMart: An Open Source Tool for Exploring and Sharing Research Data without Compromising Data Integrity

Frank J. Farach, PhD[1], Owen McGettrick[1], Charles Tirrell[1], Clark Evans[1], Alejandro Mesa[1], Leon Rozenblit, JD, PhD[1]
[1]Prometheus Research LLC, New Haven, CT

Abstract

Behavioral and biomedical researchers value data integrity, data exploration, and data sharing, but traditional research databases have unacceptable trade-offs among these important data management functions. This case study explores the design, development, and user-acceptability of RexMart, an open-source application designed to maximize data exploration and sharing without compromising data integrity.

Introduction

Clinical research places great demands on data integrity, exploration, and sharing. The stakes are high: Low-integrity data cannot be trusted to yield valid inferences; data that cannot easily be explored are difficult to analyze; and data that are difficult to share hinder collaboration and reanalysis (e.g., meta-analyses). In our prior qualitative research, researchers rated these issues among the most important to address in data management.[1] Unfortunately, many data management solutions optimize for one of these dimensions at the expense of others. Researchers often manage data using spreadsheets or statistical software because their denormalized structure makes certain kinds of data exploration and sharing convenient, but this approach jeopardizes data integrity, data-analytic flexibility[2], and scalability. Conversely, normalizing data optimizes data integrity but constrains the design and performance of query interfaces.[3] This case study explores the design, development, and evaluation of RexMart, an open-source application designed to optimize data integrity, exploration, and sharing.

Methods

We interviewed system designers and examined the informatics literature for design strategies both within and outside research information systems. Several architectures were proposed and compared against internal and user requirements. Once an architecture was selected, we developed the first major release of RexMart and conducted acceptance testing with four users at a major research university. The open-source code can be found at https://bitbucket.org/rexdb/rex.explore_ui-provisional and https://bitbucket.org/rexdb/rex.explore-provisional.

Results

Architectural analysis revealed a promising solution that, to our knowledge, has not been implemented elsewhere for research data. Specifically, a well-normalized relational data warehouse stores research and study management data, providing robust control over data integrity. Users access data via a data mart, a point-in-time snapshot of data from the warehouse that has been reorganized into a dimensional relational data model. The schema of the resulting data mart is configurable by analysts so that each on-demand data mart can be re-normalized or denormalized to fit analytic needs. The first major release of this open-source application allowed users to create data marts; search for variables by column and table metadata; produce summary statistics; filter columns by value; build basic queries through the UI; and run, save, and share queries. In acceptance testing, although users wanted to be able to share databases with other users, they liked the metadata search, summary statistics, and saved query functionality.

Discussion

Data integrity, exploration, and sharing are critical to scientific progress but are difficult to maximize within the same database. By storing research and study management data in a data warehouse and providing a user-friendly way to create data marts from the warehouse for exploration, we were able to deliver a major release of RexMart that provided a satisfactory data exploration environment without compromising data integrity. Future work will focus on implementing database sharing, improving data analysis features, and providing native interoperability with data formats and models that are commonly used in clinical research, such as CDISC ODM, HL7 CDA, and dbGaP.

References

1. Johnson SB, Farach FJ, Pelphrey K, Rozenblit L. *Assessing the data management needs of clinical researchers* (in preparation, 2014).
2. Wickham H. Tidy data. *http://vita.had.co.nz/papers/tidy-data.pdf (accessed August 1, 2014).*
3. Simsion GC, Witt GC. Essentials of data modeling, 3rd ed. San Francisco: Morgan Kaufmann; 2005.

First Feasibility of a Surveillance Platform Combining Community-Submitted Symptoms and Specimens for Molecular Diagnostic Testing

Jennifer Goff[1,2], Aaron Rowe, PhD[3], Rumi Chunara, PhD[2,4]

[1]Boston University School of Public Health, Boston, MA; [2]Boston Children's Hospital, Boston, MA; [3]Integrated Plasmonics Corporation, San Francisco, CA ;[4]Harvard Medical School, Boston, MA

Introduction

Modern informatics-based participatory surveillance systems are one approach to obtaining disease information in real-time and at scale. Internet-based Influenza surveillance systems have been developed and deployed nationally for years in 10 countries in Europe as well as Australia and here in the United States through Flu Near You[1, 2, 3]. However, none of these Internet-based systems employ molecular techniques to rapidly verify the intelligence data they garner. By combining Internet-based participatory surveillance with advanced molecular methods for high sensitivity and specificity viral identification, the aims of this project are to not only serve to advance epidemiological knowledge of respiratory viral infections (such as symptom profiles, spatio-temporal distribution and risk groups) and validate existing approaches that only incorporate self-reported symptom information, but also to demonstrate a new type of informatics platform that combines crowdsourced diagnostic samples and symptom reports, enabling individuals to understand what infection they may have as well as what is circulating in their community.

Methods

We began by building a kit that includes an off-the-shelf saliva collection kit and nasal swab with viral collection medium. These kits were paired with customized instructions and instructional videos. A cohort of lay volunteers in the Boston area was recruited through a variety of modern channels. Enrollment was driven largely by announcements on local radio stations and word of mouth. Paid online advertisements and social media were also considered as means of recruiting volunteers to the study. These volunteers reported symptomatic information to the Flu Near You website and then performed self-testing and specimen collection upon becoming symptomatic with influenza like illness. Our instructions called for performing the tests within the first 48 hours of cold or flu symptoms. Samples were tested by multiplex PCR.

Results

Adherence to the study protocol by subjects was good with some indicators that there is room for improvement in our instructional materials and kit contents. Between December 1 2013 and March 1 2014, 295 participants enrolled in the study and received a kit. Participants submitted kits an average of 2.30 days (95 CI: 1.65 to 2.96) after symptoms began (Figure 1). Of those reporting symptoms, 55 participants completed the diagnostic samples by March 1 and specimens were sent for nucleic acid analysis. 6 out of 30 specimens tested to-date were positive for Influenza (Influenza A, 2009 H1N1), 13 tested positive for at least 1 coronavirus and 1 for respiratory syncytial virus (RSV). Nasal and saliva samples gave the same result for all samples except one influenza sample (no positive detection in saliva).

Conclusions

These preliminary findings present multiple initial conclusions: diagnostic samples can be generated by the public, and oral specimens are comparable with nasal specimens when used for the identification of upper respiratory pathogens, however further larger studies will increase the power of these conclusions. Volunteer recruitment, and ensuring compliance of volunteers with study protocols, are among the main challenges of this work. Revisions to the written & video instructions distributed with the specimen collection systems should yield higher levels of volunteer adherence to protocols. As well, targeted recruiting through Public Health Associations (e.g. Caucuses in the American Public Health Association) can assist in the scaling and representativeness of the participants. Additionally, all specimens collected during this study have been banked and different laboratory diagnostic platforms and DNA extraction and PCR techniques may be experimented with. Future studies will increase the amount of reported data, which will enable evaluation of the relationship between reported symptom profiles and viral etiology as well as spatio-temporal distribution of upper respiratory infection, and impact of these efforts on individual's public health and healthcare seeking behavior.

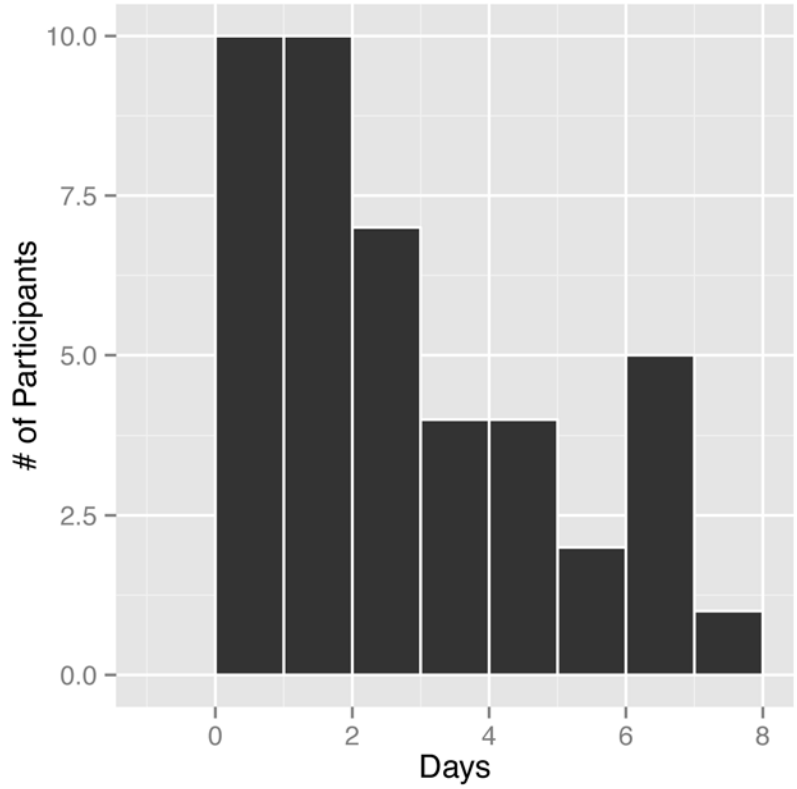


Figure 1. Days between Submission and Illness Date

References

1. Paolotti D, Carnahan A, Colizza V, Eames K, Edmunds J, Gomes G, et al. Web-based participatory surveillance of infectious diseases: the Influenzanet participatory surveillance experience. Clin Microbiol Infect 2014;20(1):17-21.
2. Dalton C, Durrheim D, Fejsa J, Francis L, Carlson S, d'Espaignet ET, et al. Flutracking: a weekly Australian community online survey of influenza-like illness in 2006, 2007 and 2008. Commun Dis Intell Q Rep 2009;33(3):316.
3. Chunara R, Aman S, Smolinski M, Brownstein JS. Flu near you: an online self-reported influenza surveillance system in the USA. Online J Public Health Inform 2013;5(1).

Funding

This work is being funded by the National Science Foundation award number IIS - 1343968.

Translating EHR-Based Diabetes Decision Support Tools Into the Safety Net

Rachel Gold, PhD, MPH[1,2]; Jon Puro, MPA[2]; Jennifer DeVoe, MD, DPhil[2,3]; Christine Nelson, PhD, RN[2,3]; Arwen Bunce, MA[1]; Celine Hollombe, MPH[1]; James Davis, BA[1]; Stuart Cowburn, MPH[2]; John Muench, MD[3,4]; Christian Hill, MD, MPH[5]; Meena Mital, MD[6]; Ann Turner, MD[5]

[1]Kaiser Permanente NW Center for Health Research, Portland, OR; [2]OCHIN, Inc., Portland, OR; [3]OHSU, Portland, OR; [4]OHSU Richmond Clinic, Portland, OR; [5]Virginia Garcia Memorial Health Centers, Hillsboro, OR; [6]Multnomah County Health Department, Portland, OR

Learning Objectives: Participants will learn about strategies for adapting the ALL Initiative for CHCs, how to anticipate barriers and facilitators for successful implementation, and the potential impact of such implementation.

Abstract. *Clinical decision support systems (CDSS) that harness electronic health record (EHR) data have improved diabetes care quality in integrated health care systems such as Kaiser Permanente (KP).[1,2] Little is known about how such tools can be implemented in community health centers (CHCs), though CHCs could gain substantially from adapting effective CDSS tools to their environment. Our study is the first clinical trial to study the implementation in CHCs of an EHR-based CDSS strategy developed in an integrated care setting. The diabetes quality improvement intervention – the 'ALL Initiative' – was highly effective at KP, and uses EHR-based CDSS tools (e.g. alerts, panel management data rosters, order sets). We adapted this intervention for implementation in 11 CHCs and studied its impact on rates of guideline-based cardioprotective medication prescribing among diabetic patients. Our presentation will describe: (1) How the EHR-based CDSS tools were adapted for use in CHCs; (2) How and why the adapted tools were or were not used by CHC staff, and lessons that could inform future efforts to adapt EHR-based CDSS tools for use in CHCs; and (3) Impact of the implemented CDSS tools in the study CHCs.*

Background: KP's diabetes quality improvement intervention, the 'ALL Initiative,' uses diverse EHR-based CDSS tools (automated point-of-care alerts, patient data management tools, streamlined order sets, etc.) to improve rates of guideline-based cardioprotective prescribing. We sought to assess whether and how this intervention, which was very effective in the KP setting,[1,2] could be adapted for implementation in CHCs, whose patient populations, EHR resources, workflows, and needs differ greatly from those in integrated care settings.

Methods: Eleven CHCs in the Portland, OR area participated in our 'translational' randomized trial. We worked with clinic staff and EHR designers from the OCHIN community health information network to adapt KP's intervention for CHCs. Briefly, in place of KP's panel management-based tools and top-down implementation approach, we created a 'menu' of stand-alone CDSS tools and relied heavily on practice facilitators. Next we implemented the adapted CDSS tools in the study clinics through a staggered randomized process, where six 'early' clinics implemented one year before the remaining five 'late' clinics. We used segmented regression analyses to estimate the change in prescribing rates post-implementation; the model also tested for differences in trends over time. We collected quantitative EHR data to assess the impact of the EHR-based CDSS tools on rates of guideline-based prescribing, and qualitative data to identify barriers and facilitators to use of the tools.

Results: We substantially adapted KP's EHR tools for use in the study CHCs, but were able to provide functions similar to those available at KP. Barriers to integrating the tools into the CHCs included: providers' mistrust of automated alerts and habit of ignoring them; challenges inherent to fitting CDSS tools into existing clinic workflows; that the tools address one aspect of diabetes care, rather than all aspects of patient care; developing population-level tools that allow providers leeway to make decisions for individual patients; and other factors related to using EHR tools to support practice change. Despite these barriers, implementing the ALL CDSS tools was associated with significant improvements in rates of guideline-based prescribing: the percent of 'early' CHC patients appropriately prescribed a statin increased from 59% to 68% in 12 months (a 15% relative increase); the percent appropriately prescribed ACE-inhibitors, from 69% to 76% (a 10% relative increase). The 'late' CHCs showed no concurrent change until their implementation began one year later, but one year post-implementation, the late CHCs' guideline-concordant prescribing rates had improved 57% to 68% (statins), and 64% to 74% (ACE-inhibitors).

Discussion: It is feasible to adapt EHR-based CDSS tools developed in the private sector for implementation in CHCs. Our results illustrate the impact of KP's ALL Initiative when implemented in CHCs, and the challenges involved in using EHR-based tools to support practice change and quality improvement efforts in CHCs.

Funding source: National Heart, Lung & Blood Institute, 1R18HL095481-01A1

References

1. Zhou YY, Unitan R, Wang JJ et al. Improving population care with an integrated electronic panel support tool. *Popul Health Manag.* 2011;14:3-9.

2. Feldstein AC, Perrin NA, Unitan R et al. Effect of a patient panel-support tool on care delivery. *Am J Manag Care.* 2010;16:e256-e266.

Automating Performance Measures and Clinical Practice Guidelines: Differences and Complementarities

Mary K. Goldstein MD MS,[1, 2] Samson W. Tu MS,[2] Susana Martins MD MSc,[1] Connie Oshiro,[1] Kaeli Yuen,[1] Tammy Hwang,[1] Dan Wang PhD,[1] Amy Furman PharmD,[1] Michael Ashcraft MD,[1] Paul A. Heidenreich MD MS[1,2]

[1]VA Palo Alto Health Care System, Palo Alto, CA; [2]Stanford University, Stanford, CA

Abstract

Through close analysis of two pairs of systems that implement the automated evaluation of performance measures (PMs) and clinical decision support (CDS), we contrast differences in their knowledge encoding and necessary changes to a CDS system that provides management recommendations for patients failing performance measures. We trace the sources of differences to the implementation environments and goals of PMs and CDS.

Introduction

PMs and CDS are methods to improve quality of care. PMs focus on measuring the quality of care that patients have received, while CDS focuses on providing information to assist health professionals with clinical management. We seek to characterize concretely the ways automated support for PMs and guideline-based CDS differ yet can be complementary to each other. We closely examine two pairs of PM and CDS implementations that use clinical data from the Department of Veterans Affairs (VA) VISTA system to generate performance reports or patient-specific CDS recommendations. We analyze how PMs and CDS differ or complement each other in their cohort definitions, knowledge modeling, workflow integration, use of data, and output structure.

Methods

1. We first compared a pair of systems, one designed for PM and the other designed for CDS, as follows: (a) Heart Failure (HF)-PM, an implementation of National Quality Forum (NQF) measure on the use of angiotensin-converting enzyme inhibitor (ACEi) or angiotensin receptor blocker (ARB) therapy for left ventricular systolic dysfunction (NQF 0081) [1], and (b) HF-CDS, the ATHENA-CDS system [2] that implements similar recommendations from a national guideline for the management of heart failure [3]. 2. Second, we analyzed the changes required to implement a CDS within the context of a PM system rather than in the context of patient visits. The PM system is a production "clinical dashboard" implementation of PMs that generates reports at the levels of VA service network, medical center, provider and individual patient. We integrated ATHENA-CDS system for diabetes with the dashboard to provide recommendations on the management of patients who have failed the PMs.

Results

1. HF-PM and HF-CDS use 33 and 13 criteria, respectively, to select the cohort of patients who should receive ACEi or ARB (9 identical, 4 similar, and 26 unique to the PMs). PM criteria are more explicit in ruling out patients not active in the health care system (death or not visited in last 12 months; criteria not needed in CDS system because it triggers only when a patient is being seen by a health professional). Sixteen HF-PM exclusions not in HF-CDS are conditions for which HF-CDS provides alerts that allow providers to exercise clinical judgment; HF-PM excludes these patients from the cohort because PMs are designed to specify a cohort with a high likelihood that the numerator criteria apply (reduce false positives). That is, whereas these criteria are the same, they are incorporated in the PM and CDS differently (exclusion criteria versus alerts). The output of HF-PM includes the proportion of patients meeting the measure as a percentage of those eligible for the measure. The output of HF-CDS consists of recommendations for managing patients' problems. 2. Integrating the CDS into the PM dashboard to display when the PM is not met (a) allows the CDS to issue alerts based on data computed by the PM dashboard that are important for care but only indirectly related to PM (e.g., medication possession ratios), (b) requires the CDS to change treatment targets (e.g., hemoglobin A1C for diabetic patients) when the targets differ between PMs and guidelines that the CDS implements (because, as is the case for hemoglobin A1C, guidelines recommend setting individualized targets, whereas PM select a target that will identify high risk patients), (c) requires that the CDS trigger and display of CDS recommendations be coordinated with the existing PM dashboard user interface.

Discussion

We use the implementations of two pairs of PM and CDS to demonstrate concretely how PM and CDS differ (e.g., in cohort definitions) and yet can complement each other (e.g., synergy of data and patient selection). We adapted an experimental CDS system to provide recommendations and displays consistent with a production PM dashboard.

References

1. National Quality Forum. NQF-Endorsed Measures (QPS) 2014 [cited 2014 March 13]. Available from: https://http://www.qualityforum.org/Measures_Reports_Tools.aspx.

2. Goldstein MK, Coleman RW, Tu SW, Shankar RD, O'Connor MJ, Musen MA, et al. Translating research into practice: organizational issues in implementing automated decision support for hypertension in three medical centers. J Am Med Inform Assoc. 2004 Sep-Oct;11(5):368-76. PubMed PMID: 15187064.

3. Yancy CW, Jessup M, Bozkurt B, Butler J, Casey DE, Drazner MH, et al. 2013 ACCF/AHA Guideline for the Management of Heart Failure: A Report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. Circulation. 2013 October 15, 2013;128(16):e240-e327.

Acknowledgement

Project supported by VA HSR&D grant IIR 11-071 and VA HSR&D Quality Enhancement Research Iinitiatve (QUERI) grants RRP 11-428 and RRP 12-447. Views expressed are those of the authors and not necessarily those of the Department of Veterans Affairs or other affiliated institutions.

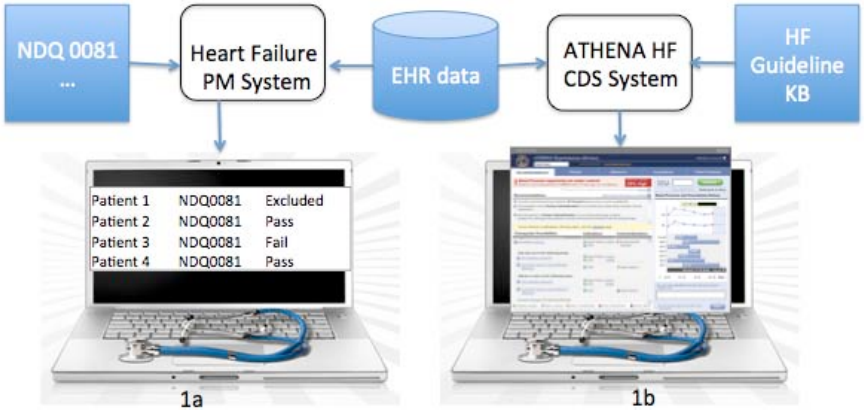


Figure 1. In the first pair of systems, Heart Failure PM (1a) is designed to generate reports on whether a cohort of patients satisfies the NQF 0081 performance measure. ATHENA HF CDS System (1b) is designed to generate detailed management recommendations for guideline-based care at the time of a patient encounter.

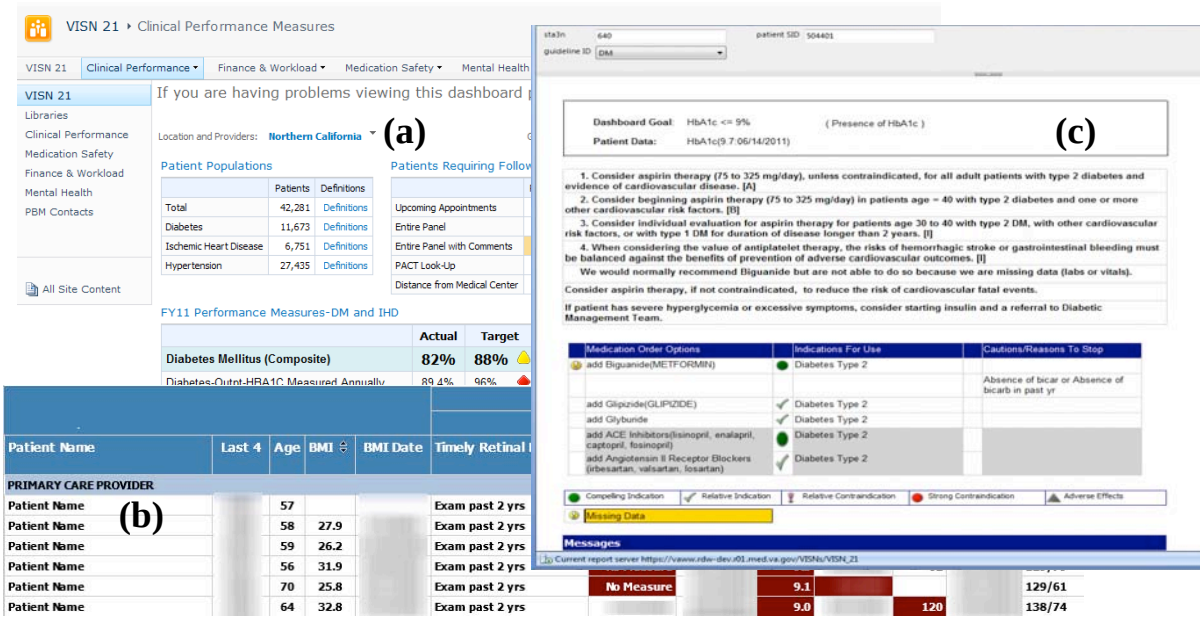


Figure 2. In the second pair of systems, Dashboard PM displays patients' performance with respect to VA PMs as a population-based summary (a) and as individual patient records (b). For selected PMs, a user can bring up ATHENA CDS recommendations on how to manage the treatment of patients who fail to meet the PMs (c).

Federating Air Quality Data with Clinical Data

Ramkiran Gouripeddi, MBBS, MS[1,2], Naresh Sundar Rajan, MS[1,2], Randy Madsen, BS[2], Phillip B. Warner, MS[2], Julio C. Facelli, PhD[1,2]
[1]Department of Biomedical Informatics, [2]Center for Clinical & Translational Science, University of Utah, Salt Lake City, UT

Abstract: *Air Quality (AQ) has been associated with various clinical conditions and requires further clinical research. Linking AQ data with clinical data is often challenging and requires diverse expertise. In this preliminary work we show the feasibility of using OpenFurther in federating these heterogeneous data for performing clinical research.*

Introduction: AQ have been associated with various adverse health effects including asthma, cardiovascular problems, respiratory infections and cancers and impaired glucose tolerance during pregnancies[1–4]. The Salt Lake Valley (SLC), Utah is prone to winter inversions where colder surface temperatures trap fine particulate matter ($PM_{2.5}$) which poses serious health concerns[5]. The summer months in the valley have increased ozone (O^3) levels. Researchers at the University of Utah are embarking on various clinical studies to understand associations between the peculiar AQ patterns in SLC and clinical conditions such cerebral venous thrombosis, exacerbations of idiopathic pulmonary fibrosis, suicide, reproductive outcomes and various cancers.

Methods: To understand the requirements for federating AQ data with clinical data we elicited use-cases from clinical researchers. Primary needs are to understand the risks associated with being exposed with various air pollutants. Manifestations following exposure could occur immediately or after a lag phase and could persist over long durations. Pathophysiology and mechanisms of many of these manifestations are not well understood at this time. The federation needs to support the spatio-temporal variations of air pollutant concentrations and location of an individual and the timing of the occurrence of conditions. Although current research associates single pollutant and clinical conditions, future areas of research could include exposures to multiple pollutants. Requirements for the granularity of AQ data vary from hourly readings to monthly averages depending on the study. Difficulties in understanding and integrating AQ data with clinical data is a limitation in performing research.

OpenFurther[6,7] is an informatics platform that supports federation and integration of data from heterogeneous and disparate data sources. It supports clinical and translational research by bringing data directly to researchers without requiring the technical expertise to query large databases or knowledge about the data source. The main components of OpenFurther include a terminology/ontology server (TS) that stores local and standard terminologies as well as inter-terminology mappings; an in-house developed metadata repository (MDR) that stores metadata artifacts for each data source and the relationships between different data models; a query tool a researcher can leverage to design a clinical research query; a federated query engine that orchestrates queries between the query tool, MDR, TS and the data sources; data source adapters to facilitate interoperability with data sources; and administrative and security components (Figure 1). We evaluated the feasibility of federating limited datasets from the Utah Department of Environmental Quality[8] using OpenFurther.

Conclusion: We were able to select different cohorts of patients living in SLC counties and having clinical conditions (e.g. asthma) occurrences that were related to the temporal variations of air pollutant concentration. Future work will include the ability to select data models that provide best estimates for areas and times when air monitor measurements are unavailable along with uncertainties. We will extend this framework to federate other AQ data sources such as a web-service from the Environmental Protection Agency's Air Quality Datamart[10] making it applicable to all regions of continental United States; balloon[11] and satellite-derived aerosol optical depth measurements[4]; and model a high resolution spatio-temporal AQ grid that cross-links to patient locations and condition occurrences**.**

Acknowledgements: Funded by Grants UU Air Quality Seed Grant, UL1RR025764 and 3UL1RR025764-02S2 from NCRR/NCATS. Apelon, Inc.

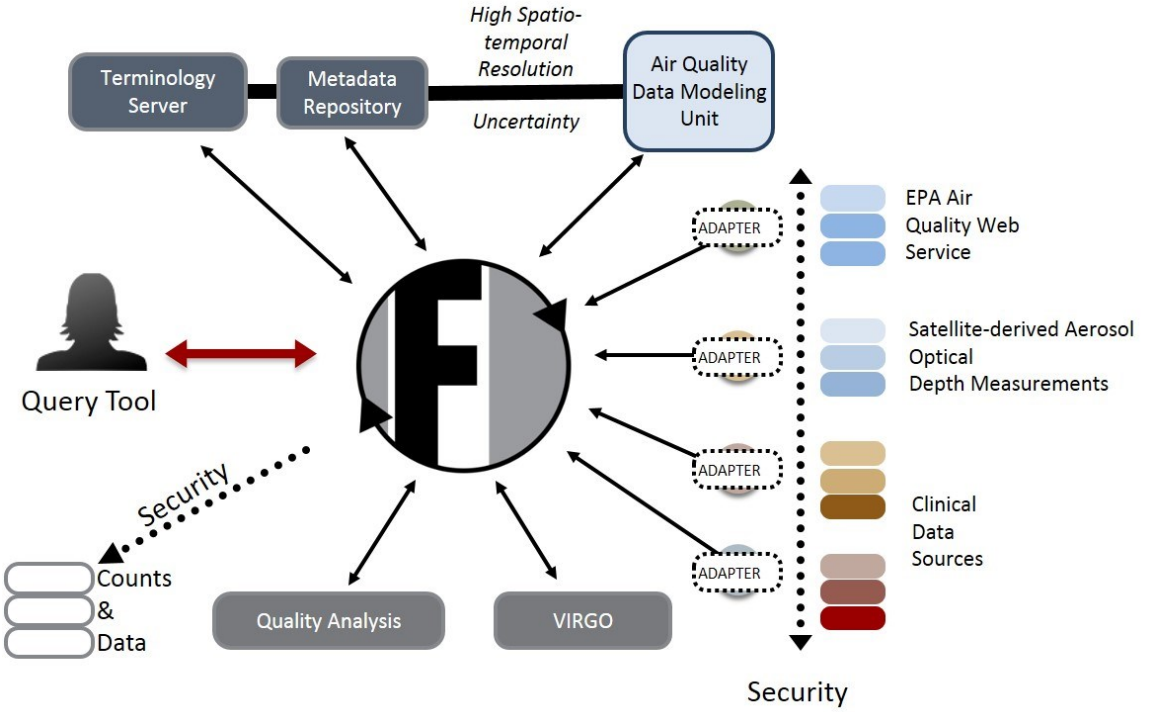


Figure 1: OpenFurther architecture for federating air quality and clinical data.

References

1. Kesten S, Szalai J, Dzyngel B. Air quality and the frequency of emergency room visits for asthma. Ann Allergy Asthma Immunol Off Publ Am Coll Allergy Asthma Immunol. 1995 Mar;74(3):269–73.
2. Weisel CP, Zhang J, Turpin BJ, et al. Relationships of Indoor, Outdoor, and Personal Air (RIOPA). Part I. Collection methods and descriptive analyses. Res Rep Health Eff Inst. 2005 Nov;(130 Pt 1):1–107; discussion 109–127.
3. Fleisch AF, Gold DR, Rifas-Shiman SL, et al. Air Pollution Exposure and Abnormal Glucose Tolerance during Pregnancy: The Project Viva Cohort. Environ Health Perspect. 2014 Feb 7 [cited 2014 Mar 7]; http://ehp.niehs.nih.gov/1307065/
4. Kloog I, Koutrakis P, Coull BA, Lee HJ, Schwartz J. Assessing temporally and spatially resolved PM2.5 exposures for epidemiological studies using satellite aerosol optical depth measurements. Atmos Environ. 2011 Nov;45(35):6267–75.
5. Utah Concludes Winter Inversion Season, Residents Proactively Engaged. [cited 2014 Mar 11]. http://www.deq.utah.gov/News/docs/2014/03Mar/DAQ_NewRelease_AirQualityStats_draftv2.pdf
6. Bradshaw RL, Matney S, Livne OE, et al. Architecture of a Federated Query Engine for Heterogeneous Resources. AMIA Annu Symp Proc. 2009;2009:70–4.
7. Gouripeddi R, Schultz ND, Bradshaw RL, et al. FURTHeR: An Infrastructure for Clinical, Translational and Comparative Effectiveness Research. American Medical Informatics Association, 2013 Annual Symposium; 2013 Nov 16; Washington, D.C. http://knowledge.amia.org/amia-55142-a2013e-1.580047/t-10-1.581994/f-010-1.581995/a-184-1.582011/ap-247-1.582014
8. Utah Department of Environmental Quality. [cited 2013 Jul 17]. http://www.airmonitoring.utah.gov/
9. US EPA O. Air Quality System (AQS) Data Mart, Technology Transfer Network (TTN), US Environmental Protection Agency. [cited 2014 Mar 12]. http://www.epa.gov/ttn/airs/aqsdatamart/index.htm
10. Whiteman CD, Hoch SW, Horel JD, Charland A. Relationship between particulate air pollution and meteorological variables in Utah's Salt Lake Valley. Atmos Environ. 2014 Sep;94:742–53.

Patient Web-Portals: Can Internet Access Explain Differences in Use Among Patients with Chronic Conditions?

Ilana Graetz[1], Courtnee Hamity[2,3], Vicki Fung[4], Nancy Gordon[3], Mary Reed[3].

[1]Department of Preventive Medicine, University of Tennessee Health Science Center, Memphis, Tennessee; [2]School of Public Health, University of California at Berkeley, Berkeley, California; [3]Division of Research, Kaiser Permanente Northern California, Oakland, California; [4]Mongan Institute for Health Policy, Massachusetts General Hospital; Harvard Medical School, Boston, Massachusetts.

Introduction: Patient web-portals have the potential to improve access to care and patient engagement. Meaningful use incentive payments will require that physicians provide patients online access to their health records and the ability to exchange secure emails. The digital divide could limit access to web-based portals among disadvantaged groups. We examined the association between patient characteristics and internet access, and how they relate to use of a patient web-portal.

Methods: A stratified random sample of adult members of an integrated delivery system who were in a chronic disease registry (asthma, diabetes, CAD, heart disease, and hypertension) completed surveys by mail, by internet-based survey, or by telephone interview. During the study period, all members could access a web-portal to send a secure email to a healthcare provider via a web browser. Study participants reported how often they used the internet, if they used to their own computer or smartphone to access the internet, and if they used the web-portal to email a provider in the last year. All analyses were weighted for sampling proportions. We used multivariate logistic regression to assess the association between patient characteristics (sociodemographic and health status) and internet access. To examine characteristics associated with sending secure email to healthcare providers, we examined two model specifications: 1) adjusting for patient characteristics only, and 2) adjusting for patient characteristics, plus frequency of internet use and devices used.

Results: Among 1041 respondents (87% response rate), 59% were white, 44% were age 65+, 27% had household income <$40,000, and 29% had a high school education or less. Overall, 71% used the internet regularly (daily, weekly or monthly), 72% used their own computer and/or 21% used a smartphone to access the internet, and 56% used the web-portal to email a provider. In multivariate analyses, respondents with lower income and education were less likely to use their own computer or smartphone for internet access ($p<0.05$). Blacks and Hispanics were less likely than whites to access the internet using their own computers ($p<0.001$) and respondents age 65+ (vs. 18-44 years) were less likely to use smartphones to access the internet ($p<0.01$). After adjusting only for sociodemographic characteristics, those who were male, age 55+ (vs. 18-44 years), Black or Asian (vs. white), with lower income and education were less likely to have emailed a provider ($p<0.05$). After also adjusting for internet use and devices, only differences by education and gender remained ($p<0.01$). Regular internet use and access via a personal computer were associated with emailing a provider ($p<0.01$); smartphone access was not, but the portal mobile application was not available during the study period.

Discussion: Regular internet use and having a personal computer could explain differences in web-portal use to email providers by age, race, and income. Education and gender-related differences in use of email remained even after controlling for internet access. As the availability and use of patient web-portals increase, it is important to understand which patients have limited access and what barriers they face. Improving internet access and making web-portals available across multiple platforms, including mobile, may reduce disparities in portal use.

Refining a Patient Risk Assessment using Adjusted Clinical Groups (ACG) with Outpatient Lab Results

Kimberly Gudzune MD MPH, Klaus Lemke PhD, Hadi Kharrazi MD PhD, Jonathan Weiner DrPH
The Johns Hopkins University; Baltimore, MD; USA

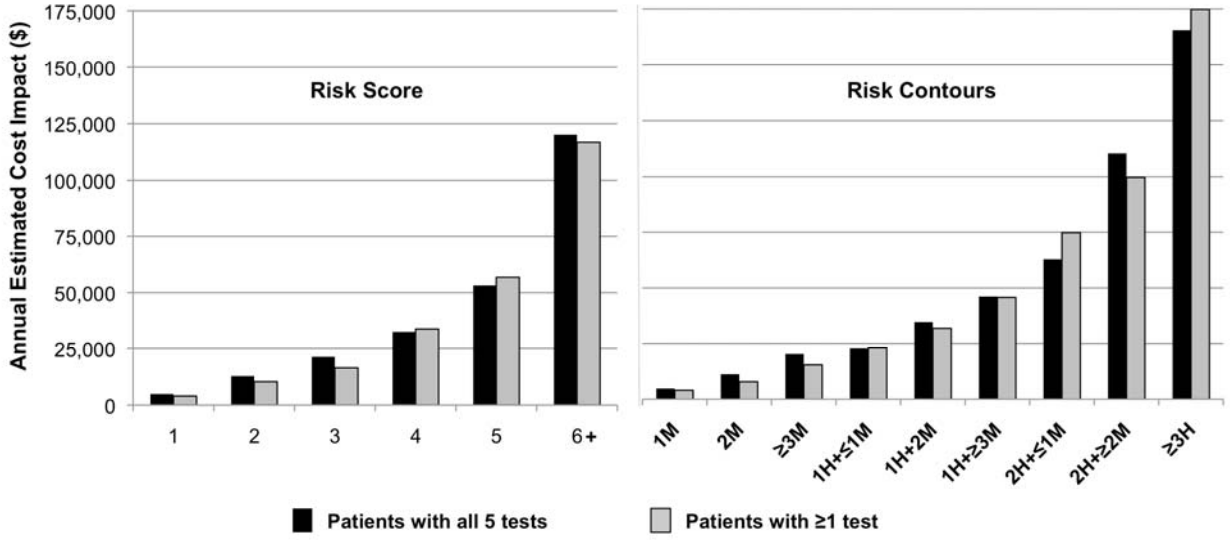
Introduction: Researchers and insurers have relied upon claims-based risk adjustment like the Johns Hopkins Adjusted Clinical Groups (ACG) system; however, electronic medical records (EMR) may provide a new source of data to enhance these claims-based systems. Prior studies have demonstrated that single outpatient laboratory tests can predict prospective costs [1-2], and risk scores based upon admission laboratory tests can predict inpatient mortality [3-6]. However, we know of no studies that have used multiple outpatient laboratory tests to predict healthcare costs. Our objective was to develop an approach to define “lab-based risk” using common outpatient laboratory results, and then test the value of these new measures on refining risk assessment.

Methods: We used 2009-2011 data from IMS Health, which includes enrollment, claims and lab information for three health plans. We examined common outpatient labs including electrolytes, liver function tests, and complete blood counts. Using regression tree analyses, we partitioned each lab test into risk tiers based on concurrent annual costs (low <$20,000, moderate $20,000 to <$60,000 high ≥$60,000). We identified tests with 3 risk tiers (albumin/protein, direct bilirubin, blood urea nitrogen, creatinine, and glucose) to use in the development and testing of two approaches to lab-based risk assessment: “risk score” and “risk contours.” The risk score assigns points to each lab based whether results fall into the high or moderate risk range (2 points and 1 point, respectively). The score ranges from 0-10 points. The risk contours approach designates a level based on counts of both high and moderate risk lab values, resulting in 9 contours that span from “0 high and 1 moderate” to “3+ high” risk results. We analyzed these approaches within two populations: patients who had all five tests (n=7,226) and patients who had at least one of the five tests (n=62,181). We used multivariate linear regression analyses to examine the impact of adding lab-based risk scores or risk contours on concurrent cost prediction beyond the traditional ACG claims-based risk assessment (base model).

Results: Among patients who had all five tests, the estimated cost impact of risk scores ranged from $4,892 (1 point) to $120,146 (6+ points)(Figure). The estimated cost impact of risk contours ranged from $4,786 (1 moderate risk only) to $165,563 (3+ high risks)(Figure). Table displays the differences in R-square for these models as compared to the base model (0.315 and 0.326, respectively, versus 0.224). Among patients who had at least one of the five tests, the estimated cost impact of risk scores ranged from $4,002 (1 point) to $116,803 (6+ points)(Figure). The estimated cost impact of risk contours ranged from $3,983 (1 moderate risk only) to $174,808 (3+ high risks)(Figure). Table displays the differences in R-square for these models as compared to the base model (0.300 and 0.310, respectively, versus 0.232).

Discussion: Both lab-refined risk assessment models explained more cost variation as compared to an ACG only model; however, the risk contours approach performed better than the risk score method. To our knowledge, this is the first study to use multiple outpatient laboratory results to predict healthcare costs. Our results suggest that outpatient laboratory data have utility in improving claims-based risk prediction. Future studies will need to investigate the role of EMR measures beyond lab results, including vital signs and other diagnostic testing, on risk assessment and predictive modeling.

Figure. Adjusted Annual Estimated Cost Impacts of the Base Model + Risk Score Method and the Base Model + Risk Contours Method.



M = moderate risk; H = high risk.

Table. R-square values for regression models using the different lab-refined risk assessment methods

	Patients with all 5 tests (n=7,226)	Patients with ≥1 tests (n=62,181)
Base Model*	0.224	0.232
Base Model + Risk Score	0.315	0.300
Base Model + Risk Contours	0.326	0.310
*Base model includes age, gender, and risk adjustment with ACG resource utilization bands.		

References:

1. McBrien KA, Ravani P, Manns BJ, et al. Health care costs in people with diabetes and their association with glycemic control and kidney function. Diab Care 2013; 36:1172-80.
2. Wagner EH, Sandhu N, Newton KM, et al. Effect of improved glycemic control on health care costs and utilization. JAMA 2001; 285:182-9.
3. Escobar GJ, Green JD, Scheirer P, et al. Risk-adjusting hospital inpatient mortality using automated inpatient, outpatient, and laboratory databases. Med Care 2008; 46:232-9.
4. Escobar GJ, Gardner MN, Greene JD, et al. Risk-adjusting hospital mortality using a comprehensive electronic medical record in an integrated health care delivery system. Med Care 2013; 51:446-53.
5. Jarvis SW, Kovacs C, Badriyah T, et al. Development and validation of a decision tree early warning score based on routine laboratory test results for the discrimination of hospital mortality in emergency medical admission. Resus 2013; 84:1494-9.
6. Mohammed MA, Rudge G, Watson D, et al. Index blood tests and national early warning scores within 24 hours of emergency admission can predict the risk of in-hospital mortality: a model development and validation study. PLoS ONE 2013; 8:e64340.

Perspectives on Care Coordination and Meaningful Use in the Emergency Department Setting

Saira N. Haque, PhD, MHSA[1], Debbie Travers, PhD, RN[2], S. Trent Rosenbloom, MD[3], Jonathan S. Wald, MD[1,4]

[1]RTI International, Research Triangle Park, NC; [2] University of North Carolina-Chapel Hill, Chapel Hill, NC; [3]Vanderbilt University, Nashville, TN; [4]Harvard Medical School, Boston, MA

Abstract

Although the emergency department (ED) has not been a focus for Meaningful Use (MU) efforts in Stage 1, the care coordination objectives for Stage 2 and proposed updates in Stage 3 have implications in the ED setting. More information about how these MU objectives affect the ED setting is needed. To learn more about MU in the ED, we conducted semi-structured interviews with a variety of stakeholders in the EDs of two academic medical centers. These interviews were designed to elicit information about the utility and feasibility of the objectives in the ED setting, electronic health record (EHR) changes needed to facilitate meeting the objectives, and ways to change them. The interviews were coded using NVivo. Perspectives were consistent within and across both EDs. Suggestions included a modified medication reconciliation process and the importance of automated notifications and care summary transmissions.

Introduction

The emergency department (ED) is designed for rapid, acute care with quality metrics that focus on minimizing patient waits and throughput time. To facilitate coordinated care, the ED must exchange relevant information with community caregivers who are caring for a given patient[1]. Such efforts as the patient-centered medical home reinforce this view. The proposed Stage 3 care coordination Meaningful Use (MU) objectives include these measures to facilitate care coordination: medication reconciliation, provision of targeted care summaries for transitions and after-visit consults and referrals, and notification of significant events. Because MU efforts have not traditionally focused on the ED, stakeholders must learn more about considerations with these objectives unique to the ED.

Methods

We conducted semi-structured in-person interviews of key stakeholders at two EDs at academic medical centers in North Carolina and Tennessee. Participants included physicians, nurses, case managers, pharmacists, and ED administrators. The results of the interviews were transcribed and coded using NVivo to identify themes. A sample of 10% of the interviews was double-coded to ensure consistency.

Results

Staff across both EDs felt that the objectives were useful and would benefit patient care. However, they expressed concerns about how the objectives would be put into practice in the ED. They regarded medication reconciliation as important, but stated that a full reconciliation was incompatible with ED workflow. Participants felt a focused review of pertinent medications was sufficient, with a full reconciliation being deferred to the patient's primary care doctor, inpatient admitting service, or pharmacist. For care summaries and notifications for post-ED visits referrals, participants discussed the importance of short, targeted summaries that could be sent automatically. The need for user-friendly EHRs that fit ED workflows and processes was a consistent theme across objectives. Another theme was the utility of health information exchanges to support care coordination, especially for patients with primary providers outside the local healthcare system.

Conclusion

Participants commented that MU was directed to primary care providers, and they felt the care coordination objectives reflected that direction. They suggested scalable requirements so that they could meet the fast-paced care requirements in the ED while also meeting the spirit of the regulation. The findings point to the need to consider a variety of settings and associated information needs and workflows in the MU objectives.

References

1. Shapiro JS, Kannry J, Lipton M, Goldberg E, Conocenti P, Stuard S, Wyatt BM, Kuperman G. Approaches to patient health information exchange and their impact on emergency medicine. Ann Emerg Med. 2006;48(4):426-432.

2. Farley HL, Baumlin KM, Hamedani AG, Cheung DS, Edwards MR, Fuller DC. et al. Quality and safety implications of emergency department information systems. Ann Emerg Med. 2013;62(4):399-407.

Acknowledgements

Funding was received from AHRQ through Contract HHSA 290-2010-00024i, Task Order 5. The views expressed are solely those of the authors and do not reflect the official positions of the institutions or organizations with which they are affiliated or the views of the project sponsors. The opinions expressed are solely those of the authors.

Enhancing the TURF Framework with a Workflow Ontology

Craig Harrington, MS, MSSW[1], Cui Tao, PhD[1], Keith Butler, PhD[2], Jiajie Zhang, PhD[1]
[1]University of Texas at Houston Health Science Center, Houston, TX;
[2]University of Washington, Human Centered Design & Engineering, Seattle, WA

Abstract

The TURF Framework enables usability analysts evaluate EHRs by analyzing work tasks, users, representations, and functionality. Currently the TURF framework does not have a focus on evaluating clinical workflows across multiple tasks or across individuals [1]. Because clinical workflows are vital to the successful implementation of an EHR, this study will augment the capabilities of the TURF usability framework by incorporating a workflow ontology. A novel feature of the workflow ontology is its use of Butler's Cognitive Work Product (CWP), which complements the TURF framework through its ability to classify workflow actions and resources as contributing to the end goal of the workflow or as overhead. The combination of the TURF framework and the workflow ontology provides a more comprehensive method of evaluating the usability of EHRs.

Introduction

The surge in adoption of electronic health records (EHRs) is not without challenges. As highlighted in a recent NIST report and other sources, a key usability issue is the negative impact that EHRs are having on clinical workflows [2, 3]. This study defines a clinical workflow as "a modular sequence of tasks, with a distinct beginning and end, performed for the specific purpose of delivering clinical care" [4]. To address the challenges of clinical workflows in EHRs, our study shows the value of an ontology as an addition to Zhang's TURF framework for evaluating the usability of EHRs.

Ontologies represent domain knowledge declaratively using formal rules and constraints. After defining the abstract nature of a domain, instances of the domain can be created and validated according to the rules and constraints. We have developed a clinical workflow ontology (CWO) for EHR workflow. The declarative reasoning of the CWO complements the nature of clinical workflows to deviate from templated processes. The CWO further enhances TURF through the application of Butler's Cognitive Work Product (CWP), which views workflows involving cognitive processes as changing the CWP from an initial state to a goal state [5]. Necessary activities are identified as advancing the CWP towards the end goal; other activities are likely to qualify as overhead. By combining the CWO with the TURF framework analysts can evaluate EHRs with a proven framework that supports clinical workflows. With this capability, analysts can identify gaps and design solutions that maximize the usability of their EHRs.

Methods

The TURF framework describes methods to analyze the usability of an EHR based on the work tasks, user requirements, information representations, and system functionality. The CWO will enable the usability analyst to also define the steps of a workflow from the initial state to the completion of the workflow using Protégé (6), an open-source visual ontology editor. As the analyst constructs the steps of the workflow, the CWO will apply logic rules to recognize inconsistencies, missing information, or errors while classifying the activities and resources for each step as necessary or overhead. Additionally the CWO supports entry and exit conditions for each step, enabling the workflow to transition to subsequent states based on event triggers or satisfying conditions as opposed to workflows that are represented as ordered sequences of activities. These features enable the CWO to accurately model clinical workflows, where a medical crisis can short-circuit a swim lane diagram.

Results

Two pilot projects previously modeled with UML class and state diagrams serve as reference models to evaluate the resulting CWO models. By enhancing the TURF framework with the CWO, analysts will be able to leverage the logical reasoning capability of the CWO to accurately represent clinical workflows, classify activities as necessary or overhead, and identifying inconsistencies or errors in the modeled clinical workflow.

Discussion

The clinical workflow ontology has the ability to represent workflows in a declarative fashion instead of a procedural fashion. This represents a novel way to represent workflows which mirrors the way clinical workflows occur. Future considerations for the ontology are to incorporate the ability to generate timings for different paths that a workflow can take and the ability to suggest representation methods for the data.

Acknowledgements

Funding/Support: This project is supported by grant number R01HS021233 from the Agency for Healthcare Research and Quality.

This work utilized the Protégé resource, which is supported by grant GM10331601 from the National Institute of General Medical Sciences of the United States National Institutes of Health.

References

1 Zhang J, Walji MF. Turf: Toward a unified framework of ehr usability. J Biomed Inform. 2011 12;**44**(6):1056-67.
2 Koppel R, Kreda DA. Healthcare IT usability and suitability for clinical needs: Challenges of design, workflow, and contractual relations. In: Nøhr C, Aarts J, editors. Information technology in health care: Socio-technical approaches 2010: IOS Press; 2010. p. 7-14.
3 Lowry SZ, Ramaiah M, Patterson ES, et al. (nistir 7988) integrating electronic health records into clinical workflow: An application of human factors modeling methods to ambulatory care. In: NIST, editor.; March 11, 2014.
4 Office of the National Coordinator for HealthIT. Reference taxonomy of clinical workflows users guide. HealthIT.gov; 2012. p. 1-12.
5 Butler KA, Zhang J. Design models for interactive problem-solving: Context & ontology, representation & routines. CHI '09 Extended Abstracts on Human Factors in Computing Systems. Boston, MA, USA: ACM; 2009. p. 4315-20.
6 Stanford Center for Biomedical Informatics Research. Protégé. 4.3.0 ed: Stanford University School of Medicine; 2014.

Automatically Enhancing Discharge Instructions with Pictographs to Improve Patient Recall and Satisfaction

Brent Hill, RN MS PhD Candidate, Seneca Perri, PhD, Jinqiu Kuang, MS, Rebecca Morris, MSPH, Katherine Doyon, BSN, MS, Bruce E Bray, MD, Qing Zeng-Treitler, PhD
University of Utah, 421 Wakara Way, Ste 140, Salt Lake City, Utah 84108

Introduction

Hospital inpatient discharge is important for patients as they are responsible to complete hospital discharge instructions to enhance recovery and prevent adverse events.[1] Several factors limit patient understanding of discharge instructions including the typically large amount of discharge information, the influence of acute and/or chronic disease, medication side effects and poor sleep quality while hospitalized which negatively impacts cognition and memory.[1] One method to improve discharge instructions is to include illustrations, which can increase focus, comprehension, recall and adherence.[2-4]

Our team developed a system called Glyph that contains a pictograph library and each pictograph's associated terminology for medical concepts in a database.[5] The Glyph system preprocesses free text, annotates a set of concept extraction modules that locate and annotate text strings, composes images from grammar patterns, and then renders images, generated by the rule engine, for the corresponding text.[5] (Examples Table 1)

Methods

The aims of the study, to examine immediate and delayed recall and patient satisfaction with discharge instructions, were tested using a between group experimental design with pre-discharge and post-discharge measures of free recall and patient satisfaction. Patients were randomized to either receive pictograph enhanced discharge instructions or standard discharge instructions and then given a copy of their discharge instructions to review for up to 15 minutes immediately prior to inpatient discharge. Recall was assessed by asking patients what they remembered about each section of their discharge instructions. Content the patient verbalized was highlighted on a non-illustrated copy by a study nurse. Words remembered were counted, and a normalization ratio was calculated based on the actual number of words in the patient's discharge instructions. Satisfaction was assessed using a Likert-type scale ranging from 1 as completely dissatisfied to 7 as completely satisfied. Patients were called one-week post discharge and asked the same questions about their discharge instructions. One hundred and forty-four patients completed the study. Additionally, patient demographics (Table 2) were also collected.

Results

An independent-samples T-Test was conducted to determine if there were differences in group means at discharge (immediate recall) and one-week post discharge (delayed recall). The results indicate a significant recall effect for immediate recall, $t(142) = -3.1$, $p < .01$ and a non-significant recall effect at delayed recall, $t(142) = -0.26$, $p = .80$. At discharge, patients who received pictograph enhanced discharge instructions remembered more of their discharge instructions (5.8% vs. 4.3%) than patients who received standard discharge instructions when only illustrated sections of discharge instructions were compared against the same sections in standard discharge instructions.

A Mann Whitney U was conducted to evaluate patient satisfaction with discharge instructions. Two satisfaction questions were asked at discharge and one-week post discharge. There were no significant differences between the groups at discharge for either question, $z = -.53$, $p = .6$ and $z = -.99$, $p = .32$ respectively. For one week post-discharge, significant effects were found in the pictograph enhanced discharge instruction group for question 1 (understandability), $z = -2.4$, $p = .016$ but not for question 2 (inclusiveness).

Discussion

The results of this study validate Glyph-generated instructions effect on improvement in patient recall and satisfaction with discharge instructions. Specifically, pictograph enhancement improved the pre-discharge recall of instructions by 29% and post-discharge satisfaction by 5%. It demonstrates that automated illustration of patient instructions is not only feasible but also beneficial. This informatics intervention can be easily disseminated as Glyph is an automated system.

Learning Objective

After participating in this session, the learner should be better able to apply informatics applications to improve patient education.

References

1. Chugh A, Williams MV, Grigsby J, Coleman EA. Better Transitions: Improving Comprehension of Discharge Instructions. *Frontiers of Health Services Management.* Spring2009 2009;25(3):11-32.
2. Houts PS, Doak CC, Doak LG, Loscalzo MJ. The role of pictures in improving health communication: A review of research on attention, comprehension, recall, and adherence. *Patient Education and Counseling.* 5// 2006;61(2):173-190.
3. Kools M, van de Wiel MWJ, Ruiter RAC, Kok G. Pictures and text in instructions for medical devices: Effects on recall and actual performance. *Patient Education and Counseling.* 12// 2006;64(1–3):104-111.
4. Katz MG, Kripalani S, Weiss BD. Use of pictorial aids in medication instructions: a review of the literature. *American Journal Of Health-System Pharmacy.* 2006;63(23):2391-2397.
5. Bui D, Nakamura C, Bray BE, Zeng-Treitler Q. Automated illustration of patients instructions. *AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium.* 2012;2012:1158-1167.

Table 1. Non illustrated and illustrated discharge instructions.

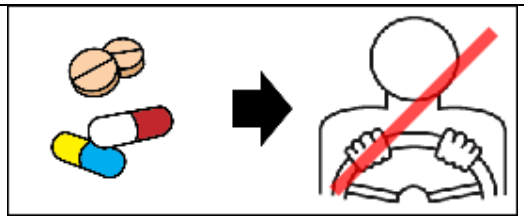
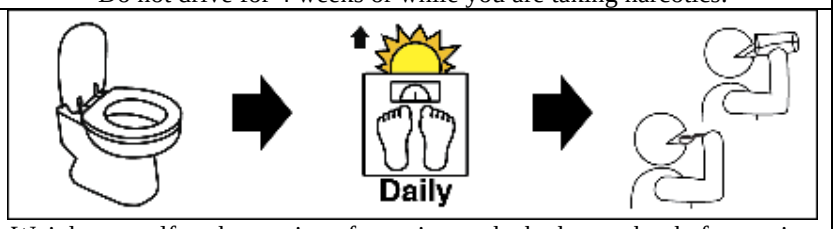
Non Illustrated	Illustrated
Do not drive for 4 weeks or while you are taking narcotics.	 Do not drive for 4 weeks or while you are taking narcotics.
Weigh yourself each morning after going to the bathroom but before eating or drinking.	 Weigh yourself each morning after going to the bathroom but before eating or drinking.

Table 2. Study participant demographics (n=144).

Characteristic	Pictograph N (%)	Standard N (%)	Total N (%)
Gender			
Male	49 (69)	52 (71)	101 (70)
Female	22 (31)	21 (29)	43 (30)
Age			
24-54	21 (30)	28 (38)	49 (34)
55-65	25 (35)	25 (34)	50 (35)
66-90	25 (35)	20 (27)	45 (31)
Race			
White	70 (99)	67 (92)	137 (95)
Black or African American	1 (1)	4 (5)	5 (4)
Native Hawaiian or Other Pacific Islander	0 (0)	2 (3)	2 (1)
Ethnicity			
Hispanic	1 (1)	1 (1)	2 (1)
Non-Hispanic	70 (99)	72 (99)	142 (99)
Education			
<5th Grade	0 (0)	1 (1)	1 (1)
5 to 8th Grade	1 (1)	3 (4)	4 (3)
9 to 12th Grade	12 (17)	23 (32)	35 (24)
>12th Grade	58 (82)	46 (63)	104 (72)

A Web-Based Resource for Medication Management of the Community-Dwelling Older Adult

Kay Jansen, DNP, PMHCNS-BC, RN [1], Amy Coenen, PhD, RN, FAAN[1], Jeeyae Choi, PhD, RN[1]
[1]University of Wisconsin-Milwaukee, College of Nursing, Milwaukee, WI, USA

Abstract

The purpose of this project was to create an evidence-based electronic resource that facilitates nurses to identify actual or potential medication management problems and to target individualized interventions to improve outcomes for older adults living in the community. A discussion of the development process and a demonstration of the web-based resource will be provided with a plan for further research to test the tool in practice.

Introduction

The need for a resource or decision support system (DSS) to assist nurses in managing medications for older adults living the community is supported by a number of factors. Some of the factors for increased risk of medication problems include (a) clinical complexity of the population, (b) desire to remain independent as long as possible, (c) multiple risk factors for problems with medication management, and (d) high costs of hospitalizations and emergency department visits[1]. The US Agency for Healthcare Research and Quality published guidelines for nurses to provide quality medication management for community-dwelling older adults. An electronic resource could help nurses translate evidence into practice and assist with documentation and evaluation of care. The intention of this study was to develop a web-based prototype for an evidence-based DSS using standardized nursing terminology.

Methods

The data sources for this study included (a) a national clinical practice guideline for medication management of the community-dwelling older adult and (b) the expert opinion of four gerontology nurses from Norway and the United States (including the guideline author) with clinical and informatics expertise. The researchers organized content using the nursing process model components of assessment, diagnosis, intervention, and evaluation. Concepts were coded using the International Classification for Nursing Practice (ICNP) to direct development of algorithms to automatically generated tailored care plans for patients based on individual assessments. Use of ICNP can also facilitate translations of the resource and data reuse for quality evaluation and program planning.

Results

An evidence-based resource to support nurses and caregivers with medication management of the community-dwelling older adult was developed for further testing. The resource is a prototype web-based clinical DSS that includes templates for documentation of individual assessments to identify risk factors. Based on the individual risk factors, a tailored care plan is generated for each patient to direct and document interventions provided for medication management. Data documented by nurses are coded in ICNP to facilitate reuse for evaluation.

Discussion and Conclusion

The medication management resource provides one example of an evidenced-based DSS for nurses. Further testing is proposed for nurses' use of the DSS in the homecare setting with older adults and persons with chronic illness. Expanding testing to include informal caregivers and family members, as well as patients, is also planned. This tool could be the first of several web-based resources nurses could access using ICNP. The integration with other electronic health record data to support evidence-based clinical decision making would be essential to optimal use, for example accessing pharmacy and laboratory data for medication reconciliation and monitoring drug dosing.

Acknowledgement

The authors would like to thank the International Council of Nurses in supporting this study.

References

1. Marek, KD, Antle L. Medication management of the community-dwelling older adult. In Hughes RG, editor. Patient safety and quality: an evidence-based handbook for nurses. Rockville, MD: AHRQ; 2008. p. 1509-1536.

Evaluation studies of the Librarian Infobutton Tailoring Environment (LITE): An open access online knowledge capture, management, and configuration tool for OpenInfobutton

Xia Jing, M.B., Ph.D.[1], James J. Cimino, M.D.[2], Guilherme Del Fiol, M.D., Ph.D.[3]

[1]Department of Social & Public Health, Ohio University College of Health Sciences & Professions, Athens OH;
[2] Laboratory for Informatics Development, Clinical Center, National Institutes of Health, Bethesda, MD;
[3] Department of Biomedical Informatics, School of Medicine, University of Utah, Salt Lake City, UT

Abstract

Introduction*: The Librarian Infobutton Tailoring Environment (LITE) is a Web-based knowledge capture, management and configuration tool, with which users can build profiles used by OpenInfobutton, an open source infobutton manager, to provide electronic health record users with context relevant links to online knowledge resources. We conducted evaluation studies of LITE in order to explore users' attitudes and acceptance and to guide future development.* ***Methods****: The evaluation consisted of an initial online survey to all LITE users, followed by an observational study of a subset of users in which evaluators' sessions were recorded while they conducted assigned tasks. The observational study was followed by administration of a modified system usability scale (SUS) questionnaire.* ***Results****: Fourteen users responded to the survey and indicated good acceptance of LITE with feedback that was mostly positive. Six users participated in the observational study, demonstrating average task completion time of less than 6 minutes and an average SUS score of 72.* ***Discussion****: LITE can be used to fulfill its designated tasks quickly and successfully. Evaluators proposed suggestions for improvements in LITE functionality and user interface.*

An Infobutton is a clinical decision support tool embedded within the electronic health records (EHRs) that has been proven to be an effective tool to meet clinicians' information needs at the point of care[1,2]. Certification of "Meaningful Use" of EHRs by the Office of the National Coordinator for Health Information Technology now requires the inclusion of Infobuttons that use the *HL-7 Context-Aware Knowledge Retrieval ("Infobutton") Standard*[3]. The Librarian Infobutton Tailoring Environment (LITE)[*] is a tool that can be used to define context-specific links (CSLs) that can be served up by OpenInfobutton (an open source and centralized Infobutton manager) when an EHR user clicks on an Infobutton in a particular clinical context (task, patient demographics, user demographics, concept of interest, etc.). In order to investigate users' acceptance of LITE and to improve the usability of LITE, we conducted a multi-part evaluation.

LITE allows users to define health care knowledge resources (such as MedlinePlus or UpToDate) and then define institution-specific CSLs that describe the EHR contexts. The definition of institutions, resources and CSLs is guided by "wizards" that step the user through each process. LITE is managed by Drupal. PHP and MySQL are used in back end to support necessary programming functions.

We conducted a three-stage evaluation of LITE[4] from November 2013 to January 2014: a general survey of all users, an observational study of a subset of users, and a follow-up survey of the subset (Fig 1). The surveys were generated and managed by Survey Monkey. The observational study was conducted via Webex. Evaluators conducted the designated tasks and shared their desktops via Webex. All the screen activities were captured via BB FlashBack. Audio recording was optional during the observational study. The researcher observed the entire process and was available for assistance. The researcher also asked follow up questions if an evaluator expressed confusion or particular interest at any point. A modified system usability scale (SUS) questionnaire was administered to observational study participants. We used a published method for calculation of SUS scores[5]. All LITE users were invited for all stages of the evaluation.

The response rate for the general evaluation survey was 16.5% (14/85). The general impressions were quite positive (Fig 2). The users' assessments were that LITE fulfills its main tasks successfully and quickly. The average time for observed users to complete each task was less than 6 minutes under a very relaxed test atmosphere (Table 1). Questions and interviews were common during the tests. The average SUS score for LITE is 72 (the average SUS score from literature is 68). Users felt that of the current Infobutton standard parameters age, sex, EHR task, subject, subtopic, encounter type (.e.g., inpatient, outpatient), user type (e.g., provider, patient), and language were useful but that performer discipline was not necessary. Evaluators suggested that "race or ethnicity" and "geographic locations" be added as new parameters. The main suggestions can be classified into the following categories: navigation, content layout and organization, functionalities, annotations and instructions, interface presentation features and suggestions external to LITE. The main expectations from the evaluators about LITE were: users expected LITE to be more intelligent with more interactions with users; they expected more control over interface display; they desired greater clarity and consistency in the use of terminology within LITE interfaces; and they expected trouble shooting tips, educational materials and support. Due to the small sample size, the findings of the observational and usability sessions need to be considered with caution.

[*] http://lite.bmi.utah.edu

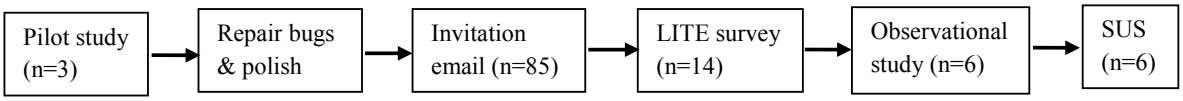


Figure 1. LITE evaluation work flow (SUS: system usability scale questionnaire)

Figure 2. Main survey results summary from the general evaluation of LITE

Table 1. Time spent completing designate tasks during the observational study (in seconds)

	Max	Min	Average	SD	Mean	Count
Institution						
Create	485.5	58.0	177.8	142.4	147.3	7
Change /verify	18.7	6.4	11.3	3.7	11.4	9
Resource						
Create	827.5	139.6	335.2	281.0	237.2	5
Modify	273.8	27.9	109.8	100.0	93.4	5
Verify	244.9	5.3	62.34	79.2	33.6	8
CSL						
Create	516.6	124.8	267.3	167.0	188.4	6
Modify/verify	166.1	27.4	90.6	70.2	78.4	3
Test of CSL	383.4	59.9	210.4	122.5	180.2	9

References

1. Cimino JJ. An integrated approach to computer-based decision support at the point of care. *Trans Am Clin Climatol Assoc*. 2007;118:273–288.
2. Del Fiol G, Haug PJ, Cimino JJ, Narus SP, Norlin C, Mitchell JA. Effectiveness of topic-specific infobuttons: a randomized controlled trial. *J Am Med Inf Assoc*. 2008;15(6):752–759.
3. Cimino JJ, Jing X, Del Fiol G. Using OpenInfobutton and the Librarian Infobutton Tailoring Environment (LITE) to Meet the Electronic Health Record "Meaningful Use" Requirement for Using the HL7 Infobutton Standard. *AMIA 2012*. 2012.
4. Kushniruk AW, Patel VL. Cognitive and usability engineering methods for the evaluation of clinical information systems. *J Biomed Inf*. 2004;37(1):56–76.
5. Brooke J. SUS - A quick and dirty usability scale. Available at: http://hell.meiert.org/core/pdf/sus.pdf.

Using Phenotype Portal for Checking Clinical Guideline Recommendation Compliance

Lara Johnstun, BS[1], Danielle Groat, BS[1], Amol Bhalla, MD, MHSA[1], Kevin J. Peterson, BA[2], Jyotishman Pathak, PhD[2], Adela Grando, PhD[1]

[1]Arizona State University, Biomedical Informatics, Scottsdale, AZ
[2]Mayo Clinic, Health Sciences Research, Rochester, MN

Abstract

Phenotype Portal is a tool that currently uses de-identifed Mayo Clinic and InterMountain Healthcare clincal data to identify cohorts for electronic clinical quality measures (eCQMs) using Meaningful Use (MU) terminologies and standards. Our goal is to test Phenotype Portal's reusability by using the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC II) clincal data, and to understand Phenotype Portal's scalability to the problem of clinical guideline recommendation compliance for the treatment of asthma exacerbation.

Introduction

The Phenotype Portal[1] is a platform for clinicians and investigators to identify patient cohorts using electronic health record data by leveraging informatics-based phenotyping processes. The use of Phenotype Portal requires the clinical data to be standardized and normalized to MU conformant terminology and value set[2] standards using Clinical Element Models (CEM). Samples of both structured and unstructured data from Mayo Clinic and InterMountain Healthcare EHRs have been used in the Phenotype Portal to demonstrate semi-automatic execution of eCQMs, which are electronic specifications of clinical quality measures required for EHR reporting for MU. Currently, the Phenotype Portal supports 50 eCQMs for identifying cohorts for clinical quality measures. Our aim is to use intensive care unit data from MIMIC II[3] clinical database to obtain preliminary data to understand if the Phenotype Portal could also be used to specify and check adherence to evidence-based clinical recommendations. MIMIC II contains de-identified critical care data from hospital information systems, detailing thousands of admissions. MIMIC II clinical database has been previously normalized to ICD-9CM, ICD-9PCS and LOINC, but lacks standard terminologies for other concept categories such as medications and patient encounter types. Pulmonary disease is among the most frequent diagnostic groups present in MIMMIC II, representing 8.84% of assigned ICD-9CM codes. With the focus on pulmonary disease, we chose to check the compliance of medical treatment recommendations from the 2007 National Heart Lung and Blood Institute "Guidelines for the Diagnosis and Management of Asthma." For instance, if a patient is admitted to a hospital for severe asthma exacerbation and is unresponsive to initial treatment with systemic corticosteroids, then adjunct treatment with heliox or IV magnesium should be considered.

Method

First, to standardize and normalize MIMIC II data in accordance to the Phenotype Portal data constraints we proposed automatic methods to normalize medications to RxNorm. SNOMEDCT and CPT codes are added to patient data to explicitly express the implicit concept of ICU care. Patient demographic values are normalized to codes consistent with value set standards. Second, to specify the chosen asthma guideline recommendations into the Phenotype Portal we created electronic specifications for asthma clinical recommendations within the eCQM framework (see Figure 1) using the Measure Authoring Tool[4] (MAT) and authored value sets (see Table 1).

Results

We imported asthma clinical guideline specifications into the Phenotype Portal to check the conformance of the asthma recommendations in the MIMIC II data and found that Phenotype Portal was able to identify cohorts of patients whose treatment conformed to asthma clinical guidelines from a sample of MIMIC II ICU patients.

Conclusions

Our research shows that Phenotype Portal can be reused with normalized clinical data different from the Mayo and InterMountain Healthcare data originally used. In particular, mapping MIMIC II data into PhenotypePortal required extra methods for normalizing medications, patient encounters and demographics.

In terms of adaptability, this preliminary work on specifying and checking compliance with clinical recommendations has helped to gain insights on the scalability of the Phenotype Portal to address a problem different from its original purpose of identifying patient cohorts for MU. Identification of cohorts for conformance with clinical recommendations allows developers of clinical decision support (CDS) to evaluate a baseline for how well clinical guidelines are being applied before the development and implementation of CDS tools. If baseline conformance cohorts reveal a need for CDS development, the effectiveness of the CDS tool can be measured following its implementation.

- **Initial Patient Population =**
  - AND: "Diagnosis, Active: Acute Asthma Exacerbation"
- **Denominator =**
  - AND: "Initial Patient Population"
  - AND: "Diagnosis, Active: Acute Asthma Exacerbation"
- **Denominator Exclusions =**
  - None
- **Numerator =**
  - AND: "Medication, Order: Oxygen Recommended for Hospitalized Asthma Exacerbation Patients"
  - AND: "Medication, Order: SABA's Recommended for Hospitalized Asthma Exacerbation Patients" concurrent with "Medication, Order: Oxygen Recommended for Hospitalized Asthma Exacerbation Patients"
  - AND: "Medication, Order: Systemic Corticosteroids Recommended for Hospitalized Asthma Exacerbation Patients" concurrent with "Medication, Order: SABA's Recommended for Hospitalized Asthma Exacerbation Patients"
  - AND: "Medication, Order: SABA's Recommended for Hospitalized Asthma Exacerbation Patients" starts before start of "Medication, Order: *Adjunct Therapies Recommended for Hospitalized Asthma Exacerbation Patients"*
- **Denominator Exceptions =**
  - None

Figure 1: Example of Asthma Guideline Recommendation Within eCQM Format
Utilizing a framework of logical, statistical and temporal constraints developed for eCQMs, MAT[4] inserts asthma exacerbation value sets within intermediate structures to create an executable specification of a recommendation from the asthma guidelines[4].

Value Set Name	Adjunct Therapies Recommended for Hospitalized Asthma Exacerbation Patients
OID	2.16.840.1.113762.1.4.1085.12
Type	Extensional
Code System	RXNORM
Code System Version	2014-06
Code	Descriptor
1158055	Magnesium Sulfate Injectable Product
1160308	Helium / Oxygen Inhalant Product
348063	Helium / Oxygen Gas for Inhalation
372696	Magnesium Sulfate Injectable Solution
6585	Magnesium Sulfate
692796	Helium / Oxygen

Table 1: Value Set for Asthma Exacerbation
Value sets that group terminology instances which are members of asthma clinical concepts defined within 2007 Guidelines for the Diagnosis and Management of Asthma[4] were authored in NIH/NLM Value Set Authority Center [2].

References

1. Pathak J, Bailey KR, Beebe CE, et al. Normalization and standardization of electronic health records for high-throughput phenotyping: The SHARPn consortium. *J Am Med Inform Assoc*. 2013;20(e2):e341-8.
2. NIH/NLM Value Set Authority Center: https://vsac.nlm.nih.gov/#. Accessed on July 18th 2014.
3. Saeed M, Villarroel M, Reisner AT, et al. Multiparameter intelligent monitoring in intensive care II: A public-access intensive care unit database. *Crit Care Med*. 2011;39(5):952-960.
4. National Asthma Education and Prevention Program. Expert Panel Report 3 (EPR-3): Guidelines for the Diagnosis and Management of Asthma-Summary Report 2007. J Allergy Clin Immunol 2007 Nov;120(5 Suppl):S94-138.
5. NQF Measure Authoring Tool: http://www.qualityforum.org/MAT/. Accessed on March 10th 2014.

Generating Problem Lists Using Medication Reconciliation

Joshua W. Joseph, MD[1], David T. Chiu, MD[1], Larry A. Nathanson, MD[1], Steven Horng, MD MMSc[1]
[1]Beth Israel Deaconess Medical Center / Harvard Medical School, Boston, MA

Abstract

Introduction*: Problem lists succinctly summarize a patient's history, and help to quickly identify important medical conditions. As patients' medical conditions and records become more complex, obtaining a useful problem list can be inefficient and inaccurate, particularly for critically-ill patients. Regulatory agencies will soon require providers to maintain an up-to-date problem list at every patient encounter. It has been suggested that medication reconciliation could be used to help clinicians meet this regulatory mandate. We hypothesized that we could generate an accurate, automated problem list from the ED medication reconciliation.*

Methods*: The objective of this study was to evaluate the sensitivity and specificity of an automatically generated problem list from the medication reconciliation. The study was approved by our institutional review board. We performed a 1 month retrospective chart review of patients admitted via the ED of an academic, tertiary care center with an annual volume of 55,000 patients. Consecutive adult patients were enrolled between 03/29/2013 - 04/30/2013. Patients who were admitted and had an inpatient discharge summary within a year were included. Patients who eloped or signed out against medical advice were excluded. Our rules-based method used the First Data Bank medication ontology to group medications into therapeutic classes. We then applied a set of clinically derived rules to those therapeutic classes to predict asthma, hypertension, diabetes, congestive heart failure (CHF), and history of venous thromboembolism (VTE). This prediction was compared to the problem list found in the EM attending note, the electronic medical record, and the last discharge summary (gold standard). We additionally trained and validated probabilistic predictive models. A logistic regression and neural network with 5 nodes in 1 hidden layer were created for both CHF and stroke risks. Sensitivity and specificity were determined by maximizing Youden's J statistic. These parameters were compared with a problem list derived from the EM attending note, the problem list in the electronic medical record (EMR) and the rule-based method.*

Results*: A total of 603 patients were enrolled from 03/29/2013 - 04/30/2013. The rules-based method had a sensitivity of 0.71 for asthma, 0.85 for hypertension, 0.75 for diabetes, 0.53 for VTE, and 0.59 for CHF. Comparatively, the attending note had a sensitivity of 0.53 for asthma, 0.59 for hypertension, 0.71 for diabetes, 0.58 for VTE, and 0.36 for CHF, while the electronic record's problem list had a sensitivity of 0.38 for asthma, 0.40 for hypertension, 0.67 for diabetes, 0.31 for VTE, and 0.42 for CHF. The logistic regression model had a higher sensitivity than all other methods for detecting CHF. Logistic regression also had a higher specificity for stroke compared to the rules-based method, and equivalent to the rules-based method for CHF; however, it had a lower specificity than either the EMR problem list or the EM attending for both stroke and CHF. The AUC for logistic regression for CHF and stroke are 0.84 and 0.85, respectively. The AUC for neural network for CHF and stroke are 0.96 and 0.98, respectively. The Hosmer-Lemeshow test was not significant p >0.05 for the logistic regression model indicating that the model was well calibrated.*

Discussion*: The rules-based method was sensitive for conditions that are normally treated with an exclusive class of medications. Probabilistic methods such as logistic regression and neural networks are more sensitive at detecting more complex diseases such as CHF and stroke history than rules-based method, problems already documented in the electronic medical record, or EM attending note. Neural networks have an even higher sensitivity at detecting history of stroke than This suggests that algorithms have the potential to facilitate more efficient and thorough problem list documentation, but further investigation is required.*

Table: Sensitivity and Specificity of an Automated Problem List and Standard Lists

	Medication Algorithm		EMR[1] Problem List		EM Attending Note	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
Asthma / COPD *(n=117; 19%)*	0.71	0.89	0.38	0.98	0.53	0.98
Hypertension *(n= 382; 63%)*	0.85	0.63	0.40	0.91	0.57	0.92
Diabetes *(n= 189; 31%)*	0.75	0.96	0.67	0.98	0.71	0.98
History of VTE[2] *(n= 112; 19%)*	0.53	0.91	0.31	0.97	0.49	0.98
CHF *(n= 153; 25%)*	0.59	0.82	0.42	0.82	0.39	0.98

[1] Electronic Medical Record
[2] Venous Thromboembolism

Table: Sensitivity and Specificity of Detection of Congestive Heart Failure and History of Stroke

	Congestive Heart Failure (*n=153; 25%*)		History of Stroke (*n=149; 19%*)	
	Sensitivity	Specificity	Sensitivity	Specificity
Neural Network	0.55	0.95	0.86	0.91
Logistic Regression	0.75	0.78	0.77	0.81
Rules Based Algorithm	0.59	0.82	0.13	0.50
EMR* Problem List	0.42	0.98	0.43	0.97
EM Attending Note	0.39	0.98	0.57	0.98

*Electronic Medical Record

Evaluation of Stage 3 Care Coordination 'Meaningful Use' (MU) Objectives among Eligible Hospitals (EH)

Hadi Kharrazi, MD PhD[1]
[1]Johns Hopkins School of Public Health, MD

Introduction

Industry experts and government officials have expressed skepticism over the effectiveness of Meaningful Use (MU) in improving health outcomes and lowering cost, and expect upcoming MU stage implementation to be complex and costly. Thus, policy makers have decided to evaluate the feasibility and usability of MU measures before finalizing stage 3 MU objectives (MU3) in 2014, and perhaps adapt recommended changes before an MU3 rollout in 2016/2017.

This federally-funded project aims at evaluating the feasibility and practicality of two specific MU3 measures that are proposed by ONC (Office of National Coordinator) to incentivize eligible hospitals (EH) to effectively adopt and use EHRs (Electronic Health Records) for inpatient/outpatient care coordination. The focal objectives of these care coordination MU3 measures are: (1) sending hospital-generated referral results to the original requestors in outpatient settings; and (2) timely notification of key patient care team members of significant life events that occurred during a hospitalization.

Methods

The study includes three phases: In the first phase, the research team analyzed the current measures and developed a matrix of pragmatic data, workflow and organizational elements that represent the two MU3 care coordination measures. This matrix is used to develop a data collection instrument to evaluate the two MU3 measures (#305 & #308). In the second phase, 10 eligible hospitals (5 in Maryland and 5 in Arkansas), representing a variety of hospital types and sizes, participated in the data collection and pilot roll-out of these two measures. In the last phase, the research team summarized the results into challenges and opportunities in adopting these two MU3 measures and reflected on the aggregated feedback collected from EHs to policy makers.

Results / Principal Findings

Results of the first and second phase of the study show mixed outcomes. None of the participating hospitals had the capability to roll-out the two MU3 care coordination measures on a system-wide scale; however, a number of them are able to collect the necessary data to evaluate the feasibility of the roll-outs. Based on the preliminary findings, the majority of the participating hospitals did not collect the necessary data required to report these two MU3 measures in an electronic format. The only participating critical access hospital dropped out of the study due to resource constraints. Common EHRs do not have the capability to automate the roll-out of these measures; however, there are innovative approaches to enhance the current EHRs and/or adjust the clinical or administrative workflows to adapt these measures in near future. Detailed results of the study, including differences found in the feasibility of these two MU3 measures based on hospital specifications, will be reported to the Agency for Healthcare Research and Quality (AHRQ) in Sep 2014.

Discussion

The overall concerns of participating hospitals with these two MU3 measures are summarized in the following categories: (1) Ambiguous Definition: Most EHs found the definition of these two measures unclear and with no specified boundaries or denominators; (2) Overlapping Measures: Some EHs found these two MU3 measures somehow overlapping with existing MU2 measures (#303) and suggested merging them; (3) Beyond EHRs: A number of hospitals considered some of the elements of these objectives to go beyond the EHR's common functionalities and suggested that other IT tools may be required to achieve these MU3 objectives; (4) Connectivity to Health Information Exchanges (HIEs): HIE connectivity can play a significant role in some of these MU3 care coordination measures and hospitals that do not have access to such connections may not be able to fully achieve these measures; and, (5) Hospital Specs / IT Infrastructure: A number of hospitals specs and underlying IT infrastructure play a role in achieving advanced care coordination measures. For example the rural and semi-rural hospitals that participated in the study could not collect the necessary raw data to measure these objectives while larger academic medical hospitals could achieve parts of the measures.

The final project report will provide policy recommendations to policymakers on how to refine these measures in order to increase their impact; propose EHR innovations that would better enable providers to meet the proposed objectives; and, suggest ways to increase the value of these two care coordination measures to providers. In conclusion, new care coordination measures of MU3 will be challenging for EHs unless EHR vendors automate some aspects of data collection and exchange, and also the clinical staff/providers adjust their workflow to enhance care coordination.

Acknowledgement

This study is funded by AHRQ ACTION-II funding mechanism.

Table 1: Summary of participating hospitals specifications (organizational, IT infrastructure, MU attestation)

#	State	U/R	Teach	Net.	Beds	Ex. Data	HIE Part.	PCP ER	Qry. Out	CM Avg.*	MM Avg.**
1	MD	U	Maj	Y	951	0.50	P	Y	Y	.960	.691
2	MD	U	-	-	371	0.75	P	N	Y	.937	.833
3	MD	U	Min	Y_A	460	0.67	P	Y	Y	.954	.606
4	MD	U	-	-	116	0.75	P	Y	Y	.935	.454
5	MD	U	-	Y_A	244	0.67	P	Y	Y	.932	.611
6	AR	U	-	-	270	0.50	TF	Y	N	.962	.539
7	AR	U	Min	Y_B	383	0.92	TF	Y	Y	.973	.608
8	AR	R	-	-	209	0.33	TF	N	N	.953	.690
9	AR	U	-	Y_B	141	0.25	NP	Y	N	.959	.613

MD: Maryland; AR: Arkansas; U: Urban; R: Rural; Teach: Teaching role; Maj: Major; Min: Minor; Net: Part of a larger network; Y: Yes; Y_X: Yes and part of network X; Ex Data: Exchange capability for labs, reports, and summary of care within hospitals entities, with outside hospitals, and outside ambulatory care centers; HIE Part.: Heath Information Exchange participation level; P: Participate in a local HIE/RHIO; TF: Have the technical framework but does not participate; NP: Neither has the framework nor participate in a local HIE/RHIO; PCP ER: Has the capability to electronically notify the primary care physician about the ER admission of their patients; Qry Out: Can automatically query patient data from outside providers; CM_{Avg}: Core Measure averages (excludes categorical measures); MM_{Avg}: Menu Measure averages (includes categorical measures; excludes unreported measures).

Table 2: Feasibility of collecting MU3 #305 data elements from various electronic sources at participating hospitals

(a) Referral Request / Initiation	Diff
(a.1) # of referrals received from an outside provider	9.2
(a.2) # of referrals received from an in-network provider	7.6
(a.3) # of referrals received from a provider within own hospital setting	3.8
(b) Results Available – Referral Performed	Diff
(b.1) # of referral results generated for a referral requested from an outside provider	9.3
(b.2) # of referral results generated for a referral requested from an in-network provider	7.6
(b.3) # of referral results generated for a referral requested from a provider in own hospital setting	6.4
(c) Requestor Identification	Diff
(c.1) # of referrals received from an outside provider that has a valid requestor contact info associated with it	7.6
(c.2) # of referrals received from an in-network provider that has a valid requestor contact info associated with it	7.0
(c.3) # of referrals received from a provider within own hospital setting that has a valid requestor contact info	6.2
(d) Results Sent Back – Referral Completed	Diff
(d.1) # of referral results sent back to the requestors when referrals are received from an outside provider	8.4
(d.2) # of referral results sent back to the requestors when referrals are received from an in-network provider	7.4
(d.3) # of referral results sent back to the requestors when referrals are received from a provider within hospital	7.2
(e) Result Receipt – Referral Loop Completed	Diff
(e.1) # of referral result-receipt acknowledgement / confirmations received from an outside provider	9.4
(e.2) # of referral result-receipt acknowledgement / confirmations received from an in-network provider	7.8
(e.3) # of referral result-receipt acknowledgement / confirmations received from a provider within hospital	7.8

Diff: Average difficulty level to measure specific data sub-element of an objective
(out of 10; ranges from very easy 1 to very hard 10; each level includes a specific technical difficulty scenario);

Extracting Drug-drug Interactions from Literature Using a Rich Feature-based Linear Kernel Approach

Sun Kim, PhD[1], Haibin Liu, PhD[1], Lana E. Yeganova, PhD[1], W. John Wilbur, MD, PhD[1]

[1]National Center for Biotechnology Information, NLM, NIH, Bethesda MD

Abstract

Introduction: When two or more drugs are being co-administered, there is a chance they will interact. This is especially true for new drugs which often have side effects that may remain unnoticed until they are available to the public. Identifying such interactions is of significant importance in the early detection of unintended or harmful drug reactions, leading to reduced number of drug-safety incidents and reduced healthcare costs. While a number of resources containing drug-drug interaction (DDI) information have been created (1, 2), the wealth of up-to-date DDI information is hidden in unstructured medical texts which are growing exponentially. This calls for developing text mining techniques to automatically identify DDIs from free text.

Methods: In this work we develop an efficient and scalable state-of-the-art system for identifying drug-drug interactions and classifying them into four predefined relation types: mechanism, effect, advice and general interaction, as described in (3). Among existing DDI extraction methods, the best performing ones are based on Support Vector Machines (SVM) with non-linear, composite kernels. However, these approaches incur more computational cost because the complexity of the underlying kernels accumulates and additional learning is required to optimize the weights for individual kernels. We propose a DDI extraction method using a linear SVM classifier with a rich-feature representation. The 'one-against-one' strategy (4) is used to further classify DDIs into four types.

Five different types of lexical and semantic features are defined to capture the characteristics of data. Lexical features include individual word features with position information and word pair features for non-adjacent words in a sentence. Integrating position information into word features is important because one sentence often involves multiple drug mentions and the position information helps differentiate the context of interacting pairs from that of non-interacting ones. Word pair features capture expression patterns involving distant words in a sentence and can effectively complement word features. Semantic features include dependency relations, parse tree structures and noun phrase-constrained coordination features. Dependency relations describe governor-dependent relations between words and are able to capture long-range dependencies between target drug mentions. Parse tree structures encode a sequence of grammatical tags representing a syntactic traverse from one drug to another. Finally, the noun phrase-constrained coordination features examine whether target entities appear in a coordination of three or more entities, an unlikely scenario for discussing interactions. The word pair and the coordination features are novel to our approach. Figure 1 illustrates the preprocessing step and features extracted for an example drug pair.

Results: We evaluated our system on the DDIExtraction 2013 challenge corpus (5) and compared it with the top scoring systems in the challenge. Our approach achieves an F1 score of 0.676 on the challenge test data, as compared to 0.651 and 0.609 reported by the two best performing teams FBK-irst (6) and WBI(7), both based on non-linear kernels. The highest F1 score reported in the challenge by linear kernel systems (8) is 0.594. Table 1 reports our DDI detection and classification performance as compared to the top three ranking teams.

Discussion: As evidenced by the results on the publicly available dataset, our method is the first linear kernel-based approach that achieves the state-of-the-art performance on both DDI detection and classification tasks. A rich feature representation appears to be beneficial as it allows a linear SVM system to achieve a competitive performance. We consider our method a strong alternative to the nonlinear, composite kernel-based approaches. The advantage of linear kernel approaches is in their simplicity and scalability. As demonstrated in (9), often linear kernels are the only practical choice for training large-scale datasets. Moreover, straightforward representations used by linear kernels enable an intuitive interpretation of results. In addition, the 'one-against-one' strategy proves vital for classifying extracted interactions into specific types. Unlike other top scoring systems, our method does not incorporate external domain-specific information, which suggests it may be more generalizable across different domains. The inherent simplicity and transparency of the proposed method could be especially beneficial if it is used as a part of a more complicated schema.

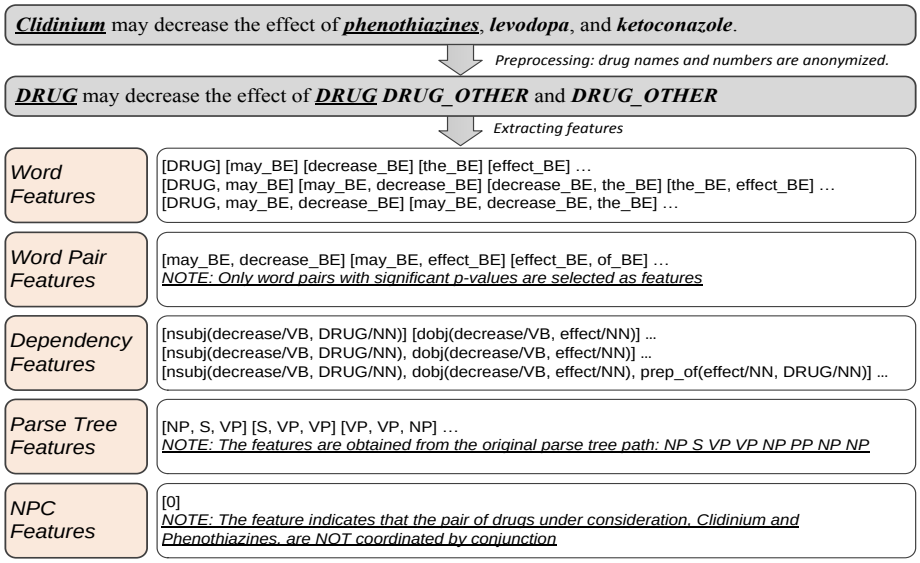


Figure 1. An example of preprocessing and feature extraction. The underlined pair of drugs, *clidinium* and *phenothiazines*, is a candidate interacting pair. 'BE' in word features denotes position (between candidate drugs).

Table 1. Performance comparison between the proposed method and top ranking approaches on the test data of the DDIExtraction 2013 challenge, evaluated by the official metrics. The performance is reported in F1 scores.

Method	Detection & Classification	Detection	Mechanism	Effect	Advice	General Interaction
Our Method	0.670	0.775	0.693	0.662	0.725	0.483
FBK-irst	0.651	0.800	0.679	0.628	0.692	0.547
WBI	0.609	0.759	0.618	0.610	0.632	0.510
UTurku	0.594	0.696	0.582	0.600	0.630	0.507

References

1. Baxter K, Claire L P. Stockley's Drug Interactions, 10th edition. Baxter K, Claire L P, editors. London: Pharmaceutical Press; 2013.
2. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, et al. DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs. Nucleic Acids Research (Database issue). 2011;39(Suppl 1):D1035–D41.
3. Segura-Bedmar I, Martinez P, Herrero-Zazo M, editors. SemEval-2013 Task 9 : Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). Second Joint Conference on Lexical and Computational Semantics (*SEM); 2013; Atlanta, GA.
4. Hsu C-W, Lin C-J. A comparison of methods for multiclass support vector machines. IEEE Transactions on Neural Networks. 2002;13(2):415-25.
5. Herrero-Zazo M, Segura-Bedmar I, Martínez P, Declerck T. The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions. Journal of Biomedical Informatics. 2013;46:914-20.
6. Chowdhury MFM, Lavelli A, editors. FBK-irst : A Multi-Phase Kernel Based Approach for Drug-Drug Interaction Detection and Classification that Exploits Linguistic Information. Second Joint Conference on Lexical and Computational Semantics (*SEM); 2013; Atlanta, Georgia.
7. Thomas P, Neves M, Rocktaschel T, Leser U, editors. WBI-DDI: Drug-Drug Interaction Extraction using Majority Voting. Second Joint Conference on Lexical and Computational Semantics (*SEM); 2013; Atlanta, Georgia.
8. Bjorne J, Kaewphan S, Salakoski T, editors. UTurku: Drug Named Entity Recognition and Drug-Drug InteractionExtraction Using SVM Classification and Domain Knowledge. Second Joint Conference on Lexical and Computational Semantics (*SEM); 2013; Atlanta, Georgia.
9. Bjorne J, Ginter F, Pyysalo S, Tsujii Ji, Salakoski T. Scaling up Biomedical Event Extraction to the Entire PubMed. Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, ACL 2010. 2010:28-36.

Evaluation of an Early Prototype of a Patient-Centered Decision Aid to Improve Accuracy of Breast Cancer Risk Perception

Rita Kukafka, DrPH, MA[1] Katherine D. Crew, MD, MS[1], Haeseung Yi, MPA[1], Tong Xiao[1] , Parijatham S. Sivasubramanian, MD[1], Alejandra N. Aguirre[1],
[1]Columbia University, New York, NY

Abstract

This study evaluated a breast cancer prevention decision aid (DA), RealRisks, which incorporates experience-based dynamic interfaces to communicate risk aimed at reducing inaccurate risk perceptions, particularly in low-numerate women. We employed a mixed methods approach to assess accuracy of risk perceptions and the acceptability of the DA using qualitative focus groups and a quantitative survey. After interacting with RealRisks, the difference in perceived vs. actual breast cancer risk significantly improved for 5-year risk (p=0.008), but not lifetime risk (p=0.20). In addition, while participants were uncertain of their risks before RealRisks, none of them expressed uncertainty after RealRisks. Not only did participants think that RealRisks was easy to use and useful, but they also became less worried about breast cancer and increased their knowledge of breast cancer.

Introduction

Breast cancer risk assessment and prevention strategies, such as chemoprevention are underutilized in the U.S.[1] Reasons for low uptake include inability to routinely screen for high-risk women in the primary care setting, inadequate time for counseling, and insufficient knowledge about risk-reducing strategies. To address patient-related barriers, we developed a prototype of a web-based decision aid, RealRisks, that incorporates heuristic or experience-based techniques for learning and problem-solving to change perceptions and behaviors.[2]

Methods

The RealRisks DA includes modules on breast cancer risk, genetic testing, and chemoprevention. Patient-provider dialogue is modeled using narrative, and participants learn about their risk by interacting with 2 games of experience-based risk interfaces, which demonstrate average 5-year and lifetime breast cancer risk. Women recruited from Northern Manhattan (N=34) participated in this study. We employed a mixed methods approach to assess improved accuracy of risk perceptions and acceptability using qualitative focus groups and a quantitative survey. Paired t-test and McNemar's test were used to compare within-individual changes in accuracy of perceived breast cancer risk. For the qualitative approach, a content analysis was conducted.

Results

Participants were mean age (53.4), Hispanic (61.8%) and low numerate (41.2%). According to the Gail breast cancer risk assessment tool (BCRAT), the mean 5-year and lifetime breast cancer risk were 1.11% (±0.77) and 7.46% (±2.87), respectively. After interacting with RealRisks, the difference in perceived and actual breast cancer according to BCRAT improved for 5-year risk (p=0.008), but not for lifetime risk (p=0.20). Accuracy of perceived breast cancer risk (defined as within ±5% of actual lifetime risk according to BCRAT) improved from 52% to 70% (p=0.10). Even in the subgroup of women with low numeracy, accurate risk perceptions improved from 45% to 70%. In particular, 4 out of 5 women who overestimated their lifetime breast cancer risk by >30% had accurate risk perceptions after exposure to RealRisks. Eighty-five percent of the participants responded that RealRisks was easy to use (90.91%), helped them understand breast cancer risk factors (87.88%), and was useful (87.88%). In addition, almost 80% became less worried about breast cancer after interacting with RealRisks (78.79%). More than 90% responded that RealRisks increased their knowledge of breast cancer (93.94%).

Conclusion

This study demonstrated a significant improvement in accuracy of perceived breast cancer risk after exposure to RealRisks in a multi-ethnic low-numerate population. Based upon feedback from our focus groups, we were able to identify information needed to fully represent the important issues of breast cancer risk to further develop our prototype for testing in a randomized controlled trial.

1. Ropka ME, Keim J, Philbrick JT. Patient decisions about breast cancer chemoprevention: a systematic review and meta-analysis. *J Clin Oncol.* Jun 20 2010;28(18):3090-3095.

2. Hau R, Pleskac TH, Hertwig R. Decisions from experience and statistical probabilities: why they trigger different choices than a priori probabilities. *J Behav Decision Making.* 2010;23:48-68.

Identifying barriers to using eHealth data for individualized clinical performance feedback in Malawi : A case study

Zach Landis-Lewis[1], Ronald Manjomo[2], Oliver Gadabu[2], Bertha Simwaka[3], Susan Zickmund[4], Gerald P Douglas[1], Rebecca S Jacobson[1]

[1]Center for Health Informatics for the Underserved, University of Pittsburgh, Pittsburgh, PA; [2]Baobab Health Trust, Lilongwe, Malawi; [3]The Global Fund to Fight AIDS, TB, and Malaria, Geneva, Switzerland; [4]Center for Health Equity Research and Promotion, VA Medical Center, Pittsburgh, PA

Introduction

Low-quality of health services and sub-optimal performance of healthcare providers in low-income countries is a critical global problem.[1,2] Audit and feedback (AF), defined as the provision of performance summaries to healthcare providers, is a commonly used intervention that can significantly improve clinical performance.[3] The use of electronic information technology in the delivery of healthcare (eHealth), which is growing rapidly in low and middle-income countries,[4,5] is creating unprecedented opportunities to provide individualized performance feedback to healthcare providers. Knowledge about barriers to the use of eHealth data for quality improvement is increasing,[6,7] but our understanding of how and when eHealth data can be successfully used to generate performance feedback is limited. An electronic medical record (EMR) system, deployed nationally in antiretroviral therapy (ART) clinics in Malawi, collects data that national ART supervision teams use to provide quarterly, clinic-level performance feedback.[8,9] The aim of this study was to identify and describe barriers to using EMR data for individual-level AF for healthcare providers in Malawi, and to consider how to design technology to overcome these barriers.

Methods

We used conceptual models for the cognitive processing of feedback[10] and implementation of evidence in clinical settings[11] to inform qualitative data collection. We distributed flyers in eight public hospitals in Malawi where the EMR is used, to recruit nurse and clinical officer participants for semi-structured interviews, observations of EMR use, and informant feedback. We interviewed a convenience sample of 32 participants (18 nurses, 14 clinical officers) in six district hospitals and two central hospitals. We audio recorded interviews and transcribed them verbatim. We collected field notes during observations and informant feedback meetings. To protect participants' rights we conducted interviews in private rooms, kept research data confidential, and did not document identifiable information. We analyzed the qualitative data using Crabtree and Miller's editing method,[12] using team input in the development of an iteratively derived codebook. We used an adjudication process to resolve differences between coders and interpreted the thematic findings using the lens of our conceptual models.

Results

We identified four barriers that can prevent eHealth data from being usable for AF in low-resource settings: provider rotations, disruptions to care processes, performance indicator stability, and user acceptance of eHealth. *Provider rotations* refers clinic staff schedules that prevent healthcare providers from working in a clinic long enough to receive quarterly performance feedback. *Disruptions to care processes* are unexpected events such as infrastructure failures, drug stockouts, and eHealth outages that are beyond the control of the individual and that limit improvement potential, reducing the relevance of performance feedback. *Performance indicator stability* refers to the average length of time that a performance indicator, once created, remains useful for measuring individual clinical performance. For example, performance indicators may become obsolete due to changes in EMR software or to the guideline recommendations on which they are based. *User acceptance*, concerning an individual's use of and attitudes toward the EMR, constrains the ability to conduct AF for individuals who do not use some EMR functions.

Discussion

To successfully use EMR data in Malawi, clinical AF interventions must adapt to provider rotations, disruptions to care processes, performance indicator stability, and user acceptance of eHealth. AF interventions may be improved by monitoring data quality errors associated with these barriers, and by failing gracefully when the quality of data becomes too low. To improve the effectiveness of feedback generated using eHealth data, a reporting tool could further facilitate adaptation by enabling a supervisor to select or tailor performance feedback messages, based on the supervisor's knowledge about the recipient and the clinical situation during the reporting period.

References

1. Holloway KA, Ivanovska V, Wagner AK, Vialle-Valentin C, Ross-Degnan D. Have we improved use of medicines in developing and transitional countries and do we know how to? Two decades of evidence. Trop Med Int Health. 2013 Jun;18(6):656-64.
2. Rowe AK, de Savigny D, Lanata CF, Victora CG. How can we achieve and maintain high-quality performance of health workers in low-resource settings? Lancet. 2005 Sep 17-23;366(9490):1026-35.
3. Ivers N, Jamtvedt G, Flottorp S, Young JM, Odgaard-Jensen J, French SD, et al. Audit and feedback: effects on professional practice and healthcare outcomes. Cochrane database of systematic reviews (Online). 2012;6:CD000259.
4. Piette JD, Lun K, Moura LA, Fraser HS, Mechael PN, Powell J, Khoja SR: Impacts of e-health on the outcomes of care in low- and middle-income countries: where do we go from here? Bulletin of the World Health Organization. 2012, 90(5):365–372 .
5. Lewis T, Synowiec C, Lagomarsino G, Schweitzer J: E-health in low- and middle-income countries: findings from the Center for Health Market Innovations. Bulletin of the World Health Organization. 2012, 90(5):332–340.
6. Weiner MG, Embi PJ: Toward Reuse of Clinical Data for Research and Quality Improvement: The End of the Beginning? Annals of Internal Medicine. 2009, 151(5):359–360 .
7. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PRO, Bernstam EV, Lehmann HP, Hripcsak G, Hartzog TH, Cimino JJ, Saltz JH: Caveats for the use of operational electronic health record data in comparative effectiveness research. Medical Care. 2013, 51(8 Suppl 3):S30–37.
8. Douglas GP, Gadabu OJ, Joukes S, Mumba S, McKay MV, Ben-Smith A, et al. Using Touchscreen Electronic Medical Record Systems to Support and Monitor National Scale-Up of Antiretroviral Therapy in Malawi. PLoS Med. 2010;7(8):e1000319.
9. Integrated HIV Program Report, January-March 2014. Government of Malawi, Ministry of Health. 17th July 2014.
10. Ilgen DR, Fisher CD, Taylor MS: Consequences of individual feedback on behavior in organizations. Journal of Applied Psychology. 1979, 64(4):349–371.
11. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC: Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implementation Science. 2009, 4:50 .
12. Crabtree BF, Miller WL: Doing Qualitative Research in Primary Care: Multiple Strategies. Sage Pubns. 1992.

A Curvilinear Path Towards Interoperability: A Curvilinear Path Towards Interoperability: Homogeneity and Disparity in Models of Health Information Exchange

James R. Langabeer II, PhD[1], Tiffany Champagne, PhD[2]

[1]University of Texas Health Science Center, Houston, TX; [2]Greater Houston Healthconnect, Houston, TX

Introduction

The benefits of sharing patient health information between providers have been well documented[1]. Improved care coordination, reduced duplication of tests and scans, and improved efficiencies are just a few of these benefits. Yet barriers and inconsistencies continue to plague the path towards interoperability. Institutional organization theory contends that organizations with similar intent and vision are similarly shaped by external forces and consequently tend to develop homogeneous strategies and become isomorphic[2]. In the case of health information exchange (HIE), billions of dollars were infused into community HIEs for a common purpose and intent, yet there has not been significant research into the extent to which they share similar strategies or approaches. In this research, we argue that the evolutionary path towards national interoperability has resulted in an adaptive change by local organizations to a barrage of barriers, which has created in effect a curvilinear trajectory. More clear, decisive action at the policy level could alter the course significantly. Using case study data from 25 regional health exchanges, we explore the disparity in organizational models in health exchanges nationwide.

Methods

This study employed qualitative methods, combining detailed informant interviews and case study data from 25 large, geographically distributed HIEs throughout the nation. The research question we sought to address centered on the degree of homogeneity in the approaches and organizational "models" being deployed by health exchanges throughout the country. Since institutional theory of organizations suggests that external forces and similarities of purpose drive organizations towards isomorphism (similarities in form and function), we would hypothesize a high level of similarity in the overall model. Yet, anecdotal evidence suggests that HIEs are evolving heterogeneously. To test our question, we created a framework of 8 factors that comprise key aspects of organizational strategy or models of HIEs. We collected and assessed data from a random sample of 25 large (10% of total population) exchanges for analyses focusing on these variables: 1) Organizational Size; 2) Reach/Scope; 3) Market Strategy; 4) Technical Architecture; 5) Maturity and Phase; 6) Vendor Platform; 7) Sustainability Model; 8) Service Offerings.

Results

We found evidence of widespread similarity in most market-based strategies and services, with some notable disparities. In terms of services offered, nearly all HIEs are approaching this homogenously, yet there was disparity in technical models and architectures. The strategy appears similar, the tactics slightly different. HIEs in this sample were relatively mature in age (mean of 6.6 years) but low in maturity (>50% not live in operational phase). Nearly all organizations (92%) have adopted a participation or transaction fee structure generated from a wide variety of stakeholders (providers, payers, systems) yet most have very little earned revenue to date outside of grant revenue. Most HIEs (72%) have adopted a hybrid or centralized data model approach, but are growing slower than anticipated due to data privacy concerns from stakeholders. A patient opt-in model of consent is the least risky, but more organizations (64%) have adopted an opt-out approach. Service offerings were largely limited to direct and continuity of care document (CCD) exchange.

Results

The seamless flow of interoperable data represents an achievable vision, but there should be greater sharing of best practices and more rigor in developing unique strategies that work in each local market. This homogeneity in models and strategy shows little local adaptation, but it also poses a significant threat to the overall path towards interoperability. Organizational theory suggest that the "planned change" of what was intended as outcomes for HIE investments has been morphed by the "adaptive change" to the reality of sustainability and local market barriers. Lacking stronger direction from policy-makers at the time of the large "stimulus" investments, very little

prescriptive assistance, compressed timelines, and lack of known best practices, HIEs have evolved in much the same way. This will create multiple failures, but also examples of likely success. Researchers should follow these strategies and performance to better understand which models are optimal, to ensure the long-term viability and success of interoperability in healthcare.

References

1. Walker J, Pan E, Johnston D, Adler-Milstein J, Bates D, and Middleton B. The value of health care information exchange and interoperability. Health Affairs 2005: W5-W18.
2. DiMaggio P and Powell W. The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. American Sociological Review 1983, 42(2), 147-160.

Combining Heterogeneous Databases to Detect Adverse Drug Reaction

Ying Li[1], Santiago Vilar[1], Ying Wei[2], Carol Friedman[1]

1. Department of Biomedical Informatics, 2. Department of Biostatistics, Columbia University, New York, NY, 10032

Introduction

Early detection of adverse drug reactions (ADRs) is critical for advancing patient safety. The rapid growth of large electronic health records (EHRs) provides an opportunity to use the information in the EHRs to hasten ADR discovery. Currently, most studies of ADR detection have either been limited to use of a single data resource to discover an ADR and then have independently used another resource for validation, or required that the ADR occurs in the two resources leading to low sensitivity.[1,2] We hypothesize that integration of evidence from multiple resources should lead to a more effective ADR detection system, and propose a methodological framework that computationally combines the adverse event reporting system of the Food and Drug Administration (FAERS)[3] with information in EHRs as well as external knowledge for ADR detection.

Methods

Our study is based on 2.7 million FAERS reports and 0.3 million EHR patients from 2004 to 2010. The framework includes three steps in Figure 1: 1) Obtaining statistics in each resource for drugs associated with an ADR of interest while adjusting for confounding effects; 2) Determining the p-values based on the empirical null distribution derived from a set of reference negative drugs and combining two p-values with weights inversely proportional to the variance associated with the statistic for that resource, i.e., more weight is assigned when variance is small; 3) Re-ranking initial signals based on similarity of their adverse drug event (ADE) profile[4] based on the Side Effect Resource (SIDER) database.[5]

The framework was applied to four serious ADRs, acute renal failure (ARF), acute liver injury (ALI), acute myocardial infarction (AMI), and gastrointestinal bleeding (GIB), and was evaluated according to a published reference standard (RS) containing known positive and negative ADRs based on available knowledge.[6] The combined statistics in step 2 is compared with statistics from each resource and from a random system based on the RS using area under the receiver operating characteristic (ROC) curve (AUC) and its relevant significance test. We also report precision and recall at the well-known cutoff value, which is 0.05 for one-sided pvalues of EHRs and FAERS and for the combined pvalue of the combined system, as well as AUC for each ADR. The re-ranking results from step 3 are then evaluated using top K precision.

Results

We restricted our positive and negative control test cases to those that had at least one person exposed and diagnosed with an ADR of interest in the EHR and at least one case report in FAERS, resulting in 104 positive and 176 negative controls. The results show that the combined system achieved an AUC of 0.73, which is not significantly different from 0.77 based only on FAERS, but is significantly better than 0.51 based only on EHRs and better than 0.53 based on a random system. The performance for the 4 ADRs are shown in Table 1. The recall of the combined system was better than or the same as the two individual resources for all 4 ADRs. Among them, ARF achieved the best recall while ALI had the worst. However, the AUC and precision of the combined system was between the individual systems.

On average, FAERS contributed 69% and the EHRs contributed 31% to the combined p-value based on variance. Table 2 shows true signals (TP) that were only detected by the combined system. Nevertheless, the combined system did not identify two TP, trandolapril and interferon beta-1a associated with ALI, which were detected by FAERS alone. There were no TPs detected using the EHR alone. We only re-ranked initial signals for which we had ADE profile information, amounting to 45 positive and 21 negative controls. Table 3 shows that re-ranking always achieved better precision in top positions of signals compared with the combined system.

Discussion and Conclusion

The main limiting factor of our study was the small EHR population we had access to, limiting EHR signal detection capability, and impairing performance of the combined system. In future work, we plan to include additional EHR data from multiple sites.

To our knowledge, the framework presented here is the first that aims to computationally combine diverse information from three different resources for ADR detection. Our preliminary results demonstrate that combining FAERS with EHR data using a computational approach was effective in achieving improved sensitivity, which is crucial for early detection. Furthermore, enhanced precision was obtained by re-ranking the signals using additional external knowledge.

Figures and Tables

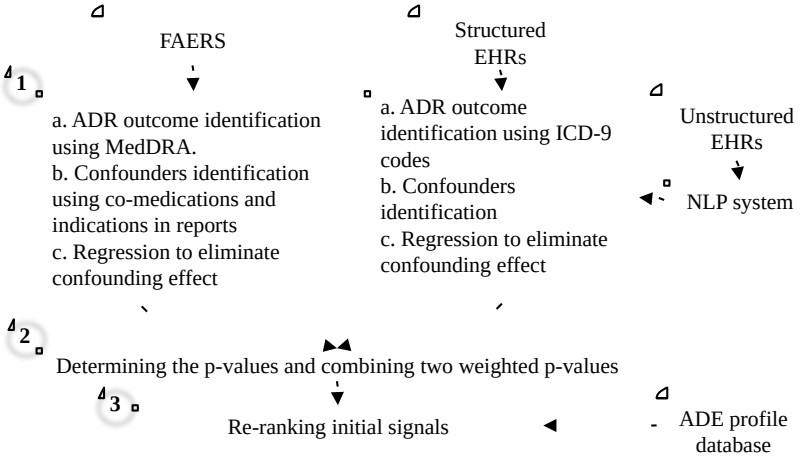


Figure 1 Methodological framework for generating, combining and re-ranking signals

Structured EHR data in this context refers to International Statistical Classification of Diseases (ICD-9) which are linked to each unique patient.

Unstructured EHR data in this context refers to EHR narratives, which are processed using the natural language processing (NLP) system.

ADR outcome is defined by Health Outcomes of Interest library provided by Observational Medical Outcomes Partnership (OMOP).

Table 1. Recall and precision at the cutoff value, and AUC of three systems for 4 ADRs

ADR	RS		Recall			Precision			AUC		
	P	N	FAERS	EHRs	COM	FAERS	EHRs	COM	FAERS	EHRs	COM
ARF	14	54	0.43	0.5	**0.79**	1.00	0.54	0.69	0.90	0.69	0.90
GIB	18	52	0.50	0.06	**0.61**	0.60	0.13	0.52	0.81	0.57	0.76
ALI	53	25	0.47	0.02	**0.47**	0.89	0.5	0.86	0.73	0.58	0.73
AMI	19	45	0.37	0.26	**0.58**	0.88	0.42	0.65	0.68	0.47	0.68

RS: reference standard; P: positive controls; N: negative controls; FAERS: signal detection system using FAERS alone; EHRs: signal detection system using EHRs alone; COM: combined system

Table 2. True signals detected only by the combined system

ARF	GIB	ALI	AMI
etodolac, captopril, piroxicam	clindamycin	felbamate	NA

Table 3. Top K precision of two systems based on the subset of the initial signals

ADR	ARF (11/5)		GIB (11/9)		ALI (19/3)		AMI (4/4)	
Top	COM	RER	COM	RER	COM	RER	COM	RER
5	1.00	**1.00**	1.00	**1.00**	1.00	**1.00**	0.80	**0.80**
10	0.80	**1.00**	0.90	**0.90**	1.00	**1.00**	NA	NA
15	0.73	**0.73**	0.67	**0.73**	0.93	**1.00**	NA	NA
20	NA	NA	0.55	**0.55**	0.90	**0.95**	NA	NA

The number inside the bracket indicates true positives and false positives in the subset; COM: combined system; RER: re-ranking signals using profile extracted from SIDER resource[5].

Reference

1. Tatonetti N, Denny J, Murphy S, et al. Detecting drug interactions from adverse-event reports: interaction between paroxetine and pravastatin increases blood glucose levels. *Clinical Pharmacology & Therapeutics*. 2011;90(1):133-142.
2. Harpaz R, Vilar S, DuMouchel W, et al. Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions. *Journal of the American Medical Informatics Association.* 2013;20(3):413-419.
3. FDA Adverse Event Reporting System (FAERS). http://www.fda.gov/cder/aers/default.htm.
4. Liu M, Wu Y, Chen Y, et al. Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs. *Journal of the American Medical Informatics Association.* 2012;19(e1):e28-e35.
5. Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P. A side effect resource to capture phenotypic effects of drugs. *Molecular systems biology.* 2010;6(1).
6. Ryan PB, Schuemie MJ, Welebob E, Duke J, Valentine S, Hartzema AG. Defining a reference set to support methodological research in drug safety. *Drug safety.* 2013;36(1):33-47.

Detecting Epilepsy Diagnosis in Clinical Notes: A Comparison of Traditional Text Classification Methods

Todd Lingren, MS, Pawel Matykiewicz, M.Phil., Yizhao Ni, PhD, Shannon M Standridge, MD, Katherine D Holland, MD, Imre Solti, MD, PHD, MA, Tracy A Glauser, MD, John P Pestian, MBA, PhD

Cincinnati Children's Hospital Medical Center (CCHMC), Cincinnati, OH

Introduction and Background

Natural language processing (NLP) is essential to exploit the information in Electronic Health Records (EHRs) that may identify many diseases, including epilepsy. Several information extraction systems, such as the clinical Text and Knowledge Extraction System (cTAKES)[1] have been developed to make the Unified Medical Language System (UMLS) more accessible for biomedical NLP research. In this study we evaluated the effectiveness of using UMLS for the identification of 37 different epilepsy-related diagnosis and treatment markers from neurology clinic notes. We then compared this method with traditional text-extraction methods that used manually annotated clinical data.

Data and Methods

Two hundred and sixty-seven neurology clinic notes from 51 patients were selected from a previous study (575.67 tokens per note on average)[2]. Two different feature types were selected for comparison. The first feature set was created via traditional method of text classification, using unigrams, bigrams, and trigrams. We applied the National Library of Medicine stopword list[3]. All words were lower-cased, all numerals were substituted with the string NUMB for abstraction, and all non-ASCII characters were removed. The notes included 7718 unique ngrams. The second feature set was based on UMLS Concept Unique Identifiers (CUIs). CUIs were obtained using cTAKES (ver. 3.0) to process the clinical notes. cTAKES employs machine learning and regular expression UMLS dictionary lookup to obtain CUIs from clinical narratives. The feature set included 624 unique CUIs. Negated and non-negated CUIs were counted as distinct features.

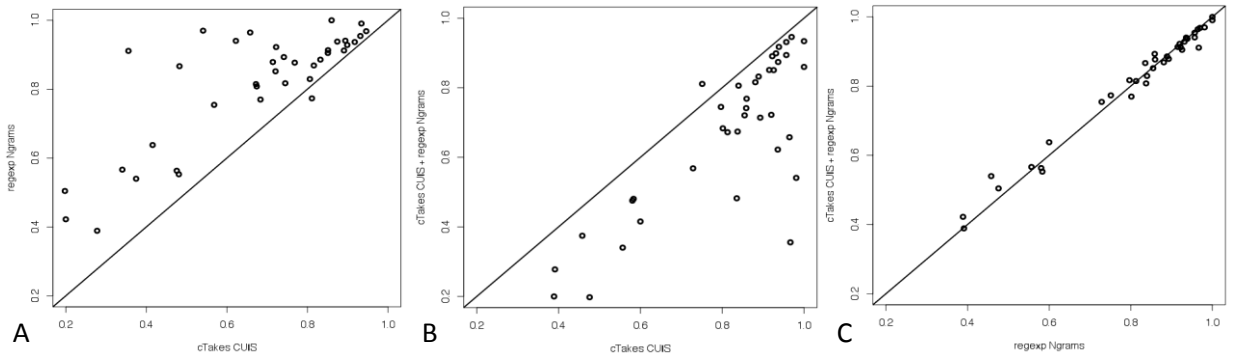
The 37 different epilepsy-specific treatment and diagnosis markers (e.g., partial epilepsy, generalized epilepsy, specified seizure past frequencies, unsatisfactory control reasons, and various drug treatment types) comprise the gold standard labels and were added by three annotators. Each note can belong to multiple classes. Average inter-annotator agreement was 0.816 (Krippendorf's alpha). Classification was performed using cosine kernel SVM one-vs-rest in 15-fold cross-validation and multidimensional grid search parameterization. We scanned three feature selection methods using rigorous internal and external cross-validation techniques[4]: information gain (IG), IG ratio, and Pearson correlation coefficient; four feature weighting functions: identity function, IG, IG ratio, and Pearson correlation coefficient; cost (C) was screened at $2^{0.5}$ increments from 1-32; percent of retained features was also done using 21 point scale. For each label the grid had total 5292 cells[5]. The average number of features used in each classification was different between feature sets. CUI features used an average of 258 features (min 26/ max 515), 2644 ngrams (280/6235), and 2167 ngrams+CUIs (376/5778).

Results and Discussion

Performance was calculated using F1 measure (macro average over all classes). CUIs performed worse (0.672, sd 0.222) (Figure 1A) and ngrams features were best (0.815, sd 0.170) (Figure 1B). Using CUIs and ngrams together performed slightly worse than just ngrams (0.813, sd 0.180, Figure 1C). Paired T-test demonstrated that classification with CUIs is statistically significantly worse than ngrams (p-value 4.734e-08) and worse than ngrams+CUIs (p-value 7.323e-08). The difference between performance of ngrams and ngrams+CUIs was not statistically significant (p-value 0.6821). Features that rank in the top ten significant features more than ten models are shown in Table 1. Six of the top ten features in ngrams+CUIs set are similar to top-10 CUI or identical to top-10 ngram features. Top ten features were calculated using absolute values of support weight vector.

Traditional approach to text classification performed well when applied to detection of diagnostic and treatment-specific markers in clinical notes. Like previously suggested[6], using UMLS CUIs did not improve the classification result in a narrow domain. CUIs seem to represent clinical concepts well (Table 1), demonstrating promise for using CUIs a higher level meaning abstraction. Including CUIs did reduce the average number of features used in classification, which can mitigate possible overtraining. Future work is necessary to augment UMLS CUIs with additional domain-specific knowledge and consider other feature types such as medication codes (e.g. RxNORM)

for targeted classification of specific drug-related markers. Additional future work will focus on replicating the results on a larger data set.



CUIS *concept unique identifiers*, regexp *regular expression*

Figure 1. Performance (F1 measure) of A) CUIs versus ngrams, B) CUIs vs. ngrams+cuis, and C) ngrams vs. ngrams+cuis

CUI features		# Models	ngrams	# Models	ngrams+CUIs	# Models
C0751495	Seizures, Focal	31	levetiracetam	17	cryptogenic[1,2]	13
C0037769	Infantile spasms	22	carbamazepine	14	epilepsy[1,2]	13
C0234378	Static Tremor	21	epilepsy	13	idiopathic[1,2]	13
C0009946	Conversion disorder	19	ma (VNS milliampere)	13	schedule	13
C0270850	Idiopathic generalized	19	medication	13	symptomatic.localization[2]	13
C0018674	Craniocerebral Trauma	18	episode	12	symptomatic.localization.related[2]	13
C0472354	Cryptogenic	17	lesion	12	diet	12
C0043096 NEGATED	Body weight decreased	16	symptomatic.localization	12	family.decided.increase	11
C0014547	Epilepsies, Partial	15	symptomatic.localization.related	12	generalized	11
C0879626	Adverse effects	15	acid.NUMB.NUMB	11	seizures[1]	11

[1]CUI feature similarity overlap [2]ngram feature overlap

Table 1. Number of models that had a particular feature in the top-ten list of features.

References

1. Savova, GK, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. JAMIA 2010;17(5): 507-513.
2. Matykiewicz, P, et al. Earlier identification of epilepsy surgery candidates using natural language processing. BioNLP Worskhop at Association for Computational Linguistics Annual Meeting 2013.
3. http://mbr.nlm.nih.gov/Download/2009/WordCounts/wrd_stop
4. Matykiewicz, P, Pestian, J. Effect of small sample size on text categorization with support vector machines. In Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP '12). Association for Computational Linguistics, Stroudsburg, PA, USA, 193-201.
5. Matykiewicz, P, et al. Comparison of corpus linguistics and machine learning techniques in determining differences in clinical notes. Submitted for peer review.
6. Demner-Fushman, D, Chapman, W, McDonald, CJ. What can natural language processing do for clinical decision support?. Journal of biomedical informatics 2009;42(5):760-772.

Authors' Contributions

TL and PM participated in preparing the data, PM also participated in programming the annotation software, and calculating the statistics; SS (MD epilepsy) participated in annotation schema design, arbitration; KH (MD epilepsy) participated in annotation schema design, arbitration; YN contributed with ideas for classification; JP conceptualized, and guided the project. All authors participated in drafting the manuscript. All authors read and approved the manuscript.

ED Noise & Cognition Interruptions: Do We Have a Jackhammer in the Cockpit?

Mary L. Little, RN, MSN[1], Osman R. Sayan, MD[2], Edward H. Suh, MD[2], Vimla L. Patel, PhD[1,3]

[1]Dept. of Biomedical Informatics, Columbia University; [2] New York-Presbyterian Hospital; [3]New York Academy of Medicine, New York, NY

Introduction

Hospital noise level research has shown sound pressure levels (SPLs) have risen 0.38 dBs each year since 1960.[1] However, few studies address this trend and its effect on patients, staff, communication, and quality of care. The impact of excessive noise on cognitive functioning influences the use of technology. This trend of increasing SPLs affecting communication and cognition, calls for human factors engineering and environmental innovations to reduce noise, maximize user oriented information retrieval, within the electronic health record (EHR), and improve staff and patient experiences.

Methods

An integrated literature review from February–March 2014, produced over 2300 SPL citations which reduced to 13 articles when combined with ED or ICU keywords. Ancestor searches produced 18 additional articles. This pilot investigation of SPLs in the ED is a part of a larger baseline workflow information study of 8 interviews, and over 34 ethnographic shift-change observations with ED attending and resident level physicians. An iPad mini was transformed into a sound meter via SPL meter apps made by SPLnFF and AudioTools; observed decibel readings (dBs), in an A-weighted frequency were causation coded. Sound booth testing to compare app accuracy and effect of an external iMM-6 iDevice Calibrated Measurement Microphone was performed. The iPad mini was held away from the observer's body by 12-18 inches to minimize interference in SPL readings. An extension of this pilot study is underway with greater measurement precision using a digital sound level meter.

Results

Several articles note an average of 73-77 dBs in ED and operating rooms, with occasional spikes of 90 dBs or more. SPLs greater than 60 dBs are associated with an increase in stay length for cataract patients,[2] decrease in mental efficiency, increase in communication based errors, and staff turnover. In the interviews, EHR tasks and inefficiencies were cited as major concerns. ED noise levels were only mentioned once, indicating the possibility of user belief in adjustment or acceptance of the ED noise levels despite literature citing the activation of the sympathetic-adrenal-cortical axis and production of stress hormones.[3] Sound booth testing showed the iMM-6 microphone falsely raised the SPL recordings by 7-8.3 dBs. Without the iMM6 microphone, average ED noise levels were 73.4 dBs. SPL recordings varied among the 3 ED areas based on patient capacity and proximity to the main corridor walking path. The loudest SPL readings occurred with overhead announcements, averaging 87-91 dBs, with a SPL_{max} reading of 108.1.

Discussion

Our initial data show that SPLs >95-100 elicited verbal complaints during the rounding process. These SPLs correspond to subway noise and occasionally a construction jackhammer at 3 feet away. Noise of this magnitude, at a critical time in team transition, affects information integrity, cognition and decision-making. There are several limitations in this pilot study which need to be addressed in future studies. While the Apple platform is extremely stable the apps may vary. Although both apps are highly rated,[4] only the AudioTools app has a recording feature. Additionally, more SPL measurements in various types of EDs and studies of the precise effects on provider cognitive functions are needed for a more comprehensive evaluation of ED noise and effect on technology use.

Sound Associations	dB Level	Hospital Noise Found at this Level
Washing Machine	40	None
Vacuum Cleaner at 1 meter	70	Staff conversations (avg), patient alarms
Garbage Disposal	80	Raised voices
Police Whistles, Subway	90	Paging announcements (avg)
Jackhammer at 1 meter	100+	Paging announcements (max)

Table 1 Comparison of Sounds. EPA hospital max is 45 dbs.

Conclusion

Aviation's sterile cockpit rule prohibits non-essential distractions during critical flight times. However, medicine has not emphasized solutions for the distracting and potentially harmful effect of excessive SPLs on cognitive functions of patients, staff, and workflow communications. This preliminary study using sound meter apps as an effective exploratory tool, highlights the need for more research in this area including interactions with EHRs for provider information retrieval, effective alerts, and safe care team transitions. Use of noise reduction solutions such as sound absorbing panels and volume limiters on audio announcements and other communication devices could be implemented. We need to keep the jackhammers out of our cockpits during the takeoff and landing period of patient handoffs in care team transition.

References

1. Busch-Vishniac IJ, West JE, Barnhill C, Hunter T, Orellana D, Chivukula R. Noise levels in Johns Hopkins hospital. The Journal of the Acoustical Society of America. 2005;118(6):3629-45..
2. Fife, D., & Rappaport, E. (1976). Noise and hospital stay. American Journal of Public Health, 66(7), 680-681.
3. Babisch W. The noise/stress concept, risk assessment and research needs. Noise and health. 2002;4(16):1.
4. Kardous CA, Shaw PB. Evaluation of smartphone sound measurement applications). The Journal of the Acoustical Society of America. 2014;135(4):EL186-EL92

Acknowledgment: Funded by 5 T15 LM007079. Research funded by James S. McDonnell Foundation (Grant No. 220020152 to Vimla L. Patel)

A Hybrid Electronic Surveillance Design Pattern for Public and Population Health

Authors John W. Loonsk, MD FACMI, Hadi Kharrazi MD PhD, Jonathan Weiner DrPH, The Center for Population Health IT, Johns Hopkins Bloomberg School of Public Health

Introduction

The integration of information technology into public health surveillance has proceeded, as in many other industries, as a combination of both the digitalization of existing paper work processes and the development of new work processes where information technology has presented opportunities for rapid advancement. Current electronic surveillance methodologies represent only an initial step in the digitalization of these public health systems and redundant and competing surveillance investments are plentiful. This analysis of existing electronic surveillance approaches has identified the elements of a new hybrid design pattern to improve integration, minimize provider infrastructure needs, and improve overall public health reporting.

Methods

An analysis of public and population health surveillance needs and existing health information technology surveillance applications and implementations was performed. Syndromic surveillance, biosurveillance, registry, case reporting, and reportable condition reporting systems were all analyzed for common information technology needs and elements. Current and developing data and technology standards were also studied and common attributes were documented.

Results

The analysis of existing surveillance approaches and applications identified numerous common elements and complementary needs. Several issues were identified that contributed to the current lack of integration and this picture for electronic public health surveillance systems. And surveillance in the U.S. can still be incomplete and untimely. There are variations in reporting regulations in different States, categorical disease programs with differing data expectations and standards, and changing case definitions in public health emergency settings. More recently, there are also health department IT resource challenges in the face of the rapid advancement of Electronic Health Records (EHRs) in clinical care and newly developing clinical care population health management activities from the Affordable Care Act (ACA) and other new payment methods.

Discussion

This analysis has identified a new hybrid surveillance digital design pattern that can address current electronic surveillance issues. The design pattern melds elements from syndromic surveillance and traditional case reporting. It uses trigger tables for the automated initiation of sensitive, but non-specific data reporting, and centralized decision logic for the identification of more specific, jurisdictional and condition-oriented reports. Because the decision logic is provided centrally, it can be more readily maintained and updated in a rapidly changing emergency situation. Since the initial reporting is automatically generated, there is the opportunity to address the historic yield issue from clinical care

case reporting. Automation of follow-up to that early indication can minimize or eliminate additional work expectations for health department staff.

Crowdsourcing ICD-11 Sanctioning Rules

Vincent Lou[1], Samson W. Tu MS[1], Csongor Nyulas MS[1], Tania Tudorache PhD[1], Robert J. G. Chalmers MB FRCP[2], Mark A. Musen MD PhD[1]

[1]Stanford University, Stanford, CA, USA, [2]The University of Manchester, Manchester, UK

Abstract

It is difficult and time-consuming for medical professionals to develop and maintain large biomedical ontologies. In this work, we developed a crowdsourcing method for obtaining sanctioning rules for the post-coordination of concepts in the draft ICD-11. Our method utilizes the hierarchical structures in the domain to improve the accuracy of the sanctioning rules and lower the cost of the crowdsourcing process. We used a Bayesian network to model workers' skills, the accuracy of their responses, and our confidence in the acquired sanctioning rules. Furthermore, we developed a method to maximize the quality of the sanctioning rules within a fixed budget for paying workers.

Introduction

The draft ICD-11[1] includes the capability for coders to specify the details of a disease using a combination of ICD codes and qualifiers such as a disease's severity or anatomic location. To ensure the integrity of coded data, such post-coordination requires *sanctioning rules* that specify legal combinations of codes and qualifiers[2]. For a large classification like ICD-11 that has tens of thousands of codes and thousands of terms in the value sets of qualifiers, manual curation of sanctioning rules by subject matter experts (SMEs) poses a formidable challenge. We developed a crowdsourcing method to generate a high-quality draft of the sanctioning rules that SMEs can verify and edit.

Methods

Crowdsourcing involves three steps: 1) defining the microtasks, 2) dividing the overall task into microtasks and publishing the microtasks on a crowdsourcing platform, and 3) collecting and combining answers[3]. Given a disease, we define a microtask as asking anonymous workers on the Amazon Mechanical Turk platform[4], in exchange for a small payment, to answer multiple-choice questions about possible anatomic locations of the disease. For the overall task of obtaining sanction rules for all diseases, we utilize the hierarchy structure of both diseases and anatomic locations to minimize the number of necessary questions. For example, if the disease cannot occur at a location, we don't need to explore the location's children. When we are acquiring the sanctioning rules for a child disease of a previously constrained parent, we start searching from the permitted anatomic locations of the parent instead of top-level locations while giving the workers the option to indicate that other anatomical locations are possible. We use Bayesian networks to model the confidence level on a sanctioning rule, given available workers responses. Initially we asked for responses from 10 workers for each question. With the Bayesian network, we stop asking for more workers once a confidence cutoff is reached. Modeling each worker's accuracy rate further reduces the number of responses needed for a given cutoff. Finally, to maximize the accuracy of sanctioning rules within a given budget, we developed methods to trade accuracy against cost by using reinforcement learning to develop an agent who constantly adjusts the confidence cutoffs during the crowdsourcing process.. Before obtaining sanctioning rules for a new disease, the agent updates the cost distributions for each probabilistic cutoff, compare the result with remaining budget, and choose the 'best' cutoff to maximize a reward function that rewards the number of "finished" diseases and penalizes the number of "unfinished" diseases that can't be tested because of budget limitation.

Results

We performed two feasibility studies using 11 sub-branches of skin diseases in the draft ICD-11. First we used hierarchical pruning techniques to generate microtasks and the 'majority vote' method to aggregating responses. Second we re-analyzed the data obtained in the first experiment using the Bayesian and budget-constraint methods to aggregate answers and to set the stopping confidence levels. Compared to a gold-standard set of sanctioning rules provided by an expert dermatologist, the sanctioning rules crowdsourced in the first experiment have sensitivity and specificity of 0.81 and 0.97 respectively. For the second experiment, on questions where cutoffs can be adjusted, a 69% cost saving was achieved with an accuracy loss of 6%.

Discussion

Our experiments show that crowdsourcing, incorporating Bayesian networks and machine-learning techniques, is a promising method to efficiently acquire sanctioning rules for ICD. Our method can be applied to acquire the constraints between any two class hierarchies.

References

1. World Health Organization. The International Classification of Diseases 11th Revision is due by 2017. 2014 [cited 2014]. Available from: http://www.who.int/classifications/icd/revision/en/.
2. Navas H, Lopez Osornio A, Gambarte L, Elias Leguizamon G, Wasserman S, Orrego N, et al. Implementing rules to improve the quality of concept post-coordination with SNOMED CT. Stud Health Technol Inform. 2010;160(Pt 2):1045-9. PubMed PMID: 20841843.
3. Mortensen JM, Musen MA, Noy NF. Crowdsourcing the verification of relationships in biomedical ontologies. AMIA Annu Symp Proc. 2013;2013:1020-9. PubMed PMID: 24551391. Pubmed Central PMCID: 3900126. Epub 2014/02/20.
4. Amazon. Amazon Mechanical Turk 2014 [cited 2014]. Available from: http://aws.amazon.com/mturk/.

Quantifying Information Redundancy in Common Laboratory Tests

Yuan Luo, MS[1], Jason Baron, MD[2], Peter Szolovits, PhD[1], Anand Dighe, MD[2]
[1]Massachusetts Institute of Technology, Cambridge, MA, USA
[2]Massachusetts General Hospital, Boston, MA, USA

Introduction: Traditionally, redundant testing can be defined as tests that clinicians cancel after becoming aware of prior results [1]. However, other factors such as bundled tests from automated instruments increasingly contribute to redundant common laboratory tests, which are thought to lead to expanding health care costs or clutter the diagnostic picture at the presence of multiple inaccurate or even inconsistent results. We hypothesize that there may be a broader source to redundant testing – information redundancy. We characterize information-redundant tests as providing diagnostic or clinical information that largely overlaps with other concurrent tests on the same patient. Understanding the potential redundancies in various laboratory tests may be useful in enhancing test selection and result interpretation. In this study, we investigated the degree of information redundancy for erythrocyte sedimentation rate (ESR), given the results of other tests commonly ordered alongside ESR. The ESR is a non-specific marker of acute phase reactant and can be increased in a variety of inflammatory as well as some non-inflammatory conditions. In this report, we predict ESR from other laboratory results, using predictability as an indicator of information redundancy.

Methods: This study used data from inpatient, outpatient and emergency department testing at the Massachusetts General Hospital for specimens collected over a 3 month period with approval from the Institutional Review Board. The final dataset included ESR results and any results from 38 other selected tests that were performed on the same accession (collection) as each ESR test. The 38 others tests, as listed in Table 1, were primarily selected to comprise the tests most commonly performed alongside ESR and included C-reactive protein, CBC parameters and routine chemistry testing. We also included patient age and gender in our dataset and excluded ESR results that were associated with no other laboratory tests.

The final dataset consisted of data from a total of 6486 test collections, which is randomly split into a training set of 4541 collections and a test set of 1945 collections according to a 7:3 ratio. Besides ESR, each collection had a median of 28 of the 38 other tests measured (20-37, IQR). We applied a two-step process to predict ESR. The first step used multiple imputation (MI) to infer results on any of the 38 other tests that were not performed on any particular collection. The second step used regression to predict ESR from age, gender and imputation-completed results of the 38 other tests. We experimented with different MI methods including unconditional mean imputation (Mean), Multivariate Imputation by Chained Equations (MICE) [2], and a random forest based MI algorithm, MissForest [3]. After imputation, we also experimented with different regression methods to predict ESR, including linear regression (LR), Bayesian linear regression (BR), random forest regression (RFR) and Lasso.

Results: Table 2 shows the performance of different imputation-regression combinations on the training and held-out test set respectively. The best performance on held-out test data is achieved at an R^2 (coefficient of determination) of 0.702 when using MICE to impute each variable conditioned on only highly correlated other variables (correlation ≥ 0.25) followed by using random forest regression on the imputation-completed test results, and demonstrates the presence of substantial information redundancy. Note that the random forests based algorithm, when used in either imputation or regression, tends to overfit. This can be seen by comparing the training-testing performance discrepancies of MissForest with other imputation methods, as well as by comparing that of random forest regressions with other regression methods.

Discussion: To put the above results into context, we randomly shuffled the ESR values. The R^2 between shuffled ESR and original ESR is 0.001 ± 0.000. The $|R^2|$ for imputation-regression on the shuffled ESR all have mean ≤ 0.05 and confidence interval ≤ 0.005. These data corroborated our hypothesis that the ESR provides information in at least some patients that overlaps substantially with the results of other concurrent tests. We selected ESR as a proof-of-concept, as it would likely provide information that substantially overlaps with other common tests. We plan to apply similar techniques to a wide-range of tests and we envision this approach will ultimately provide a useful foundation for clinical decision support. For example, defining a patient-specific "pre-test" distribution of results for each test, given the patient's existing test results, may be informative in deciding which tests to order. In addition, it will be useful to correlate imputed results with clinical information and outcomes to assess the possibility that imputed ESR may actually provide a better assessment of physiologic or pathologic states than measured ESR.

References

[1] A. K. Jha, D. C. Chan, A. B. Ridgway, C. Franz, and D. W. Bates, "Improving safety and eliminating redundant tests: cutting costs in US hospitals," *Health affairs*, vol. 28, pp. 1475–1484, 2009.

[2] S. Buuren and K. Groothuis-Oudshoorn, "MICE: Multivariate imputation by chained equations in R," *Journal of statistical software*, vol. 45, pp. 1–67, 2011.

[3] D. J. Stekhoven and P. Bühlmann, "MissForest - non-parametric missing value imputation for mixed-type data.," *Bioinformatics*, vol. 28, pp. 112–118, 2012.

Table 1 Description of explaining variables used for ESR regression.

Variable Description	Variable Description
Patient age	Mean cell volume
Patient gender	Percent monocytes
Absolute basophil count	Percent neutrophils
Absolute eosinophil count	Percent nucleated red blood cells
Albumin level	Absolute nucleated red blood cells
Alkaline phosphatase level	Plasma anion gap
Absolute lymphocyte count	Plasma blood urea nitrogen (BUN) level
Absolute monocyte count	Plasma chloride level
Absolute neutrophil count	Plasma creatinine level
Percent basophils	Plasma glucose level
Plasma calcium level	Plasma potassium level
C reactive protein (CRP) level	Plasma bicarbonate level
Percent eosinophils	Platelet count
Globulin (plasma) level	Plasma sodium level
Hematocrit	Red blood cell (RBC) count
Hemoglobin level	Red blood cell distribution width
Percent lymphocytes	Aspartate transaminase (AST) level
Mean cell hemoglobin	Alanine transaminase (ALT) level
Mean cell hemoglobin concentration	Total bilirubin (plasma) level

Table 2 Regression results with different imputation-regression combinations on training and held-out test data sets respectively. Performance metrics include mean square error (MSE) and coefficient of determination R^2. In MICE-full, imputing every variable is conditioned on all the rest variables. In MICE-sel, imputing each variable is conditioned on only highly correlated other variables (correlation ≥ 0.25). MICE and MissForest are run 100 times and their results are in the format of mean $\pm$ 0.95 confidence interval. Best performance is highlighted with bold face.

Methods		Training			Testing		
Reg	Imputation	MSE	R^2	Correlation	MSE	R^2	Correlation
LR	Mean	0.070	0.581	0.762	0.074	0.554	0.744
	MICE-full	0.069±0.000	0.585±0.001	0.765±0.001	0.073±0.000	0.558±0.003	0.748±0.002
	MICE-sel	0.055±0.000	0.674±0.001	0.821±0.001	0.058±0.001	0.651±0.003	0.807±0.002
	MissForest	0.038±0.000	0.771±0.000	0.878±0.000	0.069±0.000	0.587±0.001	0.774±0.000
BLR	Mean	0.070	0.579	0.761	0.072	0.568	0.754
	MICE-full	0.070±0.000	0.581±0.001	0.762±0.001	0.070±0.000	0.576±0.002	0.759±0.001
	MICE-sel	0.055±0.000	0.673±0.001	0.820±0.001	0.054±0.000	0.674±0.001	0.821±0.001
	MissForest	0.039±0.000	0.767±0.000	0.876±0.000	0.068±0.000	0.593±0.001	0.777±0.000
RFR	Mean	0.010	0.941	0.977	0.067	0.594	0.771
	MICE-full	0.009±0.000	0.944±0.000	0.977±0.000	0.064±0.000	0.618±0.002	0.786±0.001
	MICE-sel	**0.007±0.000**	**0.955±0.000**	**0.981±0.000**	**0.050±0.000**	**0.702±0.002**	**0.838±0.001**
	MissForest	0.005±0.000	0.967±0.000	0.985±0.000	0.075±0.000	0.550±0.001	0.755±0.001
Lasso	Mean	0.072	0.567	0.753	0.073	0.564	0.751
	MICE-full	0.071±0.000	0.575±0.001	0.758±0.001	0.071±0.000	0.573±0.002	0.757±0.001
	MICE-sel	0.056±0.000	0.665±0.001	0.815±0.001	0.055±0.000	0.671±0.001	0.820±0.001
	MissForest	0.042±0.000	0.748±0.000	0.866±0.000	0.069±0.000	0.587±0.001	0.769±0.000

Evaluate the Effectiveness of Mobile Health Intervention Program for the Senior Population Suffering from Hypertension and Hypercholesterolemia in Taiwan

Wu M.P, Director, MN[1], Chang P.L, Prof, PhD[2], Lee C.H, MD, PhD[3]
[1]Department of Nursing, Taipei City Hospital, Zhongxing Branch, Taiwan; [2]Institute of Biomedical Informatics, National Yang-Ming University, Taiwan; [3]Department of Surgery, Taipei Veterans General Hospital, Taiwan.

Introduction

The objective of this study is to evaluate the effectiveness of a mobile health program targeting seniors with a high risk of hypertension and hypercholesterolemia. Taiwan has gradually become a country with a high percentage of senior citizens. With changes in diet and living habits, the number of seniors suffering from hypertension and hypercholesterolemia is on the increase. In recent years, the use of mobile medical application for self-healthcare management has been increasing among the senior population. In this mobile Health program, the participants would be learning how to use the mobile phone and the medical healthcare application (App) for updating their relevant personal health records (for instance: measure blood pressure, pause, physiologic index, time of taking medicine, Chronic Disease Management etc.) at home. Call center will send a text message to the participants if any unusual was found in the uploaded data. This medical application system has reached the efficacy of disease prevention, health alert and early treatment by enabling participants to measure and control their own health, thus reducing unnecessary needs for medical consultations, which alternately contributes to the more efficient use of limited medical resources. It is for the aforementioned reasons that this study can be regarded as innovative and influential.

Method

The study samples were selected from residents living in the Shilin District of Taipei City, Taiwan, aged from 50 to 70 year- old adults. All of our participants were physically diagnosed with high- risk of hypertension and hypercholesterolemia. We used a structured questionnaire as our study tool, the contents of the questionnaire including participant background, Chinese version Hypertension & High Cholesterol Management Self- Efficacy Scale Questionnaires and The Summary of Hypertension & High Cholesterol Self- Care Activities Questionnaires. The participants would be taking the mobile application (App) and medical information service integrated course. Following the end of the interventional program, the participants would have to complete assessment at the baseline and there would be a 6-month follow-up with their personal cases.

Results

66 participants (mean age= 61.01) were evaluated throughout a 6- month study period. The results indicated that there were no change in BMI, Triglyceride and HDL-C. In addition, LDL-C slightly dropped. However, the SBP ($p< 0.001$), DBP ($p< 0.001$), Self- Efficacy Scale ($p= 0.031$) and Self- Care Activities Questionnaires ($p< 0.001$) had significant improvement between the baseline and the 6- month follow- up. (Table 1)

Discussion

Reliable instruments were used in this study to measure the effectiveness of the mobile health program with high-risk of hypertension and hypercholesterolemia in elderly population in Taiwan. The related application was tended to have a positive effect on the participants' self- healthcare control and helped them to maintain a higher level of healthy behavior in both physical and mental respects. This mobile health program should have a good effect on hypertension and hypercholesterolemia participants or elderly population. Although there is no statistically significant difference in LDL-C, a slightly progress can be seen from the result of the 6-month follow-up through the intervention of the mobile health program. Furthermore, we can use information relating to medical technology as well as the health education program to build a better health and medical network system in the future so that the senior participants with a high risk of hypertension and hypercholesterolemia can achieve a better quality of life.

Table and References

Table 1

	Baseline Mean ± SD	6- month follow- up Mean ± SD	t-value	95% CI	p- value
BMI	28.34±4.00	23.42±2.76	1.02	-4.83~14.87	0.313
Systolic blood pressure	127.56±17.96	120.65±15.83	3.80	3.28~10.54	0.000***
Diastolic blood pressure	78.52±11.65	73.53±10.88	3.75	2.33~7.64	0.000***
Triglyceride	126.89±109.20	126.78±80.10	0.01	-22.67~22.89	0.992
HDL- cholesterol	60.27±18.72	58.81±19.40	1.0	-1.35~4.2	0.304
LDL- cholesterol	125.77±31.07	123.58±31.53	0.60	-5.11~9.49	0.551
Self- Efficacy Scale	120.59±27.74	113.76±29.34	2.21	0.66~13.00	0.031*
Self- Care Activities Questionnaires	24.78±8.57	30.97±12.16	-3.99	-9.28~-3.09	0.000***

*p< .05 *p< .01 **p< .001 ***

References

1. Giudice R, Izzo R, Manzi MV, Pagnano G, Santoro M, Rao MA., et al. Lifestyle-related risk factors, smoking status and and cardiovascular disease. High Blood Press Cardiovasc Prev. 2012; 19(2), 85-92
2. Toledo PD, Lalinde W, Pozo FD. Interoperability of a Mobile Health Care Solution with Electronic Healthcare Record Systems. Engineering in Medicine and Biology Society, 2006 Aug 30- Sep 3, 28th Annual International Conference of the IEEE 2006;5214-5217.
3. Chung HT, Gau DL. Prevalence of the metabolic syndrome in individuals seeking for Health Examination. Cheng Ching Medical Journal. 2006; 2, 10-15.
4. Beigh SH, Jain S. Prevalence of metabolic syndrome and gender differences. Bioinformation. 2012; 8(13), 613-616.
5. Wu SF. Effectiveness of self-management for person with type 2 diabetes following the implementation of a self-efficacy enhancing intervention program in Taiwan. PhD thesis. Queensland University of Technology. 2007

Semantic Loss in Consolidated CDA Exports for Meaningful Use Stage 2

Joshua C. Mandel, MD[1,3], John D'Amore, MS[2], David Kreda[3]
[1]Boston Children's Hospital, Boston MA; [2]Diameter Health, Newton MA; [3]Harvard Medical School, Boston MA

Introduction: Stage Two of the Meaningful Use incentive program requires eligible providers and hospitals to use the Consolidated Clinical Document Architecture (C-CDA) to communicate clinical data between electronic health records (EHRs)[1]. As part of a national certification program, EHRs must be capable of producing C-CDA documents that pass specific testing criteria[2,3]. While these criteria are designed to provide a basis for interoperability across distinct electronic health record technologies, they omit several key factors for real-world interoperability. The SMART (Substitutable Medical Applications & Reusable Technology) Platforms team developed the SMART C-CDA Scorecard, a tool to evaluate C-CDA documents for adherence to best practices for interoperability. To assess the utility of this tool and catalog issues affecting real-world interoperability, the SMART Platforms team invited health information technology vendors to participate in a time-limited SMART C-CDA Collaborative.

Methods: We contacted 107 vendors of health information technology to submit a sample C-CDA document with fictional patient data. Forty-four organizations responded to the invitation and 17 technologies submitted at least one C-CDA sample. We evaluated each technology using the SMART C-CDA Scorecard, which grades performance across six domains: general issues, lab results, medications, problem observations, social history and vital signs. For each domain, multiple rubrics examine successful execution of best practices, such as "Vital signs are expressed with Unified Code for Units of Measure (UCUM)." Based on issues identified, we further elucidated error types and frequencies by manual inspection. We shared findings directly with eleven vendors who elected to participate in individual review sessions.

Results: For the technologies examined with the SMART –CDA Scorecard, the average score was 61% with a minimum of 37% and a maximum of 100%. The three most common errors observed in C-CDA submissions were 1) extraneous precision of datetime values, 2) use of erroneous codes from Unified Medical Language System(UMLS) and 3) medications not expressed with RxNorm semantic drug codes (Table 1). Manual inspection and discussion with vendors confirmed observations and provided context for errors.

Discussion: Based on our work, C-CDA documents are likely to omit key clinical information and require manual reconciliation in the second stage of Meaningful Use. In particular, we asw alarming errors in the use of dynamically bound vocabularies such as Systematized Nomenclature of Medicine (SNOMED), Logical Observation Identifiers Names and Codes (LOINC), and RxNorm. We recommend 1) that the certification process of EHRs should incorporate terminology validation when evaluating C-CDA documents, and 2) that producers and consumers of C-CDA documents should routinely implement real-time data quality services to track implementation issues as they occur. In addition, vendors found the SMART C-CDA Scorecard a useful tool in the review of their sample documents, and they expressed interest in integrating this functionality into their internal testing and development process.

Table 1: Error Frequency among 17 Technologies Submitting C-CDA Documents

Best Practice for C-CDA Implementation	Technologies with at least one violation
Sensible date and time precision	14
SNOMED, LOINC and RxNorm validate from UMLS	11
Medications expressed with RxNorm semantic drug codes	11
Lab results encoded in top 2,000 LOINC codes	8
Vital signs expressed with "required entries" template	8
Document uses C-CDA 1.1 templates when applicable	7
Problems encoded in HITSP SNOMED 16,000 subset	7
Problem statuses are internally consistent	6
Vital signs use appropriate UCUM units	6
SNOMED, LOINC and RxNorm display correct names	3
Only structured smoking status observations are used	3
Vital signs are encoded using correct LOINC codes	3
Smoking Status uses correct SNOMED subset	1
Smoking Status uses correct template	0

1. HL7 Implementation Guide for CDA Release 2: IHE Health Story Consolidation, DSTU 1.1 (US Realm). https://www.hl7.org/implement/standards/product_brief.cfm?product_id=258 Accessed March 7, 2014.

2. Medicare and Medicaid Programs; Electronic Health Record Incentive Program-Stage 2. Fed Regist [regulation on the Internet]. 2012 Sep 4 [cited 2014 Mar 10]; 77:53967 -54162. Available from: https://federalregister.gov/a/2012-21050

3. Health Information Technology: Standards, Implementation Specifications, and Certification Criteria for Electronic Health Record Technology, 2014 Edition; Revisions to the Permanent Certification Program for Health Information Technology. Fed Regist [regulation on the Internet]. 2012 Sep 4 [cited 2014 Mar 10]; 77: 54163 -54292. Available from: https://federalregister.gov/a/2012-20982

Patient-initiated secure messaging: Effects on chronic care process and clinical outcomes in a natural experiment

Sean McClellan, PhD, Palo Alto Medical Foundation Research Institute; Mountain View, CA, USA
Laura Panattoni, PhD, Palo Alto Medical Foundation Research Institute; Mountain View, CA, USA
Ming Tai-Seale, MPH, PhD, Palo Alto Medical Foundation Research Institute; Mountain View, CA, USA

Introduction: We studied the effect of patients sending secure electronic medical advice request messages to providers on process and clinical outcomes, among patients diagnosed with diabetes, hypertension, or hyperlipidemia. Prior research conducted in integrated delivery systems has shown substantial benefits of electronic messaging for patient health, but few longitudinal studies to date have examined the effects of implementing messaging in fee-for-service environments, where incentives for providers to use messaging for chronic disease management may differ. To estimate the effect of messaging, we exploited a natural experiment, where a large multispecialty practice in California eliminated a user fee for messaging of $60/year and initiated incentives for doctors to securely message patients about medical advice requests, resulting in increased rates of patient-provider messaging. We hypothesized that patients initiating any messages would have moderately improved process and clinical outcomes relative to others and that intensity of messaging would be positively associated with outcomes.

Methods: Longitudinal observational study from 3/2009-3/2012 using a "clean user" design, in which all patients in the study had activated their online portal before 2/2011, and none had paid to use messaging previously. The study included three cohorts of patients with diabetes (N=4,224), hypertension (N=15,504), or hyperlipidemia (N=21,728), based within a large multispecialty practice in California. Outcomes included both process and clinical measures for blood pressure, HBA1c, and LDL levels. Messaging was measured by whether or not patients initiated any email threads, categorized into quartiles in order to test differences in messaging intensity. The effects of messaging were identified through differences-in-differences analyses using generalized least squares on propensity score-matched samples, to account for the possibility of unobserved differences between treatment and control groups. Outcomes for patients initiating threads were compared before and after sending messages became free to patients, relative to controls. Covariates included patient age, insurance, Charlson score, and number of visits and phone calls to the group, and also characteristics of patients' assigned primary care providers (PCP), including gender, experience, continuity, and department. Standard errors were clustered within patients.

Results: Patients that messaged initiated an average of two threads per year. In multivariate analyses on matched samples, for each cohort, the probability of completing needed tests consistently increased by between 3 and 9 percentage points for patients initiating message threads, relative to controls. For example, in the diabetes cohort, the probability of receiving an HbA1c test increased by 2.8 percentage points (CI: 0.2,5.3) for patients sending only one message/year and by 5.4 percentage points (CI: 2.9,7.9) for patients sending four or more messages/year. However, we found no evidence that messaging was consistently associated with improvements in blood pressure, HBA1c, or LDL levels.

Discussion: By exploiting a natural experiment, we found that initiating secure medical advice requests was associated with clear improvements in the process of care for patients with diabetes, hypertension, or hyperlipidemia, but was not associated with improved clinical outcomes within the first year. Secure messaging may foster patient engagement and compliance above other online portal functionalities, such as reminders. However, improving clinical outcomes may require other interventions.

Analyzing and Comparing Clinical Work Systems with Cognitive Work Analysis: Lessons Learned

James L. McCormack, PhD, Paul N. Gorman, MD
Oregon Health & Science University; Portland, OR

Introduction: Recent studies have shown that information and communication technologies, including electronic health records (EHRs), can negatively impact patient safety, efficiency, effectiveness, and user satisfaction when they are poorly tailored to the local work practices and contexts of ambulatory care. Yet little is known about how independent medical practices choose to handle external clinical information (received in paper, electronic, or verbal forms), and what sociotechnical capabilities and constraints shape these choices. This presentation describes the lessons we learned applying the Cognitive Systems Engineering framework of Cognitive Work Analysis (CWA) to describe and compare four real-world clinical work systems that are used to handle external test results, referral communication, and outside care summaries.

Methods: CWA[1] was used to analyze and compare qualitative data from four independent primary care practices that included 24 unique interviews with providers and staff, and observations from 20 days in the field. They ranged in size from 1 to 8 providers; all used commercial EHRs and had attested to Stage 1 Meaningful Use. We selected CWA for three reasons. First, unlike traditional workflow studies, CWA is a formative approach that analyzes work systems in terms of the domain constraints and capabilities that shape local work practices. CWA also offered a rigorous and systematic framework to describe and compare both social and technical factors including organizational and environmental characteristics and the specific capabilities and limitations of personnel, facilities, equipment, and software. Finally, CWA has been used in industry and the military to inform the design of complex adaptive systems[2], but has rarely been applied in ambulatory care settings; a recent review found only one paper that used CWA in primary care out of a total of twenty-eight healthcare studies[3].

Results: We found that the broad perspective and unique representations offered by CWA provided a systematic framework to describe and compare the relevant constraints and capabilities across four diverse physician practices. Our use of CWA resulted in three considerations for improving the design of processes, artifacts, and technology: 1) Balance the unique affordances of paper, verbal, and electronic media; 2) Prevent technology from becoming a barrier to individual and team situation awareness; and 3) Empower and encourage staff and providers to design, adapt, and evolve local work practices.

Discussion: Our decision to use CWA presented many challenges. First, while books and papers describe the perspective, theory, and the framework, they rarely specify research procedures. Second, in contrast to previous healthcare studies[3], we attempted all five of the analytic stages described by Vicente[1] with limited time and resources. Our broad approach to data collection made it difficult to fully characterize individual cognitive strategies and competencies, the impact of team structure, and the details of specific information artifacts and media. Finally, extension of CWA by researchers and practitioners has created several variations in terminology and representational techniques from which to choose[3]. We will discuss these and other theoretical, methodological, and practical considerations in our presentation.

Conclusions and Acknowledgments: Cognitive Work Analysis, despite the many challenges, is a promising analytic framework for use in informatics studies. Although our application of CWA had several limitations, it succeeded in providing a sociotechnical map of information handling in primary care that could be used in future investigations and to inform the (re)design of processes, artifacts, and technology. This work was supported by pre-doctoral training grants from the National Library of Medicine (T15-LM-7088-18-ST) and the Oregon Clinical and Translational Research Institute (5 TL1-RR-024159-05). Special thanks to Gavan Lintern and Kenneth Funk.

References:

1. Vicente KJ. Cognitive Work Analysis: Toward Safe, Productive, and Healthy Computer-Based Work. 1st ed. CRC Press; 1999.
2. Bisantz AM, Burns CM. Applications of Cognitive Work Analysis. 1st ed. CRC Press; 2008.
3. Jiancaro T, Jamieson GA, Mihailidis A. Twenty years of Cognitive Work Analysis in health care: A scoping review. Journal of Cognitive Engineering and Decision Making. 2014 Mar; 8(1):3-22.

Application of Behavioral Economics to Design of Decision Support and Performance Feedback: A Comparative Randomized Controlled Trial

Daniella Meeker[A], Jeffrey A Linder[B], Stephen Persell[C] , Mark W Friedberg[B,D], Noah J Goldstein[E], Craig R Fox[E], Tara K Knight[F], Alan Rothfeld[G], Jason N Doctor[F]

[A]RAND Corporation, Santa Monica, CA; [B]Division of General Medicine and Primary Care, Brigham and Women's Hospital, Boston, MA; [C]Division of General Internal Medicine and Geriatrics, Institute for Healthcare Studies, Feinberg School of Medicine, Northwestern University, Chicago, IL;[D]RAND Corporation, Boston, MA; [E]Anderson School of Management, University of California, Los Angeles, CA; [F]Schaeffer Center for Health Policy and Economics, University of Southern California, Los Angeles, CA; [G]COPE Health Solutions, Los Angeles, CA 90015

Introduction

Informatics-based strategies to improve quality of care through electronic health records (EHRs) have had mixed results.[1,2] Failure to apply principles of decision psychology and behavioral economics to EHR design may contribute to these equivocal results. Recently, there has been interest in using knowledge from this field to inform design,[3] and EHR design in particular might stand to benefit.[4-6,7] We hypothesized that novel EHR-based clinical decision support (CDS) and automated performance feedback that build on lessons from behavioral economics could improve antibiotic prescribing practices for acute respiratory infections (ARIs) in primary care settings.[8]

Methods

We recruited 342 clinicians from 49 sites in three health systems and provided education in antibiotic prescribing guidelines to all participating clinicians. Each clinic site was randomized in a 2 x 2 x 2 factorial design to receive 0, 1, 2, or all 3 of the following interventions: (1) "Accountable Justifications": upon initiating an antibiotic prescription for an ARI, clinicians were prompted to record an explicit justification that appeared in the patient electronic health record; (2) "Suggested Alternatives": upon initiating an antibiotic prescription for an ARI, clinicians received an order set of non-antibiotic treatment choices; and (3) "Peer Comparison": each provider's rate of inappropriate antibiotic prescribing relative to top-performing peers was reported to the provider periodically by email. We customized antibiotic prescribing measures and decision support specifications to each health system based on preexisting workflow each of the three different EHR systems in use. We created a clinical data research network to manage the peer comparison intervention and monitor data quality and decision support. We measured antibiotic prescribing rates for likely viral ARIs over an 18 month intervention period.[8]

Results

At the time of analysis, participating clinicians had generated 26,783 visits for ARIs that were inappropriate for antibiotic treatment during the 18-month intervention. In these visits, 68% of patients were female, with an average age of 49 years. Adjusting for baseline prescribing rates, clustering, and patient characteristics, clinicians in the control group prescribed antibiotics in 29% of these visits. In comparison, prescribing rates were 21% among providers exposed to Accountable Justifications (relative reduction 25% versus control; adjusted OR 0.66, 95% CI 0.55-0.79); 29% in the Suggested Alternatives group (relative reduction 0%; adjusted OR 1.02, 95% CI 0.80-1.31); and 23% among providers exposed to Peer Comparisons (relative reduction 17%; adjusted OR 0.71, 95% CI 0.55-0.91). There were no significant interactions between suggested alternatives, accountable justifications, or peer comparison.

Discussion

Of the three interventions, the Suggested Alternatives was the least effective, and arguably most similar to conventional CDS tools using a standard alert-based design triggered by the ARI diagnosis. While only subtly different from traditional decision support, the Peer Comparisons and Accountable Justification interventions appealed to professional and social norms, which may account for their greater effectiveness. Additionally, the elicitation of reasoning in the justification process may engage a more deliberative thought system, since documented rationales are visible to others, in contrast to simple warning alerts that are easily dismissed without clear consequence. In conclusion, it is both feasible and effective to incorporate principles of behavioral economics into informatics-based interventions. These interventions may be more effective than traditional clinical decision support.

Supplemental figures

Exhibit 1 - Odds of antibiotic prescribing

Intervention	Odds Ratio	P>|z|	[95% CI]	
Suggested alternatives	1.02	0.854	0.8	1.309
Accountable justification	**0.66**	<0.001	0.555	0.792
Peer comparisons	**0.71**	0.009	0.55	0.919

References

1. Bright TJ, Wong A, Dhurjati R, et al. Effect of Clinical Decision-Support Systems A Systematic Review. *Annals of Internal Medicine.* 2012.
2. Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *BMJ.* 2005;330(7494):765.
3. Thaler R, Sunstein C, Balz J. Choice architecture. *Available at SSRN 1583509.* 2010.
4. Bourdeaux CP, Davies KJ, Thomas MJ, Bewley JS, Gould TH. Using 'nudge'principles for order set design: a before and after evaluation of an electronic prescribing template in critical care. *BMJ quality & safety.* 2013:bmjqs-2013-002395.
5. Harewood G, Clancy K, Engela J, Abdulrahim M, Lohan K, O'Reilly C. Randomised clinical trial: a 'nudge'strategy to modify endoscopic sedation practice. *Alimentary pharmacology & therapeutics.* 2011;34(2):229-234.
6. Elrod J, Androwich IM. Applying human factors analysis to the design of the electronic health record. Paper presented at: Nursing Informatics2009.
7. Middleton B, Bloomrosen M, Dente MA, et al. Enhancing patient safety and quality of care by improving the usability of electronic health record systems: recommendations from AMIA. *Journal of the American Medical Informatics Association.* 2013;20(e1):e2-e8.
8. Persell SD, Friedberg MW, Meeker D, et al. Use of behavioral economics and social psychology to improve treatment of acute respiratory infections (BEARI): rationale and design of a cluster randomized controlled trial [1RC4AG039115-01]--study protocol and baseline practice and provider characteristics. *BMC Infect Dis.* 2013;13:290.

Alerts for Low Creatinine Clearance: Design Strategies to Reduce Prescribing Errors

Brittany L. Melton, PhD, PharmD[1], Alan J. Zillich, PharmD[2-5], Michael Weiner, MD, MPH[2,4,5], M. Sue McManus, PhD, NP[6], Jeffrey R. Spina, MD[7,8], Alissa L. Russ, PhD[2-5]
[1]School of Pharmacy, University of Kansas, Lawrence, KS; [2]Center for Health Information and Communication, Department of Veterans Affairs, Health Services Research and Development Service CIN 13-416; PPO #09-298 Indianapolis, IN; [3]College of Pharmacy, Purdue University, West Lafayette, IN; [4]Indiana University Center for Health Services and Outcomes Research, Indianapolis, IN; [5]Regenstrief Institute, Inc., Indianapolis, IN; [6]Department of Veterans Affairs, Nephrology Services Central Texas, Temple, TX; [7]VA Greater Los Angeles Healthcare System, Los Angeles, CA; [8]David Geffen School of Medicine, University of California Los Angeles, CA

Introduction

Creatinine clearance is a common marker for kidney function and is regularly used to determine doses of medications that may be renally cleared or nephrotoxic. The use of alerts for creatinine clearance has been shown to reduce prescribing errors associated with renally dosed medications. However, the manner in which such information is presented can affect the usefulness of the alerts. *The objective of this study was to apply human factors principles to the design of creatinine clearance alerts. We hypothesized that the redesigned alerts which incorporated human factors principles would improve usability and reduce prescribing errors compared to an original alert design.*

Methods

Twenty VA prescribers completed two 30-minute prescribing sessions for fictitious patients in a counter-balanced, crossover design. One session used the original alerts based on the VA's electronic health record system, and the other used the redesigned alerts. Each session involved three medications that required adjustment or cancellation due to reduced creatinine clearance: spironolactone, ibuprofen, and allopurinol. In the original design, the alert was presented prior to any medication selection, while the redesigned alerts appeared only after a renally dependent medication was selected and prescription information had been entered. The redesign also included additional details about potential risks to the patient. For usability, we analyzed video data and asked prescribers to rate the perceived efficiency of viewing related lab results; this item was analyzed with the Wilcoxon signed-rank test. Correct and incorrect actions for each medication were determined *a-priori* and were used to evaluate prescribing errors. We assessed prescribing errors using Wilcoxon signed-rank and McNemar tests. This study was part of a larger investigation that evaluated prescribing outcomes for various types of alerts. Previous literature has shown that 20 participants uncovers about 99% of usability issues for a given software design.

Results

For usability, prescribers did not perceive any difference in the efficiency of viewing lab test results between the two designs ($p=0.113$). However, 9 (45%) prescribers stated that they would ignore the original alerts because they were not in the process of ordering a renally cleared medication at the time the alert appeared. Participants made significantly fewer prescribing errors when using the redesigned alerts compared to the original alerts ($n=26$ and $n=47$, respectively, $p=0.001$). When using the original alerts, 15% of participants made only one prescribing error while 50% made an error for all three medications. However, with the redesigned alerts, 45% of participants made one error and only 1 (5%) participant made an error for all three medications. Prescribing errors were significantly reduced for ibuprofen and allopurinol with the redesigned alerts versus the original alerts ($p=0.008$, $p=0.012$, respectively, for each medication). There was no significant difference in prescribing errors for spironolactone.

Discussion

These results support the hypothesis that applying human factors principles to creatinine clearance alerts would reduce prescribing errors. Although others have noted that alerts should appear early in the prescribing process to reduce interruptions and support workflow, our results indicate that prescribing errors were reduced when alerts appear closer to the time of decision-making, in this case, later in the medication ordering process. Overall, incorporation of human factors principles into creatinine clearance alerts can reduce prescribing errors and thereby improve safety for patients with reduced renal function.

Congestive Heart Failure Information Extraction Framework (CHIEF) Evaluation

Stéphane M. Meystre, MD, PhD[1,4], Youngjun Kim, MS[2,4], Andrew Redd, PhD[3,4], Jennifer Garvin, PhD, MBA[1,3,4]
[1] Department of Biomedical Informatics, [2] School of Computing,
[3] Division of Epidemiology, University of Utah, Salt Lake City, Utah
[4] VA Health Care System, Salt Lake City, Utah.

Abstract: *To help assess heart failure treatment and automatically extract clinical performance measures, we developed a natural language processing application to extract this information from clinical notes at the Veterans Health Administration. This application was evaluated with a corpus of 32962 notes from 771 patients and 98.9% of the patients meeting the performance measure criteria were accurately classified.*

Introduction: Congestive heart failure (HF) is the main cause of hospitalization for patient older than 65 and affects 5.8 million patients in the U.S., with high associated healthcare costs reaching more than $35 billion in the U.S. in 2008. Efforts to improve the treatment of HF and reduce associated costs include performance measures to assess treatment adherence to recommended care. The American College of Cardiology Foundation, the American Heart Association, and the Physician Consortium for Performance Improvement published such performance measures including "Left Ventricular Ejection Fraction (LVEF) Assessment" and "Angiotensin-Converting Enzyme Inhibitor (ACEI) or Angiotensin Receptor Blocker (ARB) Therapy." The ADAHF (Automating Data Acquisition for Heart Failure) project aimed at automating the extraction of such information from clinical notes. Natural language processing (NLP) can be used to extract various types of information from clinical narratives,[1] and we developed such an application to automatically extract clinical information for HF performance measures. The evaluation of the Congestive Heart Failure Information Extraction Framework (CHIEF) when classifying each patient as meeting the performance measure requirements or not is reported here.

Methods: The CHIEF is based on the Apache UIMA framework with modules for each type of information to extract from clinical notes. After text pre-processing (sentences detection, tokenization, part-of-speech tagging), a module extracts mentions and values of LVEF using machine learning (Margin Infused Relaxed Algorithm; MIRA[2]). The next module uses a medication names dictionary with fuzzy matching to extract medication names. Another module integrates RapTAT[3] to extract medication contra-indications or other reasons the patient should not be treated with ACEIs or ARBs. Finally, a last module uses all the extracted information extracted from multiple notes along with clinical note type information to classify the patient as meeting the performance measure or not. A set of rules is used for this last task. For the development and the evaluation of CHIEF, a large corpus of clinical notes from VHA (Veterans Health Administration) patients suffering from HF was manually annotated and split into a training corpus (notes from 314 patients) and a testing corpus (32962 notes from 771 patients). After training with the former, the latter corpus was used for the evaluation reported here.

Results: As presented in Table 1, when comparing the patient classification inferred by CHIEF using rules and the automatically extracted LVEF, medication, and contra-indications information, 98.9% of the patients meeting the performance measure criteria were accurately classified. Most information was accurately extracted. Only reasons for not treating the patient with ACEIs or ARBs were often missed, only detecting about 27% of them at the patient level.

Table 1: Patient level classification results

	Sensitivity	Positive Predictive Value	F_1-measure
Was LVEF assessed ?	1 (95% CI: 0.995-1)	0.990 (0.980-0.996)	0.995
Was the measured LVEF below 40% ?	0.968 (0.946-0.983)	0.952 (0.926-0.970)	0.960
Was the patient treated with ACEI or ARB ?	0.992 (0.981-0.997)	0.885 (0.858-0.908)	0.935
Was there a reason if not treated with ACEI or ARB ?	0.269 (0.199-0.349)	0.907 (0.779-0.974)	0.415
Performance measure met ?	0.989 (0.978-0.995)	0.987 (0.976-0.994)	0.988

Binomial exact confidence intervals used; F_1-measure is the harmonic mean of sensitivity and positive predictive value (with equal weight for each).

Acknowledgments: Research supported by VA HSR&D IBE 09-069 (ADAHF). The views expressed in this article are those of the authors and do not necessarily represent the views of the Department of Veterans Affairs.

References

1. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform. 2008;:128–44.
2. Crammer K, Singer Y. Ultraconservative online algorithms for multiclass problems. The Journal of Machine Learning Research. MIT Press; 2003 Mar;3.
3. Gobbel GT, Reeves R, Jayaramaraja S, Giuse D, Speroff T, Brown SH, et al. Development and evaluation of RapTAT: A machine learning system for concept mapping of phrases from medical narratives. J Biomed Inform. 2013 Dec 4.

Patients Screening for Clinical Trials Using EHR Representation Similarities

Riccardo Miotto[1], Ph.D. and Chunhua Weng[1,2], Ph.D.

[1]Department of Biomedical Informatics, Columbia University, New York, NY, USA
[2]The Irving Institute for Clinical and Translational Research, New York, NY, USA

Introduction

Randomized controlled trials generate quality medical evidence but suffer from longstanding recruitment problems [1]. Existing solutions mostly focus on eligibility criteria processing to facilitate trial search by patients [2, 3] or patient search using electronic health records (EHR) by investigators [4]. An untapped opportunity is using "case-based reasoning" to identify patients similar to known trial participants. This study tested the feasibility of this idea.

Methods

Our methodology framework includes two parts: (1) the participants currently enrolled in a trial are used to estimate a trial representation capturing relevant patterns in the participants' EHR data; (2) other patients are ranked by their EHR data's relevance to this representation: the higher the rank, the more likely are them eligible for the trial.

We identified patients enrolled into 13 clinical trials performed at Columbia University as our gold standard. These trials covered assorted diseases such as Type 2 diabetes, HIV, and multiple myeloma. The number of participants per trial ranged from 3 to 119. We obtained the EHR data for these patients and for about 30,000 other patients selected at random from the Columbia University's clinical data warehouse for design and testing. Data for trial participants were limited to a time period ranging from before the beginning of recruitment to one year after. Our data included patient clinical notes, ICD-9 diagnosis, event codes, and lab values. Clinical notes were pre-processed to extract UMLS-based tags and summarized using topic modeling to reduce the sparseness of the data [5]. Topics were generated over a random subgroup of 10,000 patients using Latent Dirichlet allocation [6].

During evaluation, for each clinical trial, we ran a 2-fold cross validation experiment, where half patients were for training and the other half was for testing, together with the rest of the random sample. The EHR-based trial representation, which was estimated from the training set and then applied to the test set, can be obtained in various ways. In this study, for the sake of simplicity, we considered concepts frequently shared by the trial participants and we averaged the set of their values, where values could represent occurrences for diagnosis and event codes, topic probabilities for notes, and mean result values for lab. Test patients were then ranked according to their similarity to the trial representation. Similarity scores referred to a weighted linear combination of cosine similarities computed over individual data types and ranged between 0 and 1 (with "1" meaning "perfect similarity"). For each test fold we measured (1) the similarity among relevant patients and their similarity to irrelevant patients and (2) the ranks of relevant patients. We then report results averaged over all fold experiments for all trials.

Results

The relevant patients were similar among them (mean similarity of 0.578) and significantly dissimilar with the rest of the collection (mean similarity of 0.129). This confirms that participants enrolled in a clinical trial share some patterns in their EHR data. These patterns can be used to identify new potentially eligible candidates and to separate them from ineligible ones given the large difference in similarity. This benefits the recommendation of potential participants to a trial as showed by the retrieval experiment results reported in Table 1. The weighted combination approach ranked about 65% of the relevant candidates within the top 10 positions. Moreover, for each fold trial, we ranked at least one relevant patient within the top five positions. This is particularly significant because, in an applicable scenario, an investigator would be required to look at only the top of the recommended list to likely pre-screen new potential participants to the target trial.

Discussion

Results demonstrate the feasibility of using EHR representation similarity with participants to expedite patient screening with little additional actions from the investigators. This approach can be used to constantly mine patient data warehouses for discovering new potential participants or it can also integrate strategies matching EHR data and eligibility criteria towards more effective systems. Results were obtained using a very basic training techniques; more sophisticated approaches (e.g., classification, learning to rank, statistical modeling) might lead to better performance. Therefore, we believe this work could also point to new research directions involving the study of novel machine learning techniques to improve the EHR-based trial representation and, consequently, to leverage the trial participant recommendation.

Acknowledgments

This research was supported by grants R01LM009886 and R01LM010815 from the National Library of Medicine, and grant UL1 TR000040 from the National Center for Advancing Translational Sciences.

Table 1: Retrieval experiment results in terms of precision-at-10 (P10), mean average precision (MAP), and mean reciprocal rank (MRR). "LwrBnd" refers to results obtained by randomly ranking the test collection, while "UppBnd" refers to the best results achievable if all the relevant documents were at the top of the ranking list. "wComb" is derived using a weighted linear combination of the cosine similarity scores achieved by events, diagnosis, note, and lab values, separately. The other algorithms simply ranked patients according to the similarity values based on only one data type (i.e., only diagnosis, only lab values, only code events, and only notes). These scores were strongly complementary with each other, leading to the significant improvement achieved by wComb.

	Ranking Algorithm	P10	MAP	MRR
Baseline	*LwrBnd*	0.004	0.001	0.005
	UppBnd	0.477	1.000	1.000
Ranking by Data Type	*onlyEvents*	0.046	0.096	0.137
	onlyDiagnosis	0.055	0.135	0.187
	onlyNote	0.106	0.186	0.267
	onlyLab	0.132	0.281	0.374
Linear Weighted Combination of Data Type Similarity	*wComb*	**0.304**	**0.558**	**0.717**

References

1. Sullivan, J., *Subject recruitment and retention: barriers to success.* Appl Clin Trials, 2004.
2. Miotto, R., Jiang, S., and Weng, C., *eTACTS: a method for dynamically filtering clinical trial search results.* J Biomed Inform, 2013. **46**(6): p. 1060-7.
3. Weng, C., Wu, X., Luo, Z., Boland, M.R., Theodoratos, D., and Johnson, S.B., *EliXR: an approach to eligibility criteria extraction and representation.* J Am Med Inform Assoc, 2011. **18 Suppl 1**: p. i116-24.
4. Cuggia, M., Besana, P., and Glasspool, D., *Comparing semi-automatic systems for recruitment of patients to clinical trials.* Int J Med Inform, 2011. **80**(6): p. 371-88.
5. Blei, D.M., *Probabilistic Topic Models.* Comm ACM, 2012. **55**(4): p. 77-84.
6. Blei, D.M., Ng, A.Y., and Jordan, M.I., *Latent Dirichlet allocation.* J Mach Learn Res, 2003. **3**(4-5): p. 993-1022.

Applying Active Learning to Word Sense Disambiguation in a Real-Time Setting

Sungrim Moon, Ph.D.[1], Yukun Chen M.S.[2], Jingqi Wang B.S.[1],
Joshua C. Denny M.D., M.S.[2,3], Hua Xu Ph.D.[1]
[1]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA; [2]Department of Biomedical Informatics, [3]Department of Medicine, Vanderbilt University, Nashville, TN, USA

Introduction: Statistical natural language processing methods often require large annotated clinical corpora. However, it is time-consuming and costly to create such corpora in the medical domain. Active learning,[1] a technology that actively selects the most informative samples for annotation in an iterative fashion, has been applied to a wide variety of language processing tasks and showed promising results on reducing annotation costs and improving classification models,[2] including clinical Natural Language Processing (NLP) classification systems.[3-6] However, prior active learning studies in medical text processing were simulations using previously annotated data sets. The effectiveness of active learning for building clinical NLP classifiers in a real-time setting has not yet been investigated. In our study, we applied an existing active learning system developed in the general English domain to build word sense disambiguation (WSD) classifiers for clinical abbreviations, and evaluated its performance using real-time annotation experiments by a physician.

Methods: In this study, we used an existing active learning system called DUALIST, [7] which provides an interface between the annotator and learning machine for iteratively annotating samples (ambiguous abbreviations in this case). The system implements an uncertainty-based querying algorithm integrated with a multinomial Naïve Bayes classifier for active sample selection. Figure 1 shows a screen shot of the DUALIST annotation interface for WSD. We tested an Active Learning (AL) mode in DUALIST, which asks an annotator to label a batch of samples (N=5 in this experiment) on the interface, trains/re-trains the WSD model and then actively selects the next batch of informative samples for annotation at the back-end. In addition, we also tested a Passive Learning (PL) mode, which randomly selects samples at each iteration, as the baseline method. We compared AL vs. PL for building WSD classifiers of clinical abbreviations. We selected six frequent ambiguous abbreviations (CA, CC, DM, LAD, PE, and RA) and collected sentences containing these abbreviations from admission notes at Vanderbilt University Medical Center. For each abbreviation, 200 sentences were randomly selected and manually annotated to serve as an independent test set, and the remaining un-annotated sentences were used in a pool for active or passive learners. A physician was asked to annotate each abbreviation at the AL or PL mode, in a random order, for five minutes. We then generated learning curves (plotting the accuracy of a WSD classifier on the test set vs. the time of annotation) for the two learning modes for comparison. The Area under the Learning Curve (ALC) was also reported as a global score for measuring the learning process.

Results: Figure 2 shows the learning curves for the six abbreviations used in our study. In most of the cases, the learning curves of the AL mode were situated above or around those of the PL mode. Table 1 presents the accuracies of the final WSD classifier after five minutes annotation, as well as the ALC scores for different learning strategies. Out of the six abbreviation experiments, AL mode achieved higher final accuracy in five and higher ALC score in four as compared to the PL mode. The AL mode showed an average ALC score of 0.80, while the PL mode achieved an average ALC score of 0.78. The average accuracies were 0.88 and 0.86 for the final classifiers by the AL and PL modes, respectively. The average annotated sample sizes were 99.17 and 128.66 for AL and PL modes, respectively. The average number of re-training iterations in five minutes by AL and PL modes were 22.33 and 27.50, respectively.

Discussion: To the best of our knowledge, this is the first study that evaluates active learning for clinical NLP classification tasks in a real-time setting. We showed that the AL mode outperformed the PL mode for four out of six clinical abbreviation WSD tasks. In the case of the abbreviation RA, AL achieved an accuracy of 0.89 in 1.82 minutes, while PL required 5 minutes for the same task. Thus, AL decreased annotation time by 64% (3.18 minutes). In terms of the size of annotated samples, AL mode used 40 samples while PL needed 116 random samples to reach the same accuracy (a reduction of 66%). The superior performance of AL is more obvious in the early stages of the learning curves. Our findings demonstrated the practical value of active learning in building WSD classifiers.

Acknowledgement: This study was supported by grant from the NLM R01LM010681.

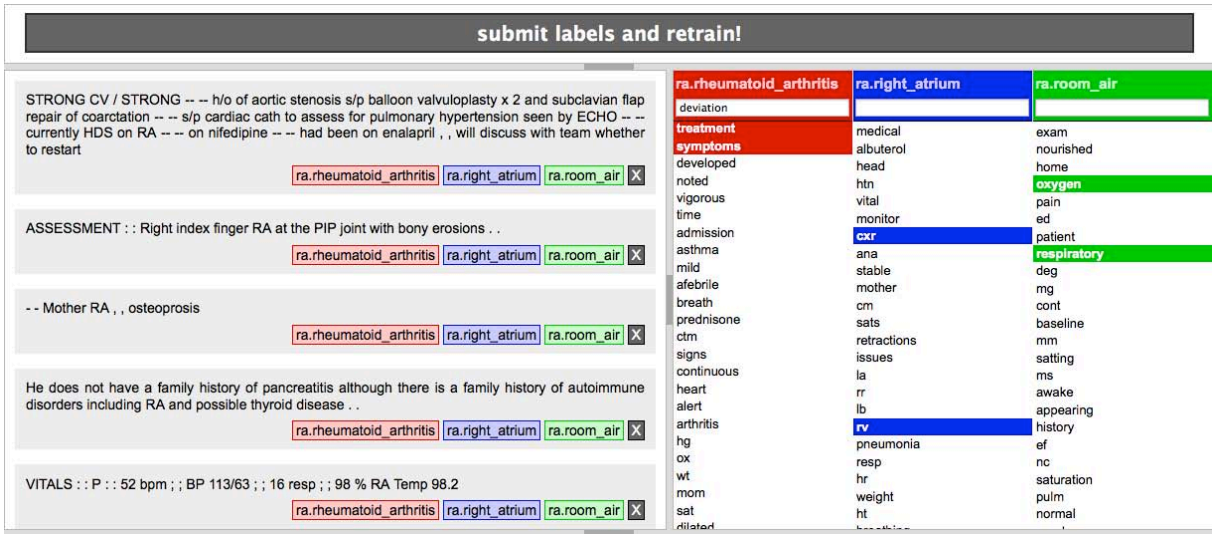


Figure 1. Interface of DUALIST

Table 1. Final accuracy and ALC score for six abbreviation experiments

	Learning mode	CA	CC	DM	LAD	PE	RA
Final Accuracy	Active learning	0.830	0.815	0.985	0.980	0.735	0.940
	Passive learning	0.835	0.785	0.970	0.975	0.725	0.890
ALC score	Active learning	0.774	0.670	0.936	0.931	0.666	0.848
	Passive learning	0.768	0.683	0.922	0.943	0.613	0.776

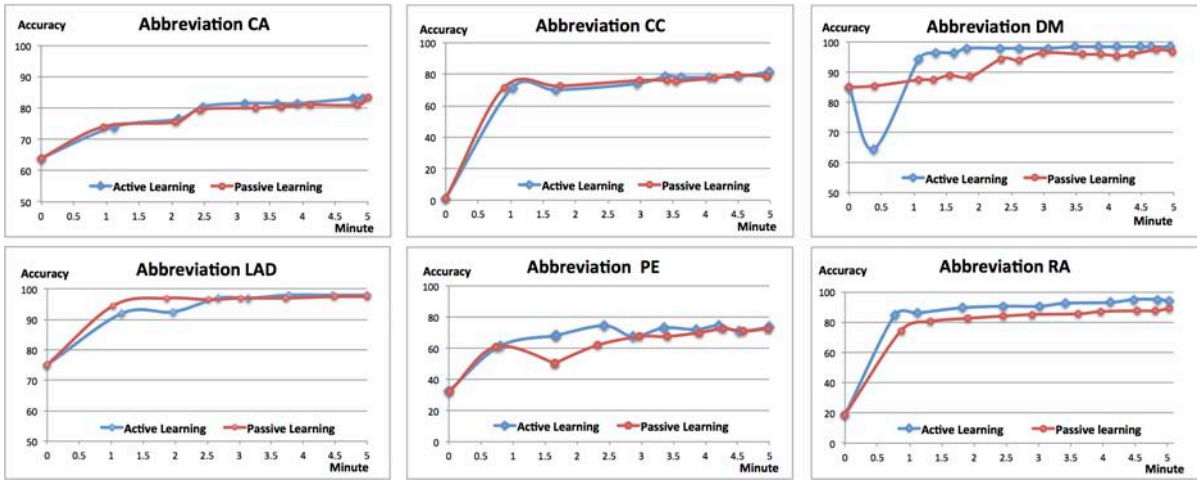


Figure 2. Learning curves that plot accuracy of a WSD classifier vs. annotation time for all six abbreviations included in this study

1. Settles B. Active learning literature survey. *University of Wisconsin, Madison.* 2010;52:55-66.
2. Olsson F. A literature survey of active machine learning in the context of natural language processing. 2009.
3. Figueroa RL, Zeng-Treitler Q, Ngo LH, Goryachev S, Wiechmann EP. Active learning for clinical text classification: is it better than random sampling? *Journal of the American Medical Informatics Association.* 2012;19(5):809-816.
4. Chen Y, Carroll RJ, Hinz ER, et al. Applying active learning to high-throughput phenotyping algorithms for electronic health records data. *Journal of the American Medical Informatics Association.* Dec 2013;20(e2):e253-259.
5. Chen Y, Cao H, Mei Q, Zheng K, Xu H. Applying active learning to supervised word sense disambiguation in MEDLINE. *Journal of the American Medical Informatics Association.* Sep-Oct 2013;20(5):1001-1006.
6. Chen Y, Mani S, Xu H. Applying active learning to assertion classification of concepts in clinical text. *Journal of biomedical informatics.* Apr 2012;45(2):265-272.
7. Settles B. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2011.

Informatics for Integrating Biology and the Bedside (I2b2) Clinical Trials (CT) Patient Ascertainment Suite

Shawn Murphy MD, Ph.D. (1,3), Nich Wattanasin MS (2), Vivian Gainer MS (2), Susanne Churchill Ph.D.(2), Issac Kohane, MD, Ph.D.(3)

1 – Massachusetts General Hospital, Boston, MA, 2 – Partners Healthcare, Charlestown, MA, 3 – Harvard Medical School, Boston, MA.

Introduction – The i2b2-CT Software Suite is an informatics tool set designed to facilitate the selection of patients from existing healthcare data (electronic health records) for clinical trials. The i2b2-CT software application may be used as a stand alone application at any single healthcare entity with an i2b2 clinical research data warehouse. It will be linked in March of 2015 through a web based research network to allow identification of eligible patients from multiple institutions through the Shared Health Research Informatics Network (SHRINE).

Methods - An investigator first investigates the feasibility of a putative clinical trial by querying his/her institution's i2b2 instance [1] for the aggregate number of patients meeting selected inclusion/exclusion criteria. If he/she determines that a trial is feasible, the investigator can proceed to requesting a Limited Data Set (LDS) under a data use agreement for rapid patient screening to build a more thoroughly characterized subset of patients to understand if the patients are suitable for the clinical trial [2]. Once the patient population has been adequately reviewed as a limited data set and the clinical trial feasibility is ascertained, the investigator and/or his/her study staff will progress to viewing individual data from each of the patients selected from the LDS review following a full review of the Institutional Review Board to allow access to Protected Health Information for contacting the patients. Access to the data and contact of patients is governed by that institution's current Institutional Review Board (IRB) practices.

Results - The i2b2-CT project developed novel software applications to provide the investigator a friendly, intuitive interface for rapid line by line viewing of digested patient attributes, and the ability to drill deep into a individualized view that can appear very similar to what an electronic medical record could produce by using Substitutable Medical Applications and Reusable Technologies (SMART) [3]. The SMART apps can be customized for each data type, and a collage of these apps can be selected and rendered in a layout that displays diagnoses, medications, and images, and genomics on a specific patient. Thus for even the most novel forms of data or routine data, the display can be adapted and controlled.

Discussion – The i2b2-CT project is used to accrue patients for clinical trials. It is also valuable as a platform for generally validating phenotypes as part of "in-silico" studies that explore associations of phenotypes with the information gathered from samples as in a BioBank. The software can adapt to Patient Surveys and Reported Outcomes as well as routine Electronic Medical Record data. It can be used for multisite trials across institutions using SHRINE [4].

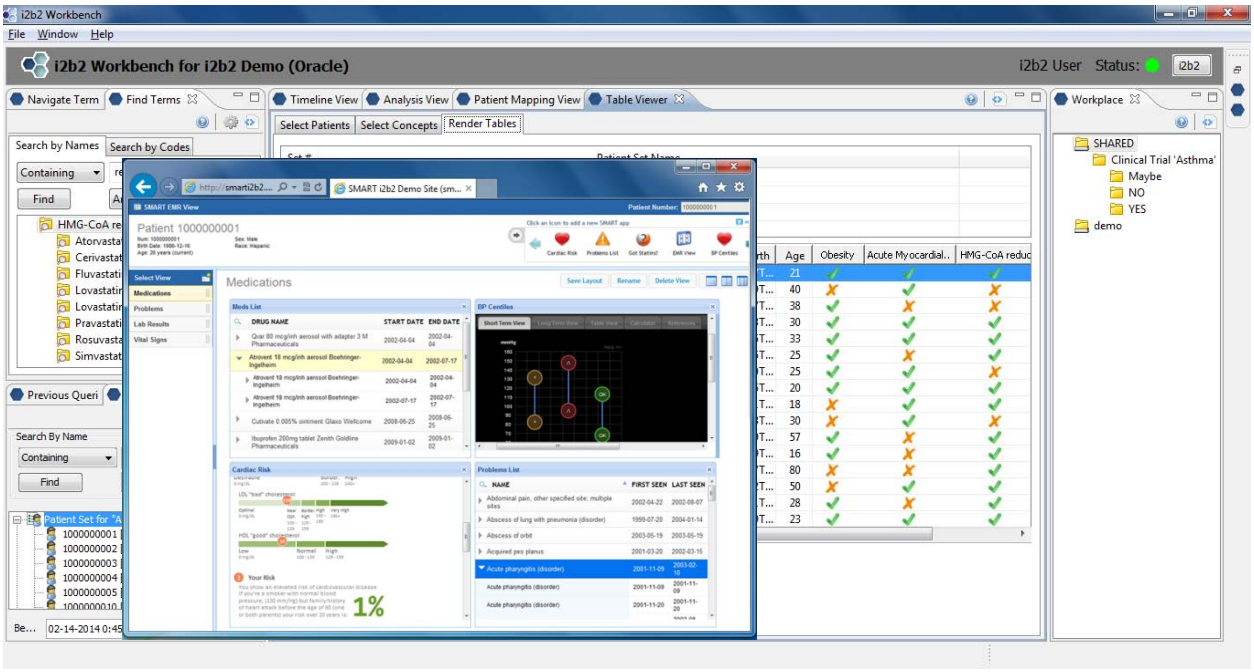


Figure shows the i2b2 Workbench where a patient set is investigated in detail using the customizable SMART electronic medical record view.

1. Murphy, S.N., et al., *Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2).* J Am Med Inform Assoc, 2010. **17**(2): p. 124-30.
2. Murphy, S.N., et al., *Strategies for maintaining patient privacy in i2b2.* J Am Med Inform Assoc, 2011. **18 Suppl 1**: p. i103-i108.
3. Mandl, K.D., et al., *The SMART Platform: early experience enabling substitutable applications for electronic health records.* J Am Med Inform Assoc, 2012.
4. McMurry, A.J., et al., *SHRINE: enabling nationally scalable multi-site disease studies.* PLoS One, 2013. **8**(3): p. e55811.

Audience will understand capabilities of i2b2-CT platform
Audience will understand privacy concerns addressed by i2b2-CT
Audience will understand capabilities of SMART platform
Audience will understand capabilities of Clinical Trials platform

Use of Ontologies for Disease Management Clinical Decision Support Systems

Seyed Ali Mussavi Rizi, MD, MHA; Abdul Roudsari, PhD
University of Victoria, Victoria, British Columbia, Canada

Introduction: Ontologies have been used in many different capacities in health information systems. In this study we aimed to perform a systematic review of the literature on using ontologies for disease management clinical decision support systems (CDSS), and development and execution of clinical practice guidelines (CPG) and clinical pathways (CP).

Methods: A systematic review of literature was performed using MEDLINE and EBSCO Academic Search Complete database. "Ontology", "ontologies", "ontological" terms were used as the primary keywords, excluding "gene ontology", and results were combined with searches for terms related to CDSS, CPG, CP using literal and MESH terms. Only papers that presented development and characteristics of an actual system were included, thus any paper that only presented a theoretical approach, details of development of a certain ontology, or review articles were excluded. The selected articles were appraised and compared on the following attributes: the rationale for using ontology in the design; tools and formalism used; clinical usage; issues and challenges; implementation stage; evaluation and follow-up studies. Application of inclusion and exclusion criteria and assessment of articles were mainly performed by the first author (SAMR) in discussions with the second author (AR).

Results: The search strategy resulted in 141 articles. After review of the abstracts, 32 articles published from 1995 up to early 2012 were selected for full review. The articles could be divided into three broad categories: clinical practice guidelines (CPG) (14) clinical pathways (CP) (4), and general clinical decision support systems or expert systems (collectively CDSS) (14). In the CPG group, the general rationale for using ontologies was developing a flexible, reusable, shareable, and computable description of the CPG and domain with some reasoning capability. In the CP group, ontologies were used to describe the CP model, the terms, semantics, relationships, also knowledge about resources, outcome, and variance. The rationale in the CDSS group was more varied and included separation of domain knowledge from solver components, increasing sharing and reusability of components, computing semantic similarity, reasoning and classification capabilities of ontologies, making implicit knowledge explicit, and development of adaptive and personalized CDSS. Protégé family of tools and OWL family of languages were most commonly used as the tool and language for development of the ontologies. Ontology engineering methodology was explicitly identified only in 5 articles. Almost all of the articles presented the results of validation or evaluation of the system, albeit primitive. Only 7 (22%) of 32 systems were deployed at the time of publication.

Discussion: Results of this study show that although ontologies could be useful in developing different types of disease management CDSS, they introduce new challenges. For example, using ontologies to separate domain knowledge and problem solving methods provides means for reusability, rapid development and update of the systems, however it shifts the problem of maintaining rules to maintaining mappings between the components and introduces new problems of incompatible and sophisticated system design with a low tendency for adoption. Another problem in using ontologies in CDSS is the required expertise and interaction between medical experts and knowledge engineers. One study evaluated, and identified that Protégé proved to be unintuitive and difficult to use for domain experts. This observation requires further research to identify the optimal tools, expertise, and composition of teams required for ontology development. Finally, comparing characteristics of the systems reviewed against published guidelines on when ontological methods are the right choice shows that some of the systems reviewed could have benefited from simple rule engines or knowledge bases, rather ontology.

This review has few limitations. First, the term "ontology" refers to different concepts and does not have a universally accepted definition. We used the literal terms for search strategy without further specification which could lead to omission of relevant articles. Second, we only focused on published academic literature which could result in potential exclusion of commercial or proprietary systems that are not described in this body of literature and limit generalizability of the results. Lastly, inclusion, exclusion, and assessment of articles were mainly performed by one author that could introduce bias into the results.

Automated Clinical Trial Eligibility Pre-Screening: Increasing the Efficiency of Participant Identification for Clinical Trials

Yizhao Ni, PhD, Stephanie Kennebeck, MD, Constance M McAneney, MD, Judith W Dexheimer, PhD, Todd Lingren, MS, Qi Li, PhD, Haijun Zhai, PhD, Imre Solti, MD, PhD, MA
Cincinnati Children's Hospital Medical Center (CCHMC), Cincinnati, OH

Introduction and Background

Manual eligibility screening (ES) for a clinical trial typically requires a labor intensive review of patient records that utilizes many resources. Many barriers remain for automated screening and the progress seems to be reaching a plateau[1]. To address these barriers, we assembled state-of-the-art natural language processing (NLP), information extraction (IE) and machine learning (ML) technologies to develop an automated ES system. The objective is to develop a high sensitivity automated pre-screening system to identify patients who meet core eligibility characteristics to markedly streamline and focus the pool of candidates for manual screening.

Data and Method

We focused on clinical trials for pediatric patients who visited the Emergency Department (ED) at CCHMC between 1/1/2010 and 8/1/2012. We collected the eligibility criteria for 13 disease-specific ED clinical trials (inclusion criteria involved one or more diseases). We randomly sampled 600 patients who visited the ED during that timeframe and extracted 15 EHR attributes to represent their clinical status at their dates of visit (Tab. 1). Two physicians then reviewed the chart for each patient and the criteria of the open trials on the patient's date of visit and created a set of gold standard trial-patient matches (Inter Annotator Agreement=96.81%). The physicians reviewed a total of 3061 trial-patient pairs and found 74 matches (average eligible rate 2.42%).

We utilized the state-of-the-art NLP, IE and ML algorithms to build the ES system (Fig. 1). Given a clinical trial and the patient candidates, (1) the system first applied logical constraint filters (LCFs) to rule out ineligible patients based on structured attributes (e.g. age) derived from the trial criteria; (2) the text-based attributes of the pre-filtered patients were then processed, from which the medical terms and assertions were extracted using the clinical NLP tool cTAKES[2] and stored as patient pattern vectors; (3) the same process was applied to the trial criteria to construct the trial pattern vector; in addition, (4) the vector was extended with informative patterns extracted from EHRs of previously enrolled patients using ML techniques so as to capture potential hyponyms (e.g. football, soccer) relevant to the trial criteria (e.g. sport-related trauma) in the patient EHRs; finally (5) IE algorithms matched the trial vector with the patient vectors using TF-IDF similarity and returned a ranked list of patients based on the similarity scores.

We used three evaluation metrics to measure performance: (1) mean average precision (MAP) commonly calculated in information retrieval; (2) workload defined as the number of patients required to be reviewed from the system output to get all eligible patients (i.e. recall=100%); and (3) the recall curve on different cut-offs of the system output. For system comparison, the baseline was used to simulate the process without automated pre-screening. It was implemented by randomly shuffling the patient list for a trial. We then compared its performance with three variants of the ES system that cumulatively integrated the proposed components: (1) LCF that only used the LCF component to exclude ineligible patients and randomly shuffled the rest of the patient list for a trial; (2) LCF+NLP: the ES system specified in Fig. 1 without the ML component; (3) LCF+ NLP+ML: the ML component was also included. To train the ML component, we collected EHRs of 4,202 patients historically enrolled in the 13 trials during the study period and randomly sampled 1% to 100% of the data for training (c.f. Fig. 2.c). None of these additional patients were included in our 600 patient set.

Results and Discussion

Fig. 2.a shows the systems' average performance over all trials. The LCF had good capability in excluding ineligible patients (workload reduction 48%, 49 vs 95 screened patients). However, without information from clinical text, it was unable to match descriptive criteria (e.g. diagnosis) with patients' clinical status. By applying the NLP and IE algorithms on clinical narratives, the LCF+NLP system further improved the performances (workload reduction 86%). For LCF+NLP+ML we observed consistent improvement in performance when more training data was used (Fig. 2.b). In case of 2% training data (85 samples), it outperformed LCF+NLP significantly (Fig. 2.a). This result implied the great potential of ML in boosting the performance of the ES system. Finally the LCF+NLP+ML system (trained on 2% of the data) achieved 90% recall by screening the top 24% of the system output (Fig. 2.c), suggesting that the screening efficiency could be improved by 400% while missing only 10% of eligible patients.

Table 1. Structured (S) and text-based unstructured (US) attributes extracted from patients' EHRs.

Age (S)	Gender (S)	Spoken language (S)
Glasgow coma scale (S)	Vitals (S)	Acuity (1-5, 1:urgent; 5: not urgent) (S)
Patient pregnancy Y/N (S)	Guardian presence Y/N (S)	Chief complaint (US)
Diagnosis (US)	ED clinical notes (US)	Patient medical history (US)
Patient surgical history (US)	Patient family medical history (US)	Patient medication history (US)

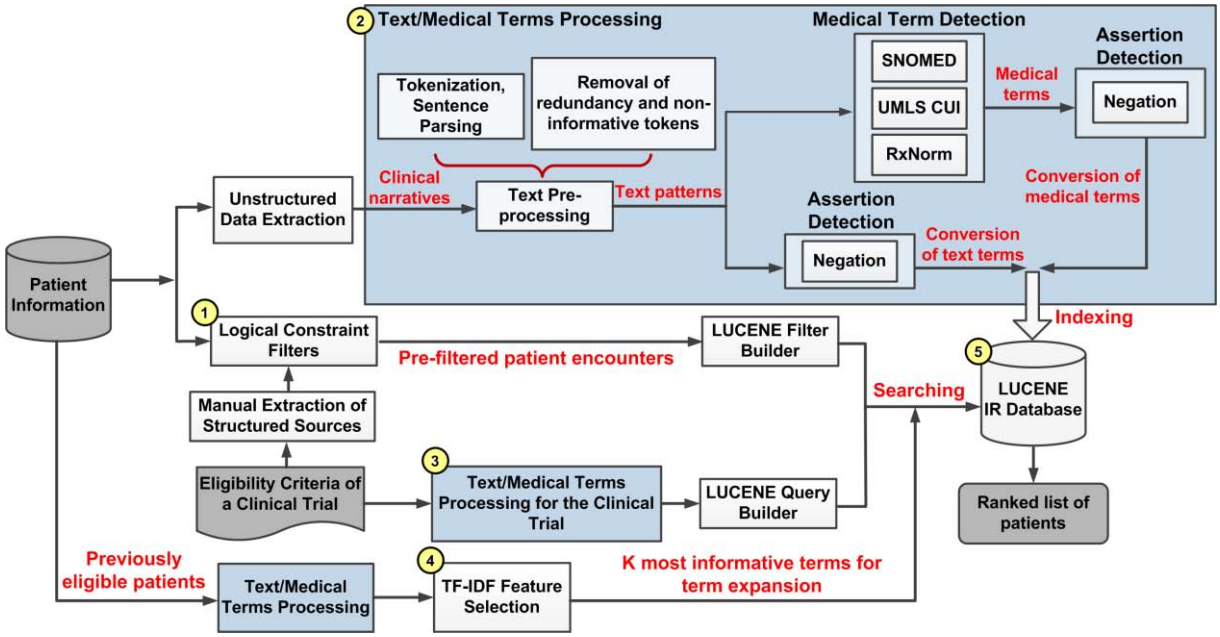
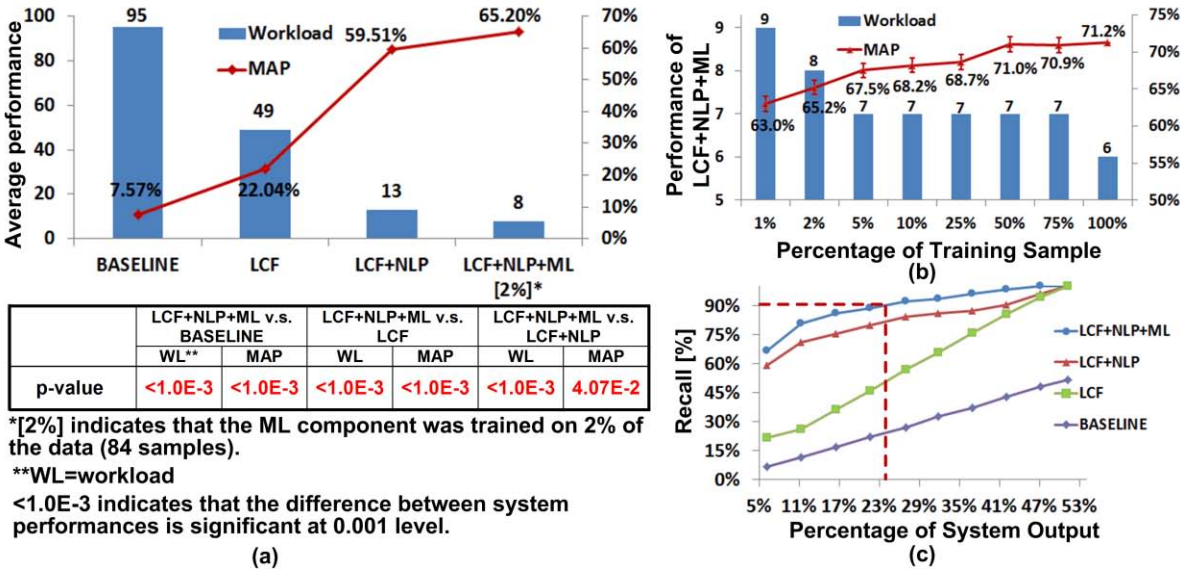


Figure 1. The architecture of the proposed automated ES system.



	LCF+NLP+ML v.s. BASELINE		LCF+NLP+ML v.s. LCF		LCF+NLP+ML v.s. LCF+NLP	
	WL**	MAP	WL	MAP	WL	MAP
p-value	<1.0E-3	<1.0E-3	<1.0E-3	<1.0E-3	<1.0E-3	4.07E-2

*[2%] indicates that the ML component was trained on 2% of the data (84 samples).
**WL=workload
<1.0E-3 indicates that the difference between system performances is significant at 0.001 level.
(a)

Figure 2. System performances, including performance curves of LCF+NLP+ML with different training samples.

References

1. Edinger T, Cohen AM, Bedrick S, et al. Barriers to retrieving patient information from electronic health record data: failure analysis from the TREC medical records track. AMIA Annu Symp Proc 2012;180-188.
2. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;**17**(5):507-51.

An Electronic Patient Safety Checklist Tool for Interprofessional Healthcare Teams and Patients

Kumiko Ohashi, RN, PhD[1], Patricia C. Dykes DNSc, RN, FAAN, FACMI [1, 2], Diana L. Stade[1], Eddy Chen, MD[3], Anthony F. Massaro, MD[4,] David W. Bates, MD, MSc, FACMI[1] Lisa S. Lehmann, MD, Ph.D[1]

1General Internal Medicine, Brigham and Women's Hospital, Boston, MA; [2]Center for Excellence in Nursing Practice, Brigham and Women's Hospital, Boston, MA; [3]Department of Medical Oncology, Dana-Farber Cancer Institute/Brigham and Women's Hospital, Boston, MA; [4]Department of Pulmonary Medicine, Brigham and Women's Hospital, Boston, MA

Introduction: Fragmentation of healthcare poses a threat to patient safety, and leads to inefficient care, which is low value care. Several studies have demonstrated the usefulness of checklists to improve the quality of care such as increasing treatment compliance and reducing complications and medical errors in the clinical setting. In the current model of care, however, physicians and nurses have separate patient care checklists to organize, manage, and hand off critical patient-based tasks. Physicians and nurses do not have integrated checklists to share their documents with each other or with other healthcare providers. The goal of this study is to facilitate and support interaction and communication among all the members of a patient care team by developing, implementing and evaluating an electronic Patient Safety Checklist Tool. Physicians and nurses will enter data into the Checklist Tool and the contents will be shared with multidisciplinary team members and patients and family to review and provide input. The Patient Safety Checklist Tool is an electronically integrated interdisciplinary suite of tools that will incorporate adaptations of existing validated safety checklists for healthcare providers to use during team rounds. Information dashboards can also be generated to provide a better overview of a patient's care for healthcare providers and patients than the current method of tracking a patient's hospitalization. Information will flow from electronic patient platforms to a Rounding Checklist and from the Rounding Checklist to the Care Team & Patient Dashboards. The Patient Safety Checklist Tool would allow all members of a healthcare team to add, edit, and view a patient's safety items (e.g., deep venous thrombosis (DVT) prophylaxis, patient falls, gastrointestinal bleeding prophylaxis etc.) and plan of care (e.g., planned procedures, imaging, labs, family meetings, etc.).

Methods: 1) A problem analysis phase consisted of the following: study team performing workflow observations, document review, interviews with healthcare providers, workflow analysis, team communication, utilization of safety bundles, and to do lists, and nine focus groups. Clinical documentation (e.g. physicians' notes, nursing plan of care, flow sheets, personal to do checklists, and notes on white boards) were reviewed to determine the contents of the Patient Safety Checklist Tool. 2) Using knowledge gained in the problem analysis phase we designed the Patient Safety Checklist Tool**.**

Results: The pilot content of the Patient Safety Checklist Tool is shown in Figure 1. Based on healthcare provider and patient feedback, patient risk/status reminder icons were developed for display in a web-based patient portal with detailed educational contents for patients and care partners. As a first pilot test, we implemented the Patient Safety Checklist Tool to support team rounding, communication, and documentation in a medical ICU.

Discussion/Conclusion: We found that there were some challenges with developing/implementing an interdisciplinary Patient Safety Checklist Tool particularly early on in the process. The key of successful implementation was assigning responsibility for specific content and ensuring completeness to avoid duplicated documentation or missing information. The Checklist Tool has potential to facilitate efficient and collaborative patient management. It allows healthcare providers, patients and families to access to the daily plan of care, and to eliminate patient harms. The checklist tool also focused on specific tasks that are essential to the provision of high quality and safe patient-centered care. Furthermore, there is potential for enhancing communication between providers, patients and families regarding specific patient safety information.

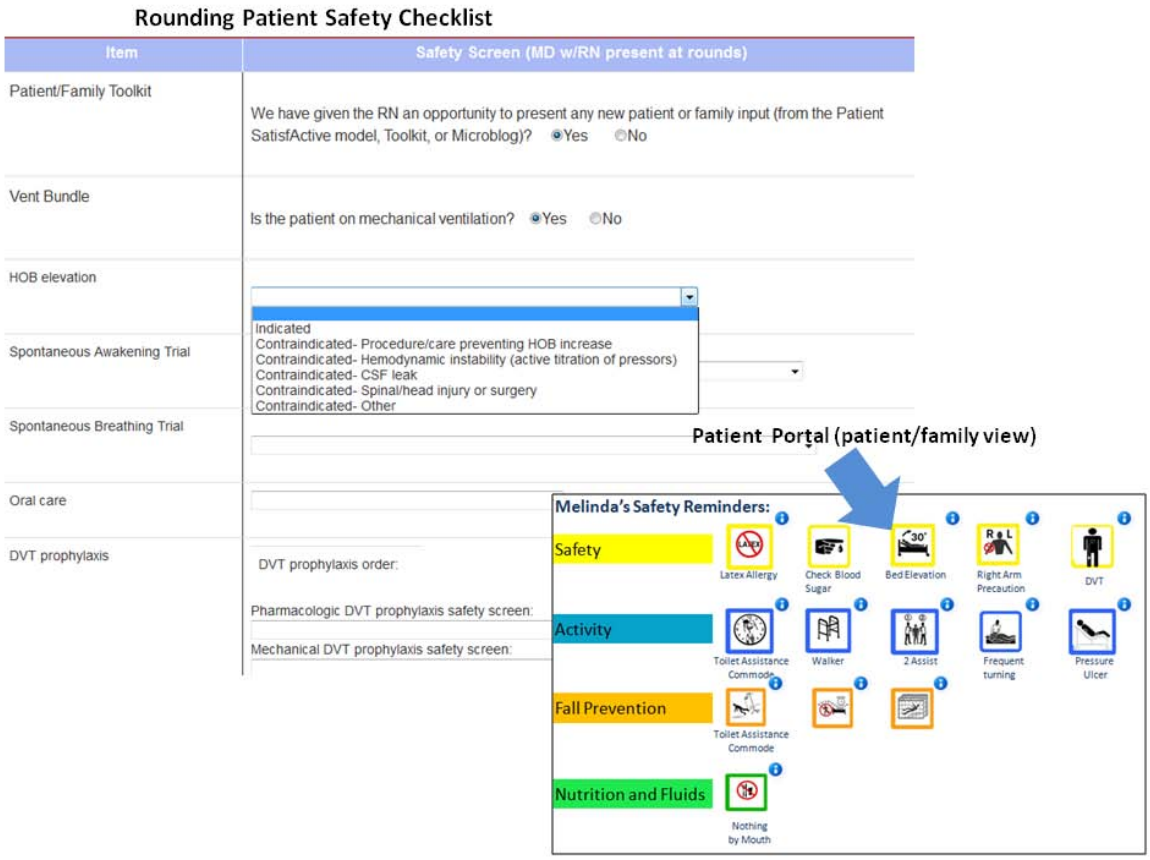


Figure 1. Overview of Patient Safety Checklist Tool Elements

Feasibility of Distinguishing Older Adults with Fall Risk Using Motion Analysis

Amy Papadopoulos, DSc[1], Cindy Crump[1], Christine Tsien Silvers, MD, PhD[1,2], Bruce Wilson[1]; [1]AFrame Digital, Inc., Reston, VA; [2]Children's Hospital Informatics Program, Boston, MA

Introduction. Falls are the leading cause of hospital admissions for trauma in the U.S. and the primary etiology of accidental deaths in persons over 65 years.[1] While it is of great importance to reduce the overall number of falls and fall-related injuries, accomplishing this goal will require more dynamic methods for "getting ahead" of a fall through a greater emphasis on risk mitigation and prevention. One approach is by improving assessment and classification of a person's level of fall risk at any given point in time, along with recognizing on a more dynamic basis when that fall risk increases to enable earlier intervention with preventive measures. The purpose of this study was to establish the feasibility of accurately and unobtrusively determining fall risk using motion data collected from places on the upper body to differentiate between fall-prone and non-fall-prone older adults within a laboratory environment.

Methods. As part of an IRB-approved protocol, 30 subjects (10 young adults; 10 older, stable walkers; 10 older, known fallers) were recruited. For each subject, motion data were collected at Virginia Tech's Locomotion Laboratory using Inertia-Link wireless sensors (LORD MicroStrain, Williston, VT) taped to a battery pack and attached to the left wrist and trunk. The sensors collected both acceleration and angular velocity at those points. An electro-optical motion capture system with Qualysis software captured 100 image frames/second from each of six cameras while subjects walked three times each (1) at a normal pace on a 15.5-meter walkway equipped with force plates, and (2) on a treadmill for two minutes. Data were analyzed using software developed in MATLAB (MathWorks, Natick, MA). Stride time and stride length were computed using data from the optical system and force-plate anteroposterior heel position measurements for a single stride. Sensor data peak-to-peak values were calculated and correlated with baseline stride time. Peak forearm swing angle from the wrist gyroscope and peak acceleration from wrist and trunk sensors were calculated and correlated with baseline stride length. As a means for recognizing stability from only wrist acceleration data gathered while walking, a new method was attempted. This method involves first calculating the vector sum of the acceleration data gathered while the participant walked on the treadmill. The autocorrelation value was then calculated as a 2-second window taken from the beginning of the resultant data was cross-correlated with the remaining resultant data from the same individual. As the window was passed over the data, peaks in autocorrelation were noted, the time between peaks calculated, and the maximum peak value stored. The average time between peaks and the maximum autocorrelation value were used as features alone and together to differentiate fallers from other elderly subjects by using the Mahalanobis distance to determine to which group of data, fallers or non-fallers, the features were closest. Leave-one-out cross validation was used to test the new method, and the areas under the receiver operating characteristic curves (AUC) were determined.

Results. The feasibility of using a socially-acceptable wrist-based device to determine gait characteristics was demonstrated by establishing the correlation between arm-swing time and stride time (0.74 wrist gyroscope, 0.74 trunk acceleration, 0.66 wrist acceleration). Stride length was found not to correlate with the collected sensor values. The possibility of using a wrist-based sensor to differentiate fall-prone from non-fall-prone individuals was demonstrated using the autocorrelation method (AUC of 0.82 for the elderly only, 0.86 for all participants).

Discussion. Wrist acceleration data collected during walking were able to differentiate fall-prone from non-fall-prone individuals. This, combined with the previously established ability to automatically differentiate walking from other activities of daily living,[4] makes unobtrusive fall-risk assessment feasible via a small, wireless wrist device.

Acknowledgements. The authors would like to thank Professor Thurmon Lockhart of Virginia Tech and the National Institute on Aging (NIA) (1R43AG029721-01) for support. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIA or the National Institutes of Health.

References

1. Centers for Disease Control and Prevention, Home & Recreational Safety. "Falls among older adults: an overview." http://www.cdc.gov/HomeandRecreationalSafety/Falls/adultfalls.html, accessed December 2, 2013.
2. Papadopoulos A, Vivaldi N, Silvers C. Wrist-based accelerometers successfully differentiate walking from other activities. Journal of Mobile Technology in Medicine. 2012:1(4S).

Missing Evidence for HIT Transformation in a Review of the Literature

Andrew B. Phillips[1], PhD, RN and Jacqueline A. Merrill[2,3], PhD, MPH, RN
[1]MGH Institute of Health Professions, School of Nursing, Boston, MA, [2]School of Nursing, [3]Department of Biomedical Informatics, Columbia University, NY, NY

Abstract
The complexity of healthcare creates barriers to transformation and inhibits a linear path towards change. This study examined the healthcare literature for evidence of technology transformation applying the theory of punctuated equilibrium (PE). Through an analysis of 477 articles from 2004 – 2012 we found no evidence to support technology transformation in health care. We hypothesize that the system creates barriers to authentic consumer engagement needed for change.

Introduction/Background
The complexities of the contemporary healthcare market create barriers to health technology transformation. Multiple payer-provider relationships and delivery models, the knowledge gap between patients and providers, and the complexities of medicine itself are representative of these barriers. Complexity inhibits a predictable linear path toward transformation. Instead healthcare behaves as a complex adaptive system, where change *emerges*, often exhibiting unexpected patterns with unintended consequences. An understanding of how change occurs within this complex system is important to the evaluation of delivery system reform. The landscape created by our nation's policy emphasis on *adoption* of health information technology (HIT) calls for an investigation of evidence of the transformative nature of HIT in healthcare.

Methods
We conducted an integrative review of the literature to examine published evidence for technology transformation in healthcare. Our framework was Gersick's theory of punctuated equilibrium (PE), which defines three distinct components of market transformation in complex systems: deep structure, equilibrium, and revolution (**Figure 1**)[1]. Five databases (MEDLINE/PubMed, Business Source Complete, Social Science Research Network, Web of Knowledge, and Factiva) were searched for publications from 2004 through 2012. Search terms encompassed multiple meanings for "technology" and "transformation" using MESH terms, key words, and free text. The Web of Knowledge database was used for an ancestry search on three foundational articles describing technology transformation and punctuated equilibrium in markets. Factiva was a source for relevant articles in the popular press. Descriptive and evaluative data were extracted and coded based on a uniform classification schema developed from Gersick's PE framework. Interator reliability between 3 coders was calculated for a subset of the data. Directed content analysis was used to evaluate and synthesize the coded data.

Results
From a total of 4,166 candidate publications, 477 met the study's inclusion criteria. 202 addressed Deep Structure, 224 Equilibrium, and 147 Revolution. Interator reliability was 0.73. Across the three components of PE ten themes were identified. The themes, descriptions and publication counts are summarized in **Table 1**: variations in the environment; market complexity; regulations; flawed risk and rewards; theories of acceptance and diffusion; barriers; ethical considerations; competition and sustainability; environmental elements, and internal elements.

Discussion
While there may be unpublished evidence pertinent to the study's conclusions not identified by this review, the 477 publications described a system tolerant only to incremental change. Contemporary technologies, workflows, imbalances, disparities, and system structure suggest a risk adverse environment with little movement toward actual system transformation. The theory of PE posits the need for *disruption* to counteract the resistance generated by known market characteristics (deep structure) in the absence of a clear superior alternative. We found no evidence of disruptive innovation. Instead, we found elements of deep structure perpetuating an equilibrium state of market fragmentation (e.g. vendor defense of market share creates a barrier to interoperable, user-responsive systems) and resistance to potentially transformative paths (e.g. sensible regulation or uniform standards). Our review identified reinforcing factors including uncertainties associated with highly variable delivery models and reimbursement systems, and inequitable returns on investment. Our nation's reform policies specifically were designed to avoid disrupting the employer-based health insurance system and to encourage incremental HIT adoption. Our findings, in concert with examples from other markets, suggest that consumers are critical drivers of transformation by demanding technologies and unrestricted access to data to use *as they wish* versus *as prescribed* by system gatekeepers who perpetuate market equilibrium.

Acknowledgments: Andrew B. Phillips was supported by the National Institute of Nursing Research (T32NR007969) S. Bakken PI

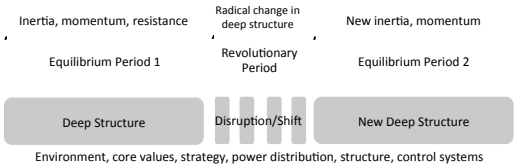


Figure 1 - Three components of Gersick's theory of punctuated equilibrium (deep structure, equilibrium and revolution)[1]

Table 1 - Summary of the Ten Themes Identified in the Literature by Punctuated Equilibrium Component. *Data sources can reflect more than one PE component or theme and not all data sources discussed a specific identified theme.

Identified Theme	Description	N*
Deep Structure		**202**
1. Variations in the environment	The environment of healthcare is defined by factors that influence the adoption and use of HIT, including provider location, size, and HIT vendor capabilities.	85
2. Market Complexity	Healthcare operates within a complex environment characterized by patient confusion, multiple social interactions, data complexity and complex reimbursement systems.	27
3. Regulatory	Regulations guide privacy and security, reporting, reimbursement, liability and standards.	43
4. Flawed Risk and Reward	Incentives to adopt HIT are flawed; inure more to payers and patients than to providers adopting the systems. Fee-for-service reimbursement norms create further disincentives.	24
5. Theories of Acceptance and Diffusion	Several models help describe the patterns of adoption and diffusion of technology within healthcare, including the Technology Acceptance Model and the Diffusion of Innovation theory among others.	27
Equilibrium		**224**
6. Barriers	Data sources highlighted the cost of HIT, lack of human and capital resources, and resistance to change from practitioners as barriers to transformation.	134
7. Ethical Considerations	Ethical considerations contributing to equilibrium include an obligation for technology to do no harm, benefit everyone and not limit ability to practice autonomously.	5
8. Competition and Sustainability	The market economy of the US demands a value driven business case for HIT adoption.	34
Revolution		**147**
9. Environmental Elements	Patient engagement and new models of care represent potential influencers of revolution within healthcare.	74
10. Internal Elements	Change requires effective management, practitioner champions, a shared vision, and a favorable organizational culture.	66

[1]Gersick CJG. Revolutionary change theories: A multilevel exploration of the punctuated equilibrium paradigm. Academy of Management Review. 1991:10-36.

Automated Operative Skill Assessment Using IR Video Motion Analysis

Mihail Popescu, PhD, Christopher Cooper MD, Stephen L. Barnes, MD, FACS
University of Missouri, Columbia, MO

Abstract

One of the greatest challenges encountered during surgery training is objective evaluation of the acquired interventional skills. Expert evaluation by direct observation is subjective and costly. Our automated evaluation method is based on placing IR reflective markers on the surgeon's arms and hands and capturing motion using a Vicon system. The evaluation is performed by comparing the hand traces of a trainee to those of an expert surgeon.

Introduction

As operative interventions become more complex and diverse, acquiring the appropriate skills has become increasingly difficult, especially for low volume, high acuity procedural training. One of the greatest challenges encountered during procedural training is the objective evaluation of the acquired skills[2]. Various computerized methods have been investigated for skill assessment: electromagnetic sensors[2, 3], video camera and a set of surgical gloves with color coded fingers[1] and kinematic measurements recorded by a da Vinci surgical robot[4].

Method and Preliminary Results

We used a Vicon Tracker (www.vicon.com) and gloves with IR reflective markers to simultaneously track and visualize the motion involved in operative procedures. An example of a surgical procedure and its Vicon capture is shown in figure 1. There are 7 Vicon cameras in the environment (one circled in the middle-upper portion of figure 1.left) which make the caption procedure robust to occlusions. In figure 1.left, 5 IR markers are placed in key positions (shoulder, elbow, wrist, back of the hand and index finger) on each arm of a resident surgeon that performs the procedure. In figure 1.middle the marker positions are shown in 3D using a "ball-and-stick" representation.

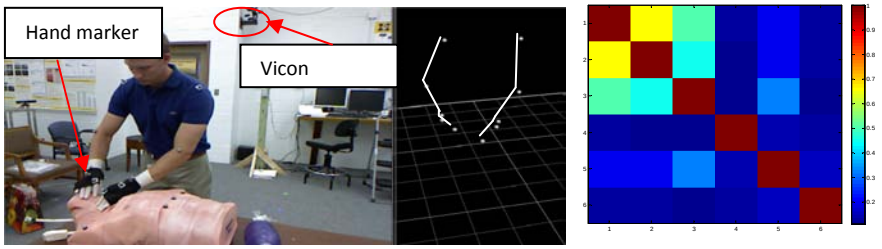


Figure 1. Surgical procedure captured using Vicon (left). 3D arms representation (middle). Similarity matrix for 6 cricothyroidotomies (right)

A cricothyroidotomy was performed 6 times by an experienced senior resident surgeon and instructor of the procedure in our simulation center. The first 3 repetitions ("good") were performed in conformity with the established procedure, while the last 3 ("bad") were conducted in an erroneous fashion. We used a procrustes analysis[5] approach to compute the similarity between two procedures (repetitions) based on the recorded marker traces. The result is shown in figure 1.right. We can see that the similarity between the first 3 repetitions (1, 2, 3) is higher than 0.5, while none of the "bad" ones (4, 5, 6) is more than 0.3 similar to any of the "good" ones. If a library of operative procedures were available, we could use the similarity approach to automatically assess the degree to which a new operative procedure is compliant (or not) to the ones from the library.

Conclusion

This preliminary experiment allows us to conclude that it is possible to automatically assess surgical skills using Vicon video data analysis through a comparison to expert performance.

References

1. Chen J, Yeasin M, Sharma R. Visual modeling and evaluation of surgical skills. Pat. Anal App. 2003;6:1-11.
2. Darzi A et al. Assessing operative skill: needs to become more objective. BMJ. 1999;318:887-888.
3. Dosis R et al. Synchronized Video and Motion Analysis. Arch Surg. Mar 2005; 140:293-299.
4. Lin HC et al. Towards automatic skill evaluation . Computer Aided Surgery. Sept. 2006;11(5):220-230
5. Gower JC, Dijksterhuis GB. Procrustes Problems. Oxford University Press. 2004.

Unsupervised Time-Series Clustering for Identifying Uncontrolled Type-2 Diabetic Patients

Patric V. Prado, B.S.[1], Chunhua Weng, PhD[2]
[1]Columbia University Mailman School of Public Health, New York, NY;
[2]Columbia University Department of Biomedical Informatics, New York, NY

Introduction

The Health Information Technology for Economic and Clinical Health (HITECH) Act, as part of the American Recovery and Reinvestment Act (ARRA) of 2009, has widened the adoption of Electronic Health Records (EHRs), ensuring the availability of a vast amount of potentially useful clinical information.[1,2] It is imperative to develop analytical methods to make sense of regularly collected EHR data and to generate actionable knowledge for clinical decision support. An example application is to identify patients with uncontrolled medical conditions and recommend them for care to clinicians. This study contributes an unsupervised time series clustering analysis method for identifying uncontrolled diabetics using three lab variables: A1C, Creatinine, and blood glucose values.

Methods

De-identified Hemoglobin A1C, Creatinine, and Glucose values were extracted from EHRs for 26,120 patients. We performed record deduplication and removed records with missing values. The dates associated with each lab value were converted to numerical distance in days from the first visit that recorded an A1C value. These dates were then binned into 120 days (to reflect the American Diabetes Association (ADA) recommendations for testing)[3], A1C (and later Creatinine) values which occupied the same 120 day period were averaged, and missing values were linearly imputed between two real data points. A distance matrix was calculated using a first order discrete wavelet transform for every patient's time series of values and then clustered using partitioning around medoids.[4,5] Using a silhouette width metric, the optimal number of clusters was estimated.[6] Glucose plasma values were similarly binned and average glucose values per time period were compared against A1C controlled/uncontrolled diabetic grouping. The data was then explored and intrinsically evaluated through plotting cluster group means and standard deviations, as well as using density plots to characterize total data distribution, weighted averages, and the proportion of every patient's A1C and Creatinine values under a cut-off signifying controlled diabetes or 'normal'. We chose 7 as the cutoff for A1C (signifying control), as recommended by the ADA standards.[3] As the value for normal Creatinine level is dependent on race, age, and gender[7]; the optimal cutoff was calculated using ROC analysis.

Results

Diabetics were clustered into controlled and uncontrolled A1C groups, with values discriminating around the ADA recommended 7% A1C value without any input as to its significance (**Figure 1**). 98.9% of the 'uncontrolled group' had their total average A1C value above 7% and 84.0% of the 'controlled group' had total values below 7% (signifying control). The proportion of patient A1C values < 7% were also calculated and graphed (**Figure 2**). About 92.6% of the uncontrolled group had less than 49% (ROC defined) of their values < 7%, while 89.1% of the controlled group had more than 49% of their values <7%. When the analysis was carried out to 20 time-points, similar results were observed. The total average Creatinine value showed the best separation between groups by the ROC defined 1.05 value (Sensitivity 95.7%; Specificity 95.92%) (**Figure 3**). There seemed to be no association between uncontrolled A1C and Creatinine values, perhaps due to data incompleteness. After expanding the analysis to include all patients, 3 clusters were prominent: a controlled group (values<1, classifies 98.4% correctly), uncontrolled group (>1, 77.8%), and a small but significant super uncontrolled group (> 2.22, 84.76%) (**Figure 4**). The glucose values, collected during the same time period as the A1C values, showed that 73.94% of the uncontrolled group had average glucose values above 154 mg/dL (corresponding to 7% A1C) and 63.66% of the controlled group had average glucose values below 154. The lack of higher correlation between glucose and A1C warrants further investigation and may be because these records are not complete and are obtained at point of care.

Discussion

These preliminary results show the promising use of time-series clustering methods for the identification of clinical groups that are reflective of more general patterns in laboratory values. As some studies show[8,9], because of the limited reliability of some point of care laboratory tests, general trends may be more clinically valuable than isolated individual measurements. Furthermore, the clustering of these patients along clinically significant values also show promise in the future utility of such a method with applications generalizable to other laboratory values and diseases.

Acknowledgments

This study was sponsored by the National Library of Medicine grant R01LM009886 (PI: Weng).

References

1. Congress OHE. Health Information Technology (HITECH Act). Health IT; 2009.
2. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc. 2013:20:117-21.
3. American Diabetes Association. Executive summary: Standards of medical care in diabetes—2009. Diabetes Care. 2009:32:S6-S12.
4. Zhang H, Ho TB, Zhang Y, Lin M. Unsupervised feature extraction for time series clustering using orthogonal wavelet transform. INFORMATICA-LJUBLJANA. 2006:30:305.
5. Reynolds A, Richards G, de la Iglesia B, Rayward-Smith V. Clustering rules: A comparison of partitioning and hierarchical clustering algorithms; JMMA. 1992:5:475–504
6. Kaufman L, Rousseeuw PJ. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York. 1990.
7. Jones CA, McQuillan GM, Kusek JW, Eberhardt MS, Herman WH, Coresh J, Salive M, Jones CP, Agodoa LY. Serum creatinine levels in the US population: Third National Health and Nutrition Examination Survey. American journal of kidney diseases. 1998:32:992-999.
8. Rebel A, Rice MA, Fahy BG. The accuracy of point-of-care glucose measurements. J Diabetes Sci Technol. 2002:6:396-411.
9. Lenters-Westra E, Slingerland RJ. Hemoglobin A1c point-of-care assay; a new world with a lot of consequences! 2009:3:418-423.

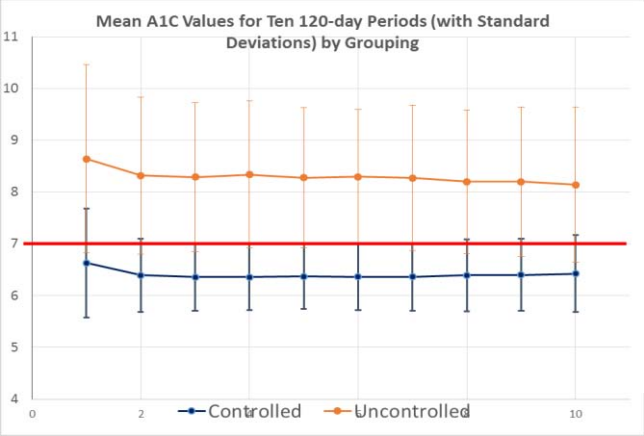


Figure 1. Mean A1C values and standard deviations per 120-day period for each grouping of patients by cluster (red horizontal line is 7% A1C value) (N=2365)

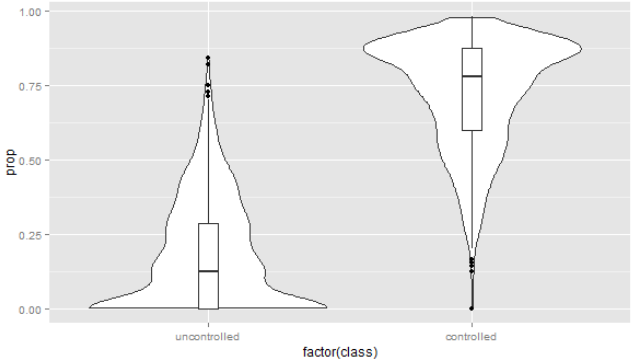


Figure 2. Box and Whisker Plot (with Overlaid Violin Plot) showing distribution of the proportion of every patient's A1C values < 7% (signifying controlled diabetes)

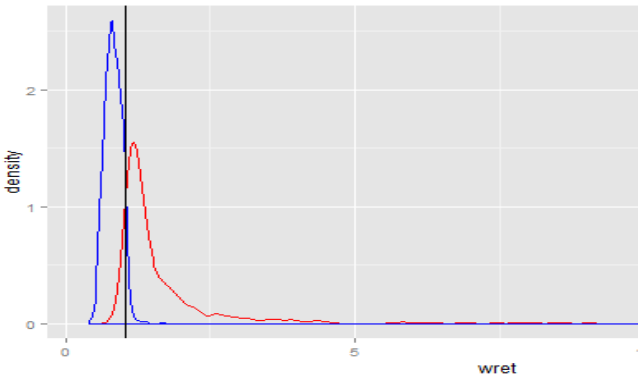


Figure 3. Density plot of mean Creatinine value for every patient by cluster (with 1.05 cutoff) (N=1601)

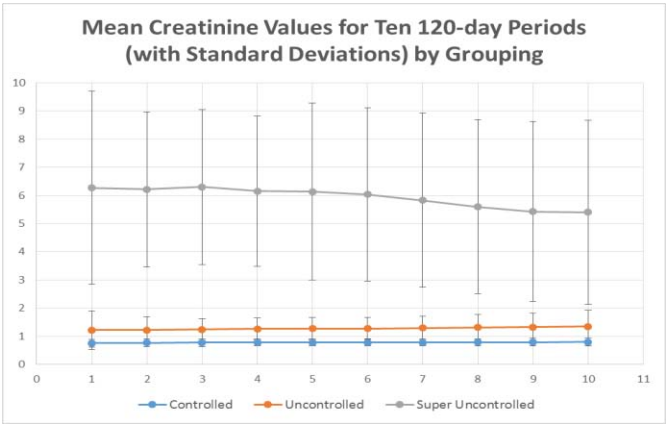


Figure 4. Mean Creatinine values and standard deviations per 120-day period for each grouping of patients by cluster (N=8788)

A Checklist for Go-Live Assessment of Bar Code Medication Administration: Incorporating Human Factors in Implementation Efforts

C. Adam Probst, PhD[1]; Caton Cadigan, MA, BA, BSN, RN, CCRN[1]; Richard Gilder, MS, RN-C, CNOR, NI[1]; Cortney Dalcour, RN, BSN[1]; Chris Matta, CPhT[1]; Donna Montgomery, RN-BC, BSN, MBA[1]; Toni Johnson-Akers, RN-BC, BSN, MBA[1] and Yan Xiao, PhD[1]

[1]Baylor Scott&White Health - Dallas, Texas

Introduction

More hospitals are implementing Bar Code Medication Administration (BCMA), which has been shown to reduce medication errors by as much as 40%[1]. The potential high cost ($40,000 per BCMA bed[2]) and difficulties in implementation (resolving coordination challenges[3]) led our organization to use human factors specialists in developing a checklist to guide go-live preparation and readiness assessment efforts. The checklist was intended to be widely disseminated to every unit at every hospital in the system, used by staff not formally trained in human factors and help avoid workflow barriers and resulting workarounds via encouraging proactive evaluation of units.

Methods

Two of the organization's fourteen facilities had implemented BCMA, with the 12 remaining required to go-live within a constricted timeframe in order to meet requirements for Meaningful Use attestation. Therefore, the system was in need of a tool to quickly determine the order in which the facilities would go-live with BCMA. However, due to variances in the purchase and installation of hardware, an assessment of each facility's current state was required in order to inform the go-live timeline. It was determined that a checklist, which could be completed quickly with little training, was an ideal tool. To that end, a human factors specialist employed multiple methods in order to determine the most critical elements in need of evaluation for go-live feasibility. First, a contextual inquiry was performed on multiple units already live with BMCA in order to identify workflow challenges and better understand key barriers to efficiency (e.g., having adequate space close to the computer to spread out medications while scanning). Secondly, time and motion studies were conducted in order to quantify workflow before and after BCMA implementation, which allowed the system to evaluate the impact to nursing efficiency created by the additional steps needed for BCMA (e.g., log in to each inpatient room computer prior to scanning). Third, a cognitive walkthrough identified bottlenecks in workflow (e.g., halting the scanning process to interact with computerized alerts). Finally, staff surveys and semi-structured interviews provided subjective assessments of workload and opportunities for improvement. Based on these studies, an infrastructure readiness assessment checklist (Fig 1) was created. The checklist, completed by unit nursing managers, consisted of 15 items that covered three domains of BCMA: inpatient rooms (e.g. placement of computers installed at the bedside), medication rooms (e.g. mapping and evaluating nursing workflow for medication retrieval) and workstations on wheels, WoWs, (e.g. assessing environmental constraints for WoW utilization). The checklist reflected known barriers identified from the studies and previous go-lives in an attempt to proactively avoid negative impacts to staff workflow by having each facility assess the most commonly encountered issues: the technology, the staff, policies and needed education.

Results

The checklist was well-received by the 12 facilities in helping to understand potential challenges in BCMA implementation. It was viewed as an effective method to incorporate the human factors findings with lessons learned from previous go-lives and determine what each facility needed to accomplish in order to proactively counter known barriers created by BCMA go-live. Our organization decided to use the checklist to help evaluate the readiness of each facility by summing all 'yes' responses to checklist items for each unit, as one data source for determining the order in which facilities would go-live with BCMA.

Discussion

The checklist approach was an effective way to incorporate human factors into preparation of BCMA go-live. The checklist may be used by those with only basic awareness training in human factors (i.e. not formally trained) and may be widely deployed for all BCMA areas and can help organizations prevent negative impact to staff workflow and threats to safety resulting from workarounds created by suboptimal implementation strategies.

BCMA Implementation Infrastructure Readiness Assessment Checklist

Instructions:

The Baylor Health Care System BCMA Implementation Readiness Assessment Checklist is comprised of 3 sections of concerns: inpatient rooms, medication rooms, and workstation on wheels (WoWs). In order to better assess the readiness of your facility's infrastructure for go-live, we ask that you complete this checklist for each care area where BCMA will be utilized (one checklist per care area).

Scoring:

For each 'Yes' response, a point is given. There are no points awarded for a 'No' response. In order to calculate a score for each section, simply total all 'Yes' responses and fill in the associated 'Score' cells. The total number of 'Yes' values can be summed to create a global score when all sections are complete. The maximum score is 15 (the maximum score for each section is 5); with a higher score indicating a more thorough BCMA infrastructure assessment.

Facility: _____

Area of Care (please circle one): Med/Surg Tele ICU NICU PACU L&D Postpartum Dialysis

BCMA Implementation Infrastructure Readiness Assessment Checklist	Yes	No
Inpatient Rooms (for workstations installed at the bedside)		
Has *every unit* been evaluated for an installation location individually, including isolation rooms, accounting input devices (e.g. dongle, mouse, keyboard)?		
Have mock-up BCMA workstations been evaluated in the different identified locations and use scenarios evaluated (e.g. facing the Pt, avoiding family/visitors, etc)?		
Has the ergonomic adjustability of the BCMA workstation been considered?		
Does the area immediately surrounding the BCMA workstation contain adequate space for users to spread out and scan, prepare (e.g. crush), and sort medications?		
Have alternative computers been secured for situations in which the primary BCMA workstation is broken or malfunctioning?		
	Score:	_____
Medication Rooms		
Has the workflow of how nurses retrieve medications been mapped and evaluated?		
Was the option of profiling the Omnicells evaluated?		
Is there a process for storing medications that are not profiled in the Omnicell?		
Is a computer workstation readily available and located next to the Omnicell?		
Has the storage and charging processes of the BCMA scanners been evaluated?		
	Score:	_____
Workstation on Wheels (WoWs)		
Has the need for a WoW, environmental considerations (e.g. door width, in-room space, carpet halls, etc), and *all* Wow : staff ratios been evaluated (i.e. RN and RT)?		
Have the WoW requirements *of each unit* been evaluated (e.g. battery charging stations, extra batteries, WoW storage, network connectivity, etc)?		
Have infection control issues (e.g. disinfecting the device) been evaluated?		
Do the WoWs have adequate space for medication preparation (e.g. crush), administration (e.g. scanning), and storage (e.g. lockable drawers)?		
	Score:	_____
	Global Score:	_____ **out of 15**

Figure 1: Baylor Scott&White Health Human Factors BCMA Infrastructure Assessment Checklist

Acknowledgments:

All nursing managers who completed the checklist for their unit at the twelve facilities.

References

1. Poon EG, Keohane CA, Yoon CS, et al. Effect of bar-code technology on the safety of medication administration. New Eng J Med. 2010; 362: 1698-1707.
2. Sakowski J.A. (2013). The cost of implementing inpatient bar code medication administration. Am J Manag Care. 2013; 19(2): e38-e45.
3. Novak LL, Anders S, Gadd CS, Lorenzi NM (2012). Mediation of adoption and use: a key strategy for mitigating unintended consequences of health IT implementation. J Am Med Inform Assoc. 2012; 19(6): 1043-1049.

Facilitating Optimal Patient & Family Engagement in Meaningful Use Stage 3

Jaclyn Rappaport, MPP, MBA [1]; Sara Galantowicz, MPH [1]; Anisha Illa, BS [1]; Andrea Hassol, MSPH [1], Charles S. Sawyer, MD, FACP [2], Jean Adams, RN, ACIO [2]; Sidney N. Thornton, PhD [3]; Shan He, PhD [3]
[1] Abt Associates Inc., Cambridge MA; [2] Geisinger Health System, Danville, PA
[3] Intermountain Healthcare, Salt Lake City, UT

Background: The Centers for Medicare and Medicaid (CMS) Electronic Health Record (EHR) Incentive Program incentivizes providers to adopt standards for "Meaningful Use" of certified EHR Technology. The Meaningful Use program is tiered into stages: Meaningful Use Stages 1 and 2 have entered implementation; objectives for Meaningful Use Stage 3 (MU3) are currently being developed by the Heath IT Policy Committee (HITPC), with a Notice of Proposed Rule-Making to be released in 2014. In partnership with the HITPC, the Agency for Healthcare Research and Quality (AHRQ) funded twelve projects to provide feedback to the HITPC to inform development of MU3 objectives; we report here on one of these AHRQ projects. Abt Associates Inc., in collaboration with Geisinger Health System and Intermountain Healthcare, focused on pilot implementation of select MU3 objectives and criteria. This presentation addresses objectives in the *Patient and Family Engagement* domain, which include: patients being able to "send summary of care documents to designated recipients"; to submit "patient-generated health information" online, and to "allow providers to receive, review, respond… and record patient-generated health data".

Methods: The two health systems pursued implementation of select MU3 objectives, customizing their HIT systems when necessary. Researchers held a series of bi-weekly, semi-structured telephone interviews with informatics leaders from both partner health systems over the course of five months to discuss implementation experience. They tracked and coded the qualitative data to identify and elucidate emerging themes. This process yielded: recommendations to improve the language of the MU3 objectives; recommended vendor functionality to support MU3; suggested evaluation metrics (including exemptions) for the objectives, and ways organizations could maximize the value of the objectives relative to their own strategic priorities. The bi-weekly calls with partners were supplemented by a one-time panel of health IT leaders from health systems across the national landscape.

Results: We found that integrating patient generated information into EHRs and information exchange poses significant challenges in terms of (1) validating the identity of the patient and their providers; (2) adjudicating and reconciling patient-provided content with existing electronic medical records, and documenting an existing patient-provider relationship; (3) obtaining and documenting patient authorization for exchange of their electronic medical information and (4) engaging patients meaningfully. The final report offered feedback to federal officials for consideration, including: ensuring patient-generated content is properly identified and reconciled into the medical record and authorized for sharing with health care teams; supporting a variety of methods for identifying patients, and for providers, using a national directory; and allowing patients to easily and securely access forums in which to exchange information about their healthcare (e.g., SMS communication or mobile applications on smart phones) with clear, patient-friendly instructions for doing so. EHR vendors, Health Information Exchanges, and other HIT service providers will all need to play a role in developing technologies and platforms supporting this evolution.

Discussion*:* The experience of MU1 and MU2 implementation underscores the need to elicit feedback from the field as part of MU3 rule-making. The goal of *Patient and Family Engagement* objectives should be to provide these stakeholders with accessible opportunities for communication, coordination and decision-making regarding their health care. Feedback on MU3 implementation highlighted that increased complexity in MU3 *Patient and Family Engagement* objectives will require that entry, validation and exchange of data using electronic tools be increasingly simple, accessible and secure. Healthcare providers, policy-makers, and vendors/ HIT service providers will need to collaborate to create innovative solutions to achieve these outcomes.

Acknowledgements: Funding was received from AHRQ through Contract HHSA290201000031, Task Order 5. The views and opinions expressed are solely those of the authors and do not reflect the official positions of the institutions or organizations with which they are affiliated or the views of the project sponsors.

Why Adherence to HL7v2 Falls Short for Microbiology Data, and What to Do About It: Implementation of a Regional Electronic Infection Control Network

Marc B. Rosenman, MD[a,b], Shahid Khokhar[b], James Egg[b], Larry Lemmon[b], Kinga A. Szucs, MD[a], S. Maria E. Finnell, MD, MS[a,b], David C. Shepherd, DO, MBA[c], Jeff Friedlin, DO[d], Xiaochun Li, PhD[b,e,f], Abel N. Kho, MD[b,g]

[a] Indiana University School of Medicine, Department of Pediatrics, Indianapolis, IN
[b] Regenstrief Institute, Indianapolis, IN
[c] Shepherd Internal Medicine, Indianapolis, IN
[d] Jeff Friedlin, Consultant, Indianapolis, IN
[e] Indiana University, School of Medicine, Department of Biostatistics, Indianapolis, IN
[f] Fairbanks School of Public Health, Indiana University, Indianapolis, IN
[g] Feinberg School of Medicine, Northwestern University, Chicago, IL

Introduction

Recent outbreaks and deaths caused by gram-negative rod multidrug-resistant organism (GNRMDRO) "superbugs" such as carbapenem-resistant *Enterobacteriaceae* (CRE) among hospitalized patients have elicited national attention. The Chief Medical Officer of England warns that antibiotic resistance could set health care back to the early 19th century. Unfortunately, in electronic health records and, especially, health information exchanges, microbiology culture data are often, in whole or in part, unwieldy text blobs, unsuitable for decision support and unanalyzable even for retrospective studies. The principal informatics problem is that most microbiology messages are not structured in standard HL7v2 format by the hospitals that send data; the messages and the database records take on a wide variety of forms and content. In consultation with infection preventionists (IPs) from diverse hospitals, our objective was to build a regional network to deliver alerts when a patient with a history of GNRMDRO is admitted to a hospital or emergency department (ED), to hasten initiation (when appropriate) of contact isolation.

Methods

We parsed data from 27 hospitals to build 1) an HL7v2 correction engine that deals with incorrect microbiology message structure and content, 2) an admission/discharge/transfer message processor with "global" identity matching, 3) decision support to identify superbugs of interest, and 4) secure email alerts to the IPs at the admitting institution. The HL7v2 correction engine builds upon our thorough review of the inbound HL7v2 structure/content patterns; it also uses open-source and project-tailored natural language processing to parse key data elements needed for the alerts: organism, antibiotics tested, minimum inhibitory concentrations, susceptibility interpretation, body source of the culture, and health care facility where drawn. The five GNRMDRO categories are 1) *Enterobacteriaceae* positive for extended-spectrum beta lactamase (ESBL-E), 2) CRE, 3) *Pseudomonas* resistant to 3 of 4 antibiotic classes, 4) *Acinetobacter* resistant to 3 of 4 antibiotic classes, 5) other GNR with pan-resistance. Because our method deals with microbiology data in general, we are also poised to send alerts for patients with MRSA and Vancomycin-Resistant Enterococci (VRE), and we can deal flexibly with other pathogens. We report here on the first 6 weeks of alerts. The 6 weeks were preceded by a 17-week run-in period in which the microbiology HL7v2 correction engine began generating its registry of patients. We were particularly interested in patient cross-over between institutions.

Results

In the first 6 weeks of the email alerts, the IPs were alerted to 63 distinct patients: 41 admitted to EDs, 21 to inpatient wards, and 1 to a pre-operative unit. Five distinct hospital systems were alerted, for patients admitted to 15 distinct hospital campuses. The GNRMDROs in the alerts were ESBL-E (N=53 [84%]: 46 *E coli*, 5 *K pneumoniae*, 1 *P mirabilis*, 1 *C freundii*), CRE (N=7 [11%]: 6 *K pneumoniae*, 1 *S marcescens*), *Pseudomonas* (N=2 [3%]), and other (N=1 [2%]). Body sources were urine 49 (78%), blood 6 (10%), wound 4 (6%), and 1 each for bronchoalveolar lavage, sputum, bile, or missing. For 19 (30%) of 63 patients, the admitting hospital system was different from where the GNRMDRO culture had been drawn. These 19 reflect 15 distinct pairs of "from, to"

locations {from the culture location, to the subsequent admission location}. Precision was 62/63, 98.4% (95% CI 95.4-100%). One patient had a correctly-parsed HL7 message with *E coli* ESBL, but the hospital sent a subsequent HL7 for the same culture, without ESBL.

Conclusion

Even in just 6 weeks, the amount of cross-over between health care facilities -- by patients colonized or infected with dangerous GNR superbugs -- is striking. The IPs find the alerts to be valuable, especially when the culture was from a hospital system different from the admitting institution. Based in part on a previous (VAX) version of MRSA and VRE email alerts in our region, this microbiology data "push" application may markedly hasten placement (at the admitting hospital's discretion) of relevant patients into contact isolation, and thereby may contribute to reducing the spread of life-threatening bacteria. Future plans include MRSA and VRE alerts, and analyses of recall and cost-effectiveness.

Successful Calculation of Kidney Failure Risk Using the Consolidated Clinical Document Architecture Standard

Lipika Samal, MD, MPH[a,b], Adam Wright, PhD[a,b], John D. D'Amore, MS[c], Beatriz Rocha, MD, PhD[a,b,d], David W. Bates MD, MSc[a,b,d]

[a]Brigham and Women's Hospital, Boston, MA, [b]Harvard Medical School, Boston, MA, [c]Diameter Health, Inc., Newton, MA, [d]Partners Healthcare System, Boston, MA

Introduction/Background

In order to deliver efficient, safe, and high quality patient care, electronic health records (EHR) must support chronic disease management. EHRs can improve disease management through clinical decision support (CDS). Today, most CDS is locally deployed within EHRs which requires specialized knowledge of EHR database structure and terminology. The limitation of this approach is that CDS tools can neither be operated independent of an EHR nor transported across practices. Another approach is to extract clinical information from the EHR and to input the data into CDS applications separate from EHRs.

Stage 2 of Meaningful Use (MU2) requires that 10% of transitions of care be accompanied by an electronic exchange of a summary document. [1] This summary of care record must conform to the Consolidated Clinical Document Architecture (C-CDA) standard. The move toward using standardized structured clinical documents brings two modernizations: 1) consistent tagged format and 2) consistent use of national vocabularies. Due to MU2, there will soon be widespread production of C-CDAs which can provide the clinical data necessary to drive CDS. We sought to develop an application to predict kidney failure risk using a C-CDA as input.

Methods

We identified a validated risk prediction model for kidney failure and implemented it as a browser-based application utilizing three open-source javascript libraries (Bluebutton.js, JQuery, and JQuery UI).[2] This application parses C-CDA documents, extracts the appropriate laboratory and demographic data, calculates both estimated glomerular filtration rate (eGFR) and kidney failure risk, and finally provides clinical decision support to the end user (Figure 1).

We tested the application using sample C-CDAs from five MU2 certified EHRs (Table 1). We manually added appropriate laboratory results to these five C-CDAs. Then, we ran further tests using five C-CDAs from the Partners EHR. The laboratory values were added to these additional test patients in the same manner as if they were generated by the laboratory information system for real patients. We compared the results for both sets of test patients against a calculator available as an online supplement to the risk prediction model.[2]

Results

We found complete agreement between risk estimates calculated manually and risk estimates calculated by the C-CDA application (Table 1).[2]

Discussion/Conclusion

We have demonstrated that a browser-based application can successfully calculate kidney disease risk by using a C-CDA as input. The application was created using open source libraries, an open repository of C-CDAs, and a risk prediction model from the biomedical literature. It can either be locally deployed on a device or be provided as a service.

We addressed several challenges during development. One challenge was to extract laboratory values which have multiple LOINC® codes. For example, eGFR may be calculated by two different equations and each has a different LOINC® code. Another challenge was that the application did not extract serum bicarbonate from a basic metabolic panel. We determined that the basic metabolic panel includes total carbon dioxide, rather than serum bicarbonate, as most clinicians refer to it. Further testing is needed to ensure all relevant concepts are appropriately mapped. A limitation is that the application has only been validated with laboratory results from the Partners laboratory information system. However, a strength of the application is that it can integrate with many potential clinical workflows since it uses common web standards. Therefore it can be validated across EHRs and institutions.

In a broader context, this C-CDA input method allows CDS development and dissemination disjoint from EHR vendor functionality and priorities. The adoption of data standards in MU2 makes possible application development for diverse clinical settings and purposes, even outside of the healthcare system.

1. Centers for Medicare & Medicaid Services (CMS), HHS. Medicare and Medicaid programs; electronic health record incentive program--stage 2. Final rule.

2. Tangri N, Stevens LA, Griffith J, Tighiouart H, Djurdjev O, Naimark D, et al. A predictive model for progression of chronic kidney disease to kidney failure. JAMA. 2011 Apr 20;305(15):1553-9.

Figure 1. User interface of kidney failure risk prediction tool.

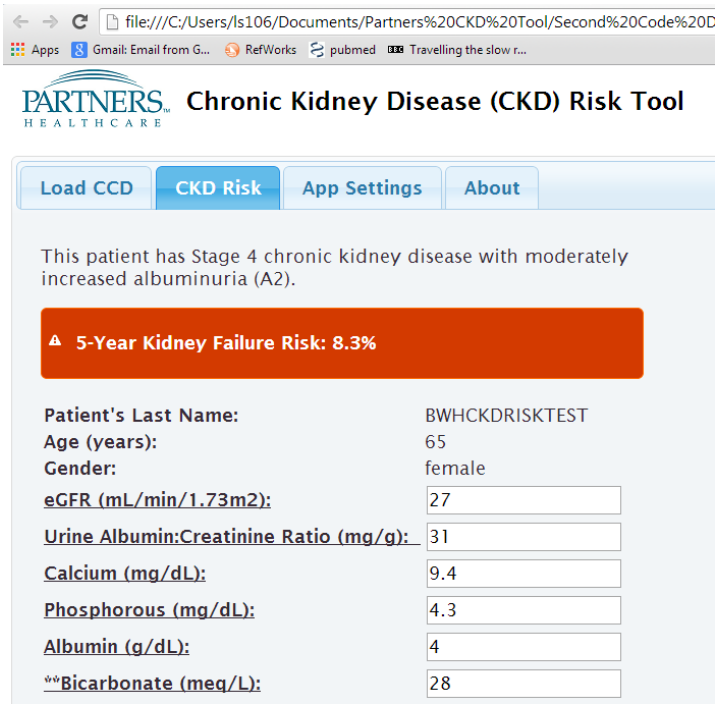


Table 1. Agreement between JAMA calculator and application.

Commercial EHR Test Patients	Age	Gender	Race	Kidney Failure Risk – JAMA calculator[a]	Kidney Failure Risk – application[b]
Allscripts™	67	Female	African American	4.4%	4.4%
Cerner™	72	Female	Asian	0.1%-98.2%[c]	0.1%-98.2%
Greenway™	76	Male	American Indian	0.6%-0.9%	0.6%-0.9%
Vitera™	88	Female	Hispanic	4.1%-25.5%	4.1%-25.5%
Partners LMR	85	Male	White	1.8%	1.8%
Additional Partners Test Patients					
ONETEST	70	Male	African American	11.7%	11.7%
TWOTEST	60	Male	White	7.8%	7.8%
THREETEST	60	Female	White	8.2%	8.2%
FOURTEST	66	Female	African American	14.3%	14.3%
FIVETEST	65	Female	Hispanic	8.3%	8.3%

The five sample CCDAs were downloaded from the SMART repository and laboratory values were added manually. https://github.com/chb/sample_ccdas

a Calculated with a downloadable Microsoft Excel calculator provided by the authors of the article describing the kidney failure risk equation [2]

b Calculated with the C-CDA application

c A risk estimate range is calculated when laboratory data is missing.

Missing Data in an Electronic Health Record-based Population Health Surveillance System

Authors

Lauren Schreibstein, MA[1]; Remle Newton-Dame, MPH[1]; Katharine H. McVeigh, PhD, MPH[2]; Sharon E. Perlman, MPH[2]; Lorna E. Thorpe, PhD[3]; Jesse Singer, DO, MPH[1]; Tiffany G. Harris, PhD[2]; Carolyn M. Greene, MD[2]

[1]Primary Care Information Project and [2]Division of Epidemiology, NYC Department of Health and Mental Hygiene, Queens, NY; [3]CUNY School of Public Health, New York, NY

Background

As EHR adoption increases throughout the United States, more data are becoming available for outcomes research and population monitoring.[1] Missing data in the EHR not only increase the risk of medical error, but also decrease EHR data quality for researchers, health systems and public health practitioners.[2] Missingness can occur in many ways. Data may appear missing because the provider did not collect information, information is recorded in a different part of the record than expected, data are in an unstructured field, or data are only entered in the paper record.[3,4]

The New York City Department of Health and Mental Hygiene (DOHMH) and City University of New York School of Public Health are developing the first EHR-based citywide chronic disease surveillance system, the NYC Macroscope, featuring key prevalence, treatment, and control indicators across 7 domains. The Macroscope is built on "the Hub," a distributed query network that includes 700 practices using one EHR product. Queries are run nightly and aggregate count results transmitted back to DOHMH. We examined missing data to quantify differences by indicator and patient population that could introduce potential bias.

Methods

For New York City residents aged 20 to 100 years who visited a Hub primary care provider in 2012 we calculated the percentages that had the following elements missing in 2012: blood pressure, smoking status, and total cholesterol lab results. To reflect national recommendations, only men 35 years and older and women 45 years and older were included for cholesterol. The percent of missing blood pressure and cholesterol was also calculated separately among those with diagnosed hypertension and hyperlipidemia, respectively.

Results

Among adults aged 20 to 100, recorded blood pressure was missing for 4.6% of patients, but only missing for 2.0% of those with an additional diagnosis of hypertension. For men 35 to 100 and women 45 to 100, cholesterol was missing for 45.5% of patients, but only missing for 32.0% of those with an additional diagnosis of hyperlipidemia. Among adults aged 20 to 100, smoking status was missing for 33.8% of patients.

Discussion

The amount of missing data varies widely across elements of the EHR. Blood pressure documentation was nearly universal, while cholesterol lab values were missing for almost half the patients recommended to have cholesterol screening. This discrepancy may be explained by a short lab lookback window (only 1 year when the screening recommendation is every five years) and the inability of some practices to receive lab results electronically. Higher risk patients had more data present in the record, which may reflect both more attention by providers and more frequent patient visits. These findings indicate that jurisdictions looking to use EHRs for population health monitoring may have the most success using more complete areas of the record like vitals or focusing on high risk populations that frequently utilize care.

References

1. Bayley KB, Belnap T, Savitz L et al. Challenges in using electronic health record data for CER. Medical Care. 2013; 51 Suppl: S80-86.
2. Kahn MG, Raebel MA, Glanz JM et al. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. Medical Care. 2012; 50 Suppl: S21-29.
3. Staroselsky M, Volk LA, Tsurikova R, et al. Improving electronic health record accuracy and increasing compliance with health maintenance clinical guidelines through patient access and input. Intl J of Medical Informatics. 2006; 75: 693-700.
4. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary use of EHR: data qualith issues and informatics opportunities. AMIA Summits Transl Sci Proc. 2010; 1-5.

Identifying Plausible Adverse Drug Reactions Using Knowledge Extracted from the Literature

Ning Shang, MS[1], Hua Xu, PhD[1], Thomas C. Rindflesch, PhD[2], Trevor Cohen, MBChB, PhD[1]

[1] School of Biomedical Informatics, The University of Texas Health Science Center, Houston, TX

[2] National Library of Medicine, Bethesda, MD

INTRODUCTION Pharmacovigilance involves continually monitoring drug safety after drugs are put on the market [1]. To aid this process, automated methods for identifying strongly correlated drug/adverse reaction (ADR) pairs from data sources such as adverse event reporting systems or Electronic Health Records have been developed [2,3]. These methods are generally statistical in nature and do not draw upon the large volume of knowledge embedded in the biomedical literature. In this paper, we investigate the ability of scalable literature based discovery (LBD) methods to identify side effects of pharmaceutical agents. The advantage of LBD methods is that they can explain the plausibility of a drug-adverse event association, thereby assisting human reviewers to validate the signal, which is an essential component of pharmacovigilance.

METHODS We drew upon two large repositories containing knowledge extracted from the biomedical literature by two natural language processing tools, MetaMap [4] and SemRep [5]. We evaluated two LBD methods that scale comfortably to the volume of knowledge available in these repositories. Specifically, we evaluated Reflective Random Indexing (RRI) [6], a model based on concept-level co-occurrence, and Predication-Based Semantic Indexing (PSI) [7], a model that encodes the nature of the relationship between concepts and thus enables reasoning over drug-side effect relationships to recover causal relationships. The patterns of relationships between concepts are referred to as "discovery patterns" [8]. An evaluation set was constructed from the Side Effect Resource 2 (SIDER2) [9], which contains known drug-adverse event relations, and models were evaluated for their ability to "rediscover" these relations.

RESULTS Our results (Table 1) demonstrate that both RRI (MAP is 0.0520) and PSI (MAP is 0.0848) can recover known drug-adverse event associations, and PSI performed better overall (AUROC is 0.6841). With discovery patterns, PSI has the additional advantage of being able to recover the literature underlying the reasoning pathways it used to make its predictions. Taking rosiglitazone and myocardial infarction as an example, the evidence supporting a causal relationship has been retrieved: *rosiglitazone* INTERACTS_WITH *apolipoproteins_b* [10,11] INTERACTS_WITH *ldl_cholesterol_lipoproteins* [12] ASSOCIATED_WITH myocardial_infarction [13], etc.

DISCUSSION In this research, an emerging, scalable method of LBD that uses distributional statistics to infer and apply discovery patterns was adapted to evaluate the plausibility of drug-adverse event relationships for pharmacovigilance. The effective application of large amounts of partially accurate biomedical knowledge to this problem was facilitated by the scalable and robust nature of approximate inference in geometric space. This approach was shown to be more effective than a comparable co-occurrence-based baseline, and has the further benefit of permitting the retrieval of evidence underlying the assertions used by the system to make its predictions. Consequently, our approach provides the means to assist with expert clinical review by providing evidence supporting the plausibility of the connection between drugs and ADRs. Furthermore, the models we have developed can be applied to filter drug-adverse event signals that are detected in spontaneous reporting systems or EHR data, a direction we plan to explore in future work.

Acknowledgements This work was supported by the US National Library of Medicine Grant (1R01LM011563), Using Biomedical Knowledge to Identify Plausible Signals for Pharmacovigilance. This work was also supported in part by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

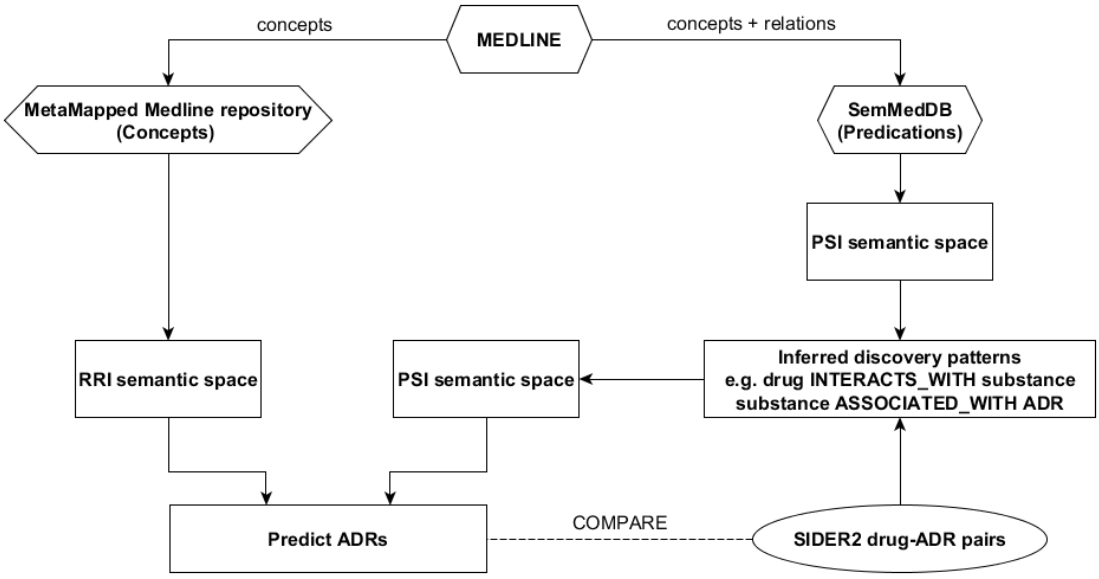


Figure 1 Research overview of detecting known ADRs using LBD distributional semantic models

Table 1 Precision and rank-based measures of detected known ADRs for different groups

Group	mean average precision (MAP)	Precision at 50 for all drugs (global precision)	Median Rank(n=3,436)		AUROC
			median	mean	
Baseline group	0.0300	0.0284	1708	1711.44	0.5021
RRI group	0.0520	0.0784	1333	1454.30	0.5508
PSI group	0.0848	0.1410	808	1108.47	0.6841

References

[1] Mann RD, Andrews EB, editors. Pharmacovigilance. 2nd ed. Wiley; 2007.

[2] Van Puijenbroek EP, Bate A, Leufkens HGM, et al. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. Pharmacoepidemiology and Drug Safety 2002;11:3–10.

[3] Wang X, Hripcsak G, Markatou M, Friedman C. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. Journal of the American Medical Informatics Association 2009;16:328–37.

[4] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp 2001:17–21.

[5] Rindflesch TC, Aronson AR. Semantic processing for enhanced access to biomedical knowledge. Real World Semantic Web Applications 2002:157–72.

[6] Cohen T, Schvaneveldt R, Widdows D. Reflective Random Indexing and indirect inference: A scalable method for discovery of implicit connections. Journal of Biomedical Informatics 2010;43:240–56.

[7] Cohen T, Widdows D, Schvaneveldt RW, et al. Discovering discovery patterns with predication-based Semantic Indexing. Journal of Biomedical Informatics 2012;45:1049–65.

[8] Hristovski D, Friedman C, Rindflesch TC, et al. Literature-Based Knowledge Discovery using Natural Language Processing. In: Bruza P, Weeber M, editors. Literature-based Discovery, Springer; 2008, p. 133–52.

[9] Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P. A side effect resource to capture phenotypic effects of drugs. Mol Syst Biol 2010;6:343.

[10] Brackenridge AL, Jackson N, Jefferson W, et al. Effects of rosiglitazone and pioglitazone on lipoprotein metabolism in patients with Type 2 diabetes and normal lipids. Diabet Med 2009;26:532–9.

[11] Sarafidis PA, Lasaridis AN, Nilsson PM, et al. The effect of rosiglitazone on novel atherosclerotic risk factors in patients with type 2 diabetes mellitus and hypertension. Metab Clin Exp 2005;54:1236–42.

[12] Vessby B, Kostner G, Lithell H, et al. Diverging effects of cholestyramine on apolipoprotein B and lipoprotein Lp(a). A dose-response study of the effects of cholestyramine in hypercholesterolaemia. Atherosclerosis 1982;44:61–71.

[13] Goldberg RB, Kendall DM, Deeg MA, et al. A Comparison of Lipid and Glycemic Effects of Pioglitazone and Rosiglitazone in Patients With Type 2 Diabetes and Dyslipidemia. Dia Care 2005;28:1547–54.

Integration of Generic Electronic Health Records: Moving from Technology Acceptance to Adaptive Structuration to Reciprocal Coordination[1]

Susan A. Sherer, Chad D. Meyerhoefer, Shin-Yi Chou, Mary E. Deily (Lehigh University)
Michael Sheinberg, Donald Levick (Lehigh Valley Health Network)

Introduction: Incentives provided through the American Recovery and Reinvestment Act spurred the adoption of electronic health records [1], which are now being integrated between multiple care settings. Our study addresses the following questions: How do users appropriate vendor-supplied electronic health records? How do they achieve coordinated care through integration of ambulatory EHR systems with hospital information systems?

Methods: We used a longitudinal qualitative multiple case embedded design [2]. Our field site was the Lehigh Valley Health Network, which implemented generic electronic ambulatory records within four obstetric (OB/GYN) practices that varied in patient demographics as well as previous level of record automation. These records were then integrated with the system at Lehigh Valley Hospital's Labor and Delivery (L&D) unit, beginning with the triage unit that evaluates patients during pregnancy. Data were collected from archival sources including meeting notes and workflow documentation and 75 one-hour interviews with both medical care providers and administrative staff in multiple locations at three points in time: (1) initial implementation of records and flow of discrete data from office system to the hospital system; (2) after hospital summary data were integrated into office records, and; (3) one year after both discrete and summary data were flowing between office and hospital record systems. We transcribed and coded all data within NVIVO, beginning with open coding, and then moving to axial and selective coding, using two coders to insure reliability and consistency [3]. Conceptual categories that described themes related to value, reengineering, facilitators, and unexpected consequences were created from the evidence [4]. We used the situated change perspective [5] to understand organizational transformation by analyzing the adaptive structuration patterns [6, 7], work-arounds [8] and the impact on coordination capabilities [9] and unintended consequences [10]. We used an iterative approach to develop a grounded theory that focused on relationships among concepts with strong inter-rater reliability as well as significant differences among respondent demographics (e.g. job title, office location) and implementation time.

Results: Performance and effort expectancy [11] were critical initial factors in the acceptance and use of these records, and varied among participants with different roles. Specific facilitating conditions that influenced these factors were awareness and champions. As the data were integrated, there were structural adaptations, including time and field shifting of data entry and avoidance of and inefficient data retrieval that led to unexpected consequences for coordination. Sequential coordination required the establishment and management of organizational policies; for example, sign-off adherence policies, since sign-offs triggered record availability. Additionally, the use of integrated information was impacted by trust in the data source. For example, users did not typically rely on discrete lab data that flowed between systems; instead they generally resorted to searching for source lab results. As the information flowed in both directions, system and process changes were instituted to achieve reciprocal coordination. System changes occurred when there were imposed-deep misalignments [12]. For example, there were data recording differences between care settings focusing on episodic vs. lifetime care. Process changes were instituted among participants who became more proactive. Reciprocal coordination required new roles and processes that took time and recognition of the importance of standardization.

Discussion: The lack of trust in shared data between multiple systems suggests that integration efforts should support rapid retrieval of source data rather than data flow between multiple systems. New systems should be facilitated via a common platform or one that provides easy retrieval and access to source data. Differences in performance and effort expectancy among roles should be evaluated before adoption, and matched with appropriate facilitators. More successful adoption occurs when complementary process changes are instituted along with the system, rather than after the system is implemented, since structural adaptations can negatively impact coordination. Data entry adaptations can lead to problems with record coordination, which can require organizational oversight. While commercial record systems may be designed to record clinical results, they require improved data retrieval capabilities or complementary investments in new processes. Cultural differences between care settings must be considered when integrating these systems. While organizational policies can support sequential coordination, reciprocal coordination requires buy-in from users, as well as mechanisms to support trust in the data. Shared system usage can support individual and organizational adherence to standardization, but this can take years to achieve.

[1] Support from AHRQ Grant PARA-08-270.

References

1. DesRoches C, Charles D, Furukawa M, Joshi M, Kralovec P, Mostashari F, Worzala C, Jha A. Adoption of electronic health records grows rapidly, but fewer than half of US hospitals had at least a basic system in 2012. *Health Affairs*. 2013;32(8):1478-85.
2. Yin RK. Case Study Research: Design and Methods, 4th edition. Los Angeles: Sage Publications; 2009.
3. Strauss A, Corbin J. Basics of qualitative research. 2 ed. Thousand Oaks, CA: Sage; 1998.
4. Glaser B, Strauss A. The discovery of grounded theory. Chicago: Aldine Publishing Company 1967.
5. Orlikowski WJ. Improvising organizational transformation over time: a situated change perspective. *Information Systems Research*. 1996;7(1):63-92.
6. Orlikowski WJ. Using technology and constituting structures: A practice lens for studying technology in organizations. *Organization Science*. 2000;11(4):404.
7. DeSanctis G, Poole MS. Capturing the complexity in advanced technology use: adaptive structuration theory. *Organization Science*. 1994;5(2):121.
8. Friedman A, Crosson J, Howard JL, Clark E, Pellerano M, Karsh B, Crabtree B, Jaen C, Cohen D. A typology of electronic health record workarounds in small-to-medium size primary care practices. *J Am Med Inform Ass*. 2014;21:e78-e83.
9. Thompson JD. Organizations in action; social science bases of administrative theory. New York: McGraw Hill; 1967.
10. Harrison M, Koppel R, Bar-Lev S. Unintended consequences of information technologies in health care - an interactive sociotechnical analysis. *J Am Med Inform Ass*. 2007;14(5):542-9.
11. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*. 1989;13(3):319-39
12. Sia S, Soh C. An assessment of package-organisation misalignment: institutional and ontological structures. *European Journal of Information Systems*. 2007;16(5):568.

How Can We Partner with Electronic Health Record Vendors on the Complex Journey to Safer Health Care?

Dean F. Sittig, PhD[1], Joan S. Ash, PhD[2], Adam Wright, PhD[3], Dian Chase RN, MSN[2], Eric Gebhardt, MBI[2], Elise M. Russo, MPH[4], Colleen Tercek, BS[2], Vishnu Mohan, MD, MBI[2], Hardeep Singh, MD, MPH[4]

[1]University of Texas Health Science Center, Houston, TX; [2]Oregon Health & Science University, Portland, OR; [3]Brigham & Women's Hospital, Boston, MA; [4]Michael E. DeBakey VA Medical Center, Houston, TX

Introduction

On January 15, 2014, the Office of the National Coordinator for Health Information Technology (ONC)[1] released the SAFER (Safety Assurance Factors for EHR Resilience) guides[2]. Our multidisciplinary group developed these 9 guides, each consisting of 10-25 recommendations covering several key facets of both EHR infrastructure (e.g., system configuration, contingency planning for downtime, system-to-system interfaces) and high-risk clinical processes (e.g., computer-based provider order entry with clinical decision support, abnormal test result reporting, patient identification, and clinician-to-clinician communication). As part of the iterative guide development and testing process, we conducted over 100 interviews and over 100 hours of field observation at eight U.S. healthcare organizations (HCOs). Following release of the guides, the EHRA (Electronic Health Record Association) responded with a detailed critique of these guides, noting that many of the recommendations would require new or improved features, functions and usability[3]. We performed an analysis of our data in an attempt to identify how HCOs were currently working or could work together with their vendors to make the EHR safer and more effective. Our goal was to create a robust foundation of knowledge to foster a HCO-vendor partnership and outline specific roles and responsibilities of EHR vendors in order for them to fully contribute to implementing and maintaining safe and effective EHR-enabled healthcare systems.

Methods

This qualitative study used rapid assessment data collection and analysis processes[4]. We identified all vendor-related quotes from over 2000 pages of interview transcript data previously collected during the SAFER project[5] using keyword searches of vendor names of the systems used at all of the sites (e.g., Cerner, Epic, LMR) coupled with generic terms such as EMR, EHR, vendor, electronic record, and HIT. To perform a content analysis, we divided our research team into 3 groups to review sub-sets of the interview transcripts. Each group member independently reviewed all the transcripts their group was assigned and then the group members worked together to reach consensus on the key themes identified. Next, the groups discussed their findings with the entire multidisciplinary research team to further clarify and refine the themes identified until a set of standard codes was agreed upon. Finally, using this combined thematic coding scheme, each group went back and reviewed and recoded (as necessary) their data. The final results were then reviewed and approved by the entire team.

Results

Table 1 lists and defines the 8 themes the team identified during the content analysis. These themes, ranging from new technical specifications to improved support and training, illustrate the complexity of creating and maintaining safe and effective EHRs. In many instances, representatives of the HCOs expressed a need to work more closely with their EHR vendors to address limitations in functionality and usability of their existing EHRs.

Discussion

Our analysis suggests that while the EHR Developer Code of Conduct[6] is a good start, further improvements in safety are likely to occur only with close partnership with vendors on the eight thematic areas we identified in our work. Within these areas, defining specific roles and responsibilities for vendors could achieve a shared understanding of how to achieve safer and more effective EHR-enabled health care.

Table 1. Key themes that EHR vendors must address if we are to achieve the safe and effective EHR-enabled healthcare system that healthcare organizations want and patients deserve.

Theme	Definition
Technical Functionality of EHRs	Features and functions that EHR developers are responsible for creating, for example CPOE functionality, system-to-system interfaces, system configuration capability (customizability)
Usability	Features of the user interface that describe the "look and feel" as well as the functions available to users
EHR Standards	Requirements, specifications, guidelines or characteristics that are used to increase availability, confidentiality, and integrity of data collection, storage, transmission, and reporting
EHR Testing	Methods to help users and organizations understand whether the EHR and all of its components are working as designed, intended, and required to provide safe and effective care
Workflow Processes	Issues related to how the EHR supports the clinical workflow by facilitating data entry, review, communication, or monitoring
Personnel to Support EHR Implementation and Use	Human resources involved in the design, development, testing, implementation, and evaluation of all aspects of the EHR-enabled healthcare delivery system
EHR Infrastructure	Computers, networks, cables, power, air conditioning, etc. required to support the EHR application
Clinical Content	Information provided to clinicians to help them select the most appropriate tests or treatments and document their actions

References

1. HHS makes progress on Health IT Safety Plan with release of the SAFER Guides. Available at: http://www.hhs.gov/news/press/2014pres/01/20140115a.html
2. Safety Assurance Factors for EHR Resilience (SAFER) Guides. Available at: http://www.healthit.gov/policy-researchers-implementers/safer
3. HIMSS Electronic Health Record Association. Comments on SAFER Guides. Available at: http://www.himssehra.org/docs/SAFER%20Guides%20Comments%20Final.pdf
4. McMullen CK, Ash JS, Sittig DF, Bunce A, Guappone K, Dykstra R, Carpenter J, Richardson J, Wright A. Rapid assessment of clinical information systems in the healthcare setting: an efficient method for time-pressed evaluation. Methods Inf Med. 2011;50(4):299-307. doi: 10.3414/ME10-01-0042.
5. Singh H, Ash JS, Sittig DF. Safety Assurance Factors for Electronic Health Record Resilience (SAFER): study protocol. BMC Med Inform Decis Mak. 2013 Apr 12;13:46. doi: 10.1186/1472-6947-13-46.
6. EHR Developer Code of Conduct. Available at: http://www.himssehra.org/ASP/codeofconduct.asp

Quantifying the Utility of Medical Tests using Longitudinal EHR Data

Stein Olav Skrøvseth, MSc, PhD[1,2], Knut Magne Augestad, MD, PhD[1,3], Shahram Ebadollahi, PhD[2]

[1]University Hospital of North Norway, Tromsø, Norway; [2]IBM Thomas J Watson Research Center, Yorktown Heights, NY; [3]University Hospitals Case Medical Center, Cleveland, OH

Introduction: Increased adoption of Electronic Health Records (EHR) has resulted in availability of sizeable longitudinal medical data. This data could potentially be leveraged to model disease progression among other things.[1] In this work we use such longitudinal EHR data to quantify the value of medical tests performed as part of the routine patient care. A test is any procedure performed to increase the knowledge about the state of a patient; both for diagnostic or screening purposes. The information content of a test can be quantified in a data-driven way using information theoretic measures, contrasted to the knowledge-driven basis often inherent in other approaches relying on the value of information.[2] We define the utility of a test to be a function of the information content of the test over the impact of that test, such as cost, risk or discomfort to the patient.

Methods: We model a patient trajectory using a set of parameters $\boldsymbol{\theta}$. Available to us is a set of tests $\boldsymbol{\xi}$, each of which gives a result $\boldsymbol{y}$ modeled by a random variable $\boldsymbol{Y}$. The purpose of any test regime is to increase our knowledge of the parameters in the model. In an information theoretic framework this is equivalent to reducing the entropy of the posterior distribution of the parameters. Using observational data we design a classification or regression machine such that the probability $\boldsymbol{\pi}(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{\xi})$ of a parameter $\boldsymbol{\theta}$ for a given value $\boldsymbol{y}$ of a test $\boldsymbol{\xi}$ can be estimated. Then, the expected information content of a test in a particular phase is the expected reduction in entropy of that test,[3]

$$\boldsymbol{\Delta}(\boldsymbol{\xi}) = \boldsymbol{S}(\boldsymbol{\theta}) - \boldsymbol{E}_{\boldsymbol{Y}}[\boldsymbol{S}(\boldsymbol{\theta}|\boldsymbol{Y},\boldsymbol{\xi})].$$

To validate the concept, we extracted the full EHR from the gastrointestinal department at the University Hospital of North Norway through 2004-2012 and identified 990 patients who underwent surgery for colorectal cancer along with all blood samples taken in the course of the patient's hospitalization. The patient model is binary with $\boldsymbol{\theta}$ denoting second surgery within 30 days post index surgery. Specifically, we coupled the information content to second surgery for anastomosis leakage (AL), a very severe complication for these patients.[4] The patients' pathways were divided into time periods relative to the point of index surgery, and for all tests of any one kind taken during that period, the information gain was computed.

Results: The information gain for C-reactive Protein (CRP) in the different temporal phases is shown in Figure 1, left. Corresponding validations were done for other tests. The maximal information content for this test related to second surgery is on days 3-4 post surgery, while tests performed prior to 1 day post surgery contain close to zero information. Figure 1, right side, shows the likelihood $\boldsymbol{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ as function of $\boldsymbol{y}$ (denoted *Value*), relaying the reasonable conclusion that higher CRP is associated with higher probability of resurgery.

Discussion: Blood tests are a very cheap and common procedure carrying little to no negative impact. However, their commonplace usage means there is a large amount of data in any EHR on which to base the current approach. The methodology can easily be expanded to other tests, in particular different modalities of radiology that carry costs in terms of economy, discomfort and radiation risks. Additionally, the simple binary patient model used in the current work can be replaced with arbitrarily complex patient models by adjusting classifiers or regression machines accordingly. The main challenge with this approach is having sufficient data for training the classifier and estimating entropy. Also, we assume that that the system is stationary such that the entropy can be consistently estimated, which may not always be the case as healthcare is a constantly evolving enterprise. When using several tests as input there may be significant overlap in their information content which can be separated out using, e.g, independent component analysis (ICA) or similar techniques.

In our analysis, the information content of CRP related to second surgery was associated with higher values of CRP taken on day three after the index surgery. CRP is known to generally increase post surgery,[5] though the connection to utility as defined here has not been made. Our results correlate well with the real world clinical knowledge that there is a utility to CRP 3-4 days after surgery.[6] The information theoretic approach provides a data-driven way for quantifying the information content of tests. Test utility may play a major role when applied to test associated with high negative imapct for healthcare or the patient.

References

1. Murdoch TB, Detsky AS. The inevitable application of big data to health care. *JAMA*. 2013;309(13):1351–1352.
2. Heckerman D, Horvitz E, Middleton B. An approximate nonmyopic computation for value of information. In: *UAI'91 Proceedings of the Seventh conference on Uncertainty in Artificial Intelligence*. Los Angeles, CA; 1991:135–141.
3. Sebastiani P, Wynn HP. Maximum entropy sampling and optimal Bayesian experimental design. *J R Stat Soc Ser B*. 2000;62(1):145–157. doi:10.1111/1467-9868.00225.
4. Kehlet H. Fast-track colorectal surgery. *Lancet*. 2008:791–793.
5. Cole DS, Watts A, Scott-Coombes D, Avades T. Clinical utility of peri-operative C-reactive protein testing in general surgery. *Ann R Coll Surg Engl*. 2008;90(4):317–21.
6. Woeste G, Müller C, Bechstein WO, Wullstein C. Increased serum levels of C-reactive protein precede anastomotic leakage in colorectal surgery. *World J Surg*. 2010;34(1):140–146.

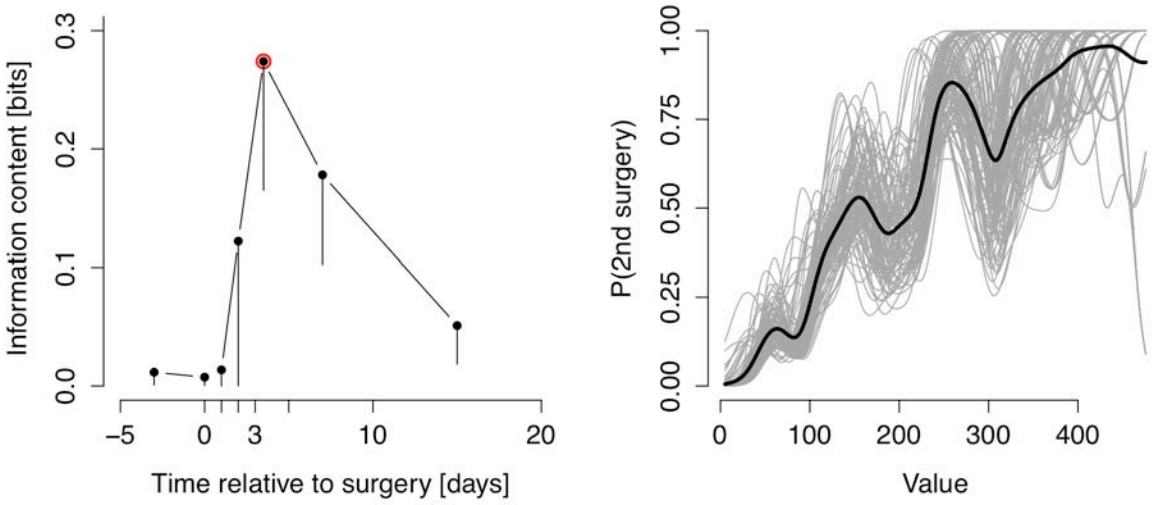


Figure 1: Left: Information content in the value of a CRP test with respect to a second surgery for anastomosis leakage in phases relative to index surgery. Time zero is the day of index surgery. Hanging lines indicate the uncertainty. Right: The classification rule in the phase of maximal information content, indicated by the red circle in the left figure. Thin lines are 100 resamples of the majority class, thick line is the mean classification rule. X-axis is the result of the test.

Identifying Metastases from Pathology Reports in Lung Cancer Patients

Ergin Soysal, MD, PhD[1], Jeremy L. Warner, MD, MS[2], Joshua C. Denny, MD, MS [2], Hua Xu, PhD[1]

[1]School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX; [2]Dept. of Biomedical Informatics, Vanderbilt University, Nashville, TN

Introduction

Metastatic patterns at the time of cancer recurrence are one of the most important prognostic factors in estimation of clinical course and survival of the patient. However, this critical information is rarely recorded in a structured format in cancer registries. This paper describes a system for identifying metastases and their location from free text pathology reports, which offers an attractive solution to this unmet clinical research need.

Methods

We selected 540 pathology reports from 262 patients with lung cancer in the cancer registry who had no metastasis at the time of initial diagnosis and later returned to clinics with metastatic lesions at an unknown site. A physician reviewed each pathology report to identify metastatic lesions. Among those, a random group of 100 patients (217 reports) were separated for use as test subjects for evaluation. Recall and precision values were calculated by comparison of the system output against a physician review on the test subjects.

Diagnosis sections of pathology reports were modeled around the specimen entity (Figure 1) in relation to other named entity classes such as body location/structure, procedure, or histological type. Location information is further classified into metastasis sites such as liver, bone and lymph nodes. Corresponding SNOMED CT® (SCT) concepts were used to create concept subsets, based on "*is a*" relationships to these generic parent concepts – e.g., the histological type subset was formed by concepts that have "*is a*" relationship to |Malignant neoplasm| 367651003 or |Malignant neoplastic disease| 363346000. The system lexicon was prepared from the SCT concept descriptions with respect to these class hierarchies, recursively. For every lexeme, probabilities were calculated for each given class. The lexicon was further enriched by synonyms from UMLS® and variations using LVG and the SPECIALIST[1].

The pathology reports were processed with *sentence boundary detection*, *tokenization* and *section header identification* modules to identify the *diagnosis sections*. Then, *part of speech tagging*, and *chunker* modules generated the phrases (Figure 2). At this stage, phrases were reprocessed for the negation and tumor grade terms using a dictionary based lookup. The identified phrases were then assigned to a class based on naïve Bayesian probabilities calculated for the lexemes forming the phrase using the physician's gold standard. For a given phrase with n lexemes $l_1, \ldots l_n$, the probability of being a member of a class C is calculated as $p(C|l_1, \ldots l_n) = \prod_{i=1}^{n} p(C|l_i)$, where the probability of class C for a given lexeme l is $p(C|l) = (p(l|C) \times p(C))/(p(l))$.

As the final step, rule-based information extraction was performed to identify metastases and their sites with respect to the specimen entity and its relationships (Figure 2). The diagnosis section had certain patterns that can be modeled as finite number of entity flows. Therefore, each entity class was accepted as a state. This approach also provided an advantage for rule based handling of negations. At the end of each path, detected entities were linked to each other with proper relationships based on generic classes, and stored as instances.

Results

The system was tested on 217 reports from 100 patients, who had metastatic disease at an unknown site. Results were compared to physician extracted information. Recall and precision values for metastasis detection were 82.69% and 87.76% correspondingly. Metastasis site detection was found to have a recall of 89.62% and precision of 93.13%. Finally, detection of lymph node involvement had a recall of 83.87% and a precision of 86.67%.

Discussion

We have demonstrated that an NLP approach using naïve Bayes can identify the site of metastatic recurrence with high precision and recall. There are many potential applications for such a tool, like recoding "*unknown metastatic site*" records in existing tumor registry data. The tool could also be generalized to any pathology narrative text programs to provide "rapid learning systems" for cancer care, with a vision of providing clinical decision support based on data collected in near-real-time at a national scale[2].

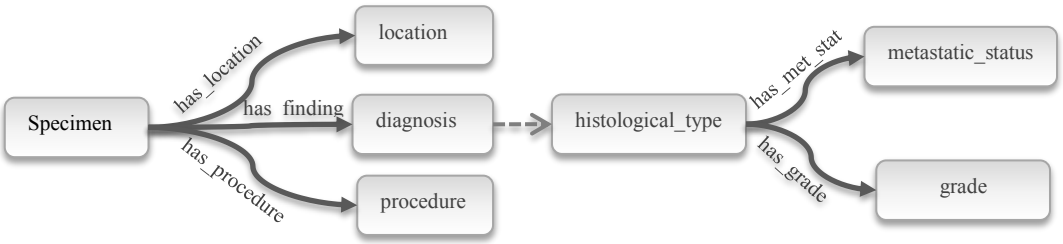


Figure 1. Entities and relationships for a specimen.

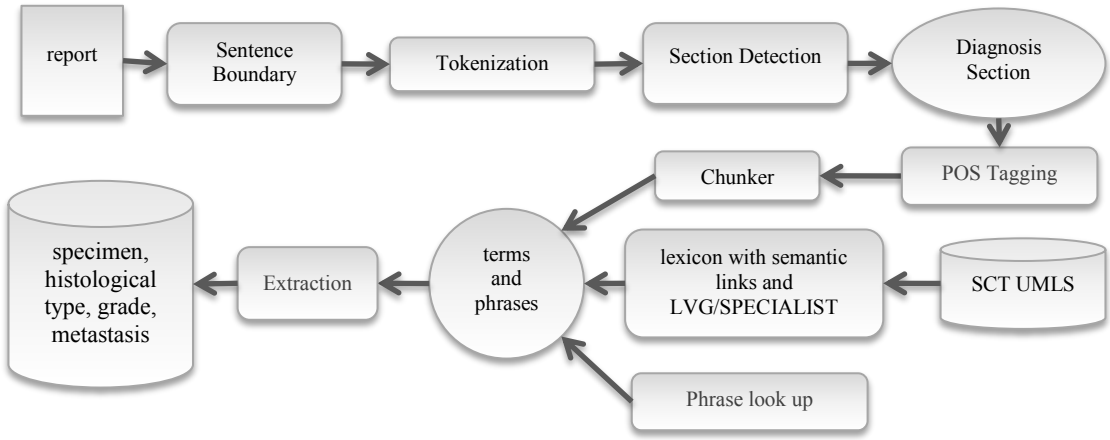


Figure 2. System anatomy. After section detection, rest of the process takes place on diagnosis section.

References

1. McCray AT, Srinivasan S, Browne AC. Lexical methods for managing variation in biomedical terminologies. *Proc Annu Symp Comput Appl Med Care.* 1994:235-239.
2. Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. *Sci Transl Med.* Nov 10 2010;2(57):57cm29.

Uncertainty and Information Seeking in a Diabetes Online Community

Si Sun, MS
School of Communication and Information, Rutgers University, New Brunswick, NJ.

Abstract
This study offers insights into the types of uncertainty that led diabetes patients to seek information in a large online community. A total of 1,440 threads in 2013 were analyzed, 453 of which reflected various types of uncertainty related to information seeking. Uncertainty was classified according to categories identified in existing literature. Among the four major types of uncertainty located in everyday settings, only two are observed in the online environment. This difference points to patients' possible concerns with and the personal difference that eliminates the necessity of online information seeking.

Background
The self-management nature of diabetes treatments combined with the complex and unpredictable symptom patterns make the experience of uncertainty an inevitable part of many aspects in patients' daily lives. This uncertainty, usually resulted from a gap in patients' knowledge could lead patients to seek and share information. Recently, these information seeking and sharing activities have increasingly carried out online. In fact, patients turn to the Internet for health information more often than they communicate with doctors. Many theories of uncertainty and their applications examine why people seek or not seek information as ways to solve a problem or cope with a disruptive event. However, little is known about how information seeking in online environment compares to that in offline settings as influenced by types of uncertainty.

Methods
The dataset used in this study includes 1440 threads from patients in the online community of Diabetic Connect, one of the largest diabetes social networking sites in the United States. The last response to these threads range from January 2013 to June 2013. Text analysis is conducted to identify the types of uncertainty patients mentioned in the thread starters, and the results are compared with Middleton's[1] categorization of uncertainty in type 2 diabetes. The uncertainty related to information seeking is identified with an indication of a cognitive gap, such as "I don't know" and "I'm totally lost", or an affective symptom[2], such as "Afraid", "overwhelmed", and "scared ". This cognitive gap or affective symptom usually follows with explicit or implicit support seeking. For example, a patient expressed need for new information "Newly diagnosed…what did you have for meals? Need some good ideas PLEASE"; another patient also noted her affective symptoms when soliciting help "I get so confused because I … keep getting told I will not feel low or have low BG when taking [medicine]. Anyone else feel lows when while taking [medicine]?" (The quotes above are slightly modified to assure patient privacy.)

Preliminary Results and Discussions
Of the 1440 threads analyzed, 453 thread starters include at least one expression that indicates uncertainty. Other threads that do not indicate uncertainty may (a) share health information collected by patients, (b) share personal experiences, (c) post Q&A on how to use website features of the online community, or (d) call for participation in community activities. The thread starters expressing some level of uncertainty encompass types of uncertainty also identified by Middleton, including (a) health-related uncertainty, such as complex and unpredictable symptoms patterns, diagnosis, treatment and regimen, and prediction and decision making; and (b) personal identity construction and maintenance. However, some types of uncertainty, including (c) social life related uncertainty, such as information sharing and perceived stigma, and (d) financial issues, are not observed in this sample.

This difference may be the result of patients' decision-making concerning whether to share information online for the purpose of information seeking, because topics such as stigma and financial issues are usually considered too sensitive for public discussions. Individual differences may also be a cause for this difference: (a) patients who perceived social stigma may avoid discussing about diabetes in public; (b) patients who interact with others online may have fewer issues with information sharing; and (c) patients who have access to the Internet may be less financially stressed. An interview study is necessary to determine which of the above speculations are true.

Implications
This investigation of the types of uncertainty prompting patients to seek information online versus those in everyday lives enables a rich understanding of the underlying reasons for patients to choose different channels when communicating about various health-related issues. This perspective could help information professionals and healthcare providers understand patients' information needs and the suitable channels for delivering support for these needs.

Reference
1. Middleton AV, LaVoie NR, Brown LE. Sources of uncertainty in type 2 diabetes: Explication and implications for health communication theory and clinical practice. *Health Communication*. 2012;27(6):591-601.
2. Kuhlthau CC. A principle of uncertainty for information seeking. *J Doc*. 1993;49(4):339-355.

Phenome-wide Association Studies Using NLP-Derived Phenotypes

Pedro L. Teixeira MS[1]; Robert Carroll MS[1]; Lisa Bastarache[1]; Peter J. Speltz[1]; Joshua C. Smith, MS[1]; Joshua C. Denny, MD, MS[1,2]
Depts. of [1]Biomedical Informatics and [2]Medicine, Vanderbilt University, Nashville, TN

Introduction: Electronic health records (EHRs) are becoming a powerful tool for clinical and genetic research. In addition to standard phenotype-based genetic studies, EHR-linked biobanks enable phenome-wide association scans (PheWAS). Using PheWAS, one can rapidly search for associations between a genotype and many phenotypes – highlighting potential pleiotropy. Prior PheWAS leveraged ICD9 billing codes as the basis for phenotypes. However, much health information may only be found within narrative EHR text – at potentially increased granularity and accuracy. This work shows that PheWAS performed using NLP-derived phenotypes from problem lists (PL), discharge summaries (DS), and history and physical (H&P) notes, without per-phenotype refinement, can replicate and sometimes improve upon the original ICD9-based method.

Methods: Vanderbilt's BioVU, a DNA biobank, is linked to a de-identified form of the EHR known as the Synthetic Derivative. We used PL, DS, and H&P notes of the first 6,260 European-American individuals entered in BioVU. Each of these documents is typically typed as narrative text. Notes were mapped to SNOMED-CT concepts using the KnowledgeMap Concept Indexer (KMCI) and SecTag, which identifies section context for concepts. We ignored concepts that were negated, possible, and those predicted to be associated with individuals other than the patient (family or otherwise). In addition, we further refined this set to the following UMLS semantic types: findings, diseases or syndromes, therapeutic or preventive procedures, signs or symptoms, neoplastic processes, pathologic functions, and congenital abnormalities. Finally, the less structured format of DS and H&P notes led to a higher rate of false positives using KMCI. We improved performance by limiting results to unambiguous matches with KMCI and the following high value sections (and their subsections): history of present illness, past medical history, hospital course, and assessment and plan. We used concepts via two routes of differing granularity. The first treats each unique CUI as a phenotype, while the second aggregates similar CUIs into a single phenotype by mapping them to ICD9-based codes and then to the existing PheWAS hierarchy of phenotypes. The mapping was constructed from a combination of the ICD9-CM to SNOMED-CT maps available through the National Library of Medicine. Both the 1:1 and 1:many mappings were used to maximize concept coverage. PheWAS was then run using the R PheWAS package for the same five SNPs on the same population as in the original PheWAS demonstration paper. These SNPs are known to be associated with rheumatoid arthritis (RA), multiple sclerosis (MS), coronary artery disease (CAD), carotid atherosclerosis (CAS), Crohn's disease (CD), and atrial fibrillation (AF). All analyses include age and gender as covariates and used logistic regression assuming an additive genetic model. Phenotypes were only considered if there was a minimum of 20 cases.

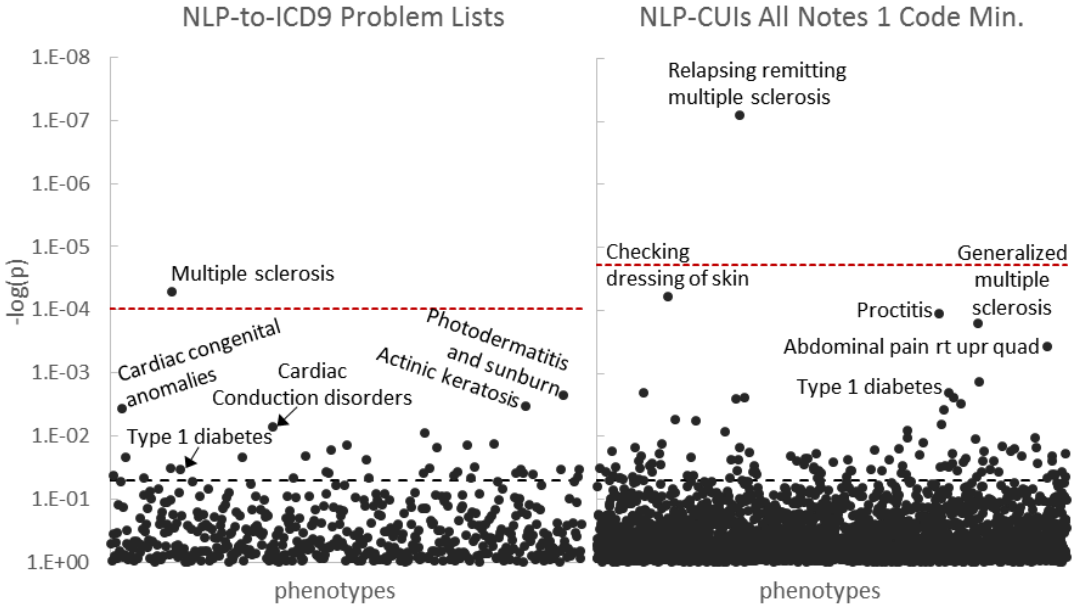
Results: Of the individuals examined 6,232 had at least one note type, 6,145 had PL, 4,026 had DS, and 5,536 had H&P. Table *1* presents the resulting p-values and odds ratios for the five SNPs across the various NLP-based methods in comparison to the original PheWAS paper or relevant GWAS results. NLP-methods have shown similar performance or better across all five SNPs. Substantial improvements in p-value were seen for both AF and RA with both granular concepts and ICD9-aggregated PheWAS concepts. Additionally, the NLP-methods demonstrated associations with concepts more granular than seen with ICD9 aggregation. One such concept was "relapsing remitting multiple sclerosis" with rs3135388 at a p-value of 8.04E-08 and an odds ratio of 5.59 in addition to its association with "generalized multiple sclerosis" at a p-value of 1.63E-04 and odds ratio of 1.75. The former is well above a Bonferroni-corrected threshold of significance of 1.90E-05. Of note, not all NLP sources performed similarly. Fewer cases are detected for NLP-based methods with discharge summaries performing particularly poorly across all diseases except AF.

Discussion: We have shown that NLP-based PheWAS can replicate associations seen with both traditional GWAS and the original ICD9-based PheWAS method. There is improvement for some diseases but issues with acronym disambiguation introduced NLP errors in some documents, especially shorter documents (such as PL) hindering accurate context-based word-sense disambiguation. In addition, the coverage changes depending on document type and condition. For example, typically chronic and less granular concepts are documented in PL. NLP-based methods will likely benefit from further increasing the number of document types included given appropriate filtering of less reliable unstructured text within the EHR.

Table 1. Comparison of NLP-Based PheWAS to Original ICD9-Based Method. The table shows a comparison of cases detected, OR, and p-values across the five SNPs examined with the various forms of NLP-based PheWAS: CUIs mapped to ICD9-based PheWAS codes or fully granular CUIs from a variety of sources. The last uses a more stringent minimum of 2 CUI occurrences to define a case.

SNP	rs3135388		rs17234657	rs2200733	rs1333049		rs6457620
Gene/region	DRB1*1501		Chr. 5	Chr. 4q25	Chr. 9p21		Chr. 6
Disease	MS	SLE	CD	AF	CAD	CAS	RA
Number of Cases							
NLP-to-ICD9 PL	94	93	160	**294**	**730**	**101**	280
NLP-CUIs PL	108	103	170	276	715	39	322
NLP-CUIs DS Filtered	20	21	76	277	24	4	92
NLP-CUIs H&Ps Filtered	76	37	178	112	139	11	251
NLP-CUIs All Notes Min 1	**136**	**110**	**207**	132	755	48	**387**
NLP-CUIs All Notes Min 2	118	105	184	74	723	41	352
Odds Ratios							
Previous OR	1.99	**2.06**	1.54	1.75	**1.20-1.47**	**1.46**	**2.36**
NLP-to-ICD9 PL	2.00	1.24	1.56	1.42	1.15	1.13	1.46
NLP-CUIs PL	1.92	1.26	1.55	1.49	1.10	0.96	1.45
NLP-CUIs DS Filtered	**2.26**	0.81	**1.72**	1.49	1.04	NA	1.55
NLP-CUIs H&Ps Filtered	2.17	1.41	1.54	1.80	0.95	NA	1.54
NLP-CUIs All Notes Min 1	1.75	1.20	1.53	1.69	1.09	0.99	1.43
NLP-CUIs All Notes Min 2	1.83	1.22	1.48	**2.08**	1.09	0.93	1.42
P-values							
NLP-to-ICD9 PL	5.17E-05	0.281	3.31E-03	4.99E-03	**0.019**	**0.390**	2.20E-05
NLP-CUIs PL	5.70E-05	**0.215**	2.72E-03	1.61E-03	0.123	0.871	6.52E-06
NLP-CUIs DS Filtered	2.09E-02	0.658	1.17E-02	2.09E-03	0.901	NA	4.23E-03
NLP-CUIs H&Ps Filtered	**2.71E-05**	0.248	2.49E-03	1.08E-03	0.671	NA	3.85E-06
NLP-CUIs All Notes Min 1	1.63E-04	0.329	**1.21E-03**	2.01E-03	0.157	0.966	**3.22E-06**
NLP-CUIs All Notes Min 2	1.19E-04	0.282	5.20E-03	**4.48E-04**	0.129	0.734	1.22E-05

Figure 1. Manhattan Plot of NLP Collapsed to ICD9 Codes and Granular CUI-Based Phenotypes (rs3135388). Phenotypes with some of the most significant p-values are labeled for NLP-derived concepts mapped to ICD9-based phenotype codes (problem lists) and granular NLP-based concepts - cases have a minimum of one code (all notes). The most significant for each is MS (5.17E-05) and relapsing remitting MS (8.04E-08) respectively.



Integrated Data Management System for Medical Registries: A Case Study using RexDB®

Charles Tirrell[a], Frank J. Farach[a], Meredith Yourd[a], Owen McGettrick[a], Oleksiy Golovko[a], Leon Rozenblit[a]
[a] Prometheus Research LLC, New Haven, CT

Summary: Disease- and procedure-specific registries are powerful tools for the improvement of long-term patient care. Unfortunately, registries are typically too expensive for smaller practices and organizations to develop and maintain. We describe an affordable and scalable alternative: an open source solution that can integrate institutional (EHR) and primary research data.

Introduction and Background: Medical registries are growing both in importance and value with rapid increases in (a) the amount of data on each patient and procedure, (b) the value of tracking outcomes, and (c) the desirability of sharing information between health professionals. Registries enable researchers to track individual patient and outcomes and evaluate cohort-level statistics that guide the improvement of patient care. Prominent examples of recent registry systems include the Intelligent Research in Sight (IRIS) Registry™ [1] and the National Cardiovascular Data Registry® (NCDR) [2]. Two difficulties prevent wider adoption of registries. First, registry applications are expensive to develop, and, where levels of change are high, even more expensive to maintain. Second, the heterogeneity and complexity of source data, especially from electronic health record (EHR) systems, makes it expensive to transform and integrate data using traditional methods. We set out to develop an approach that addresses both difficulties by using an existing open-source stack specifically designed to support heterogeneous research data under conditions of rapid protocol change.

Methods: We developed an integrated registry data management system based on the Research EXchange DataBase (RexDB), an open source (AGPLv3) stack initially developed to help for the acquisition, curation, exploration and sharing of human biomedical and behavioral research data. RexDB allows analysts (without help from programmers) to configure custom data structures, extract-transform-load (ETL) processes, and front-end application screens and workflows. ETL processes utilize HTSQL, an analyst-friendly, open source query language, which allows users to navigate, explore and manage, complex datasets with a simple query syntax. This approach allows analysts who are close to the data and to the end-users to configure auditable data transformations, in contrast to typical approaches using SQL or a programming language that relies on extensive and expensive communication between user, analyst, and programmer. In addition, the RexDB platform employs layers of functionality to support the secure management of ePHI data, as well as being developed using highly secure and industry tested open source packages that ensure the security at all levels of the application.

Results: Our initial implementation delivered two procedure-specific registries: interventional and diagnostic. Each of these registries accepted data from external EHRs, filtered the results, and transformed the data to integrate properly with a normalized, transactional data structure. A configuration mechanism allowed users of the system to configure new registries, including additional forms for data entry and visit-types with required procedures. Finally, an integrated data exploration tool allowed users to generate custom datasets based on registry participation, disease conditions and item-level data values from form data entry. Users could generate analyst-configured and ad hoc custom reports based on these datasets and share them with external researchers. Importantly, each registry took approximately three days to configure, including setting up the initial ETL processes from a single EHR system; each custom EDC form took approximately an hour per page; and each report took approximately an hour per logical condition. User-experience reports on the system will be available by the time of the meeting.

Discussion: Our initial findings suggest that the RexDB design philosophy of empowering the analyst can address both major difficulties blocking registry adoption. We have found that the level of configurability presented to analysts in RexDB is novel compared to other software platforms, where there are significant limitations placed on the ability to create custom data models, workflows, ETL processes, and screens. Additionally, the configuration format and tools that drive this customization has been shown to have a short learning curve with semi-technical individuals, which leads to a much shorter development time for deploying user-focused system customizations. We plan to apply this technology stack to additional registries for a wider perspective. Surprisingly, adapting the core RexDB study management suite to meet the needs of registry management and reporting required only a slight modification of a previously developed “research study management” model. To understand this further, we plan to explore the relationship between clinical registry and research data models in the future. This case study also suggests that using an open-source stack has the benefit of aligning incentives between the technology services vendor and the registry customer: Because the vendor cannot use intellectual property to lock in a customer, they are continually motivated to deliver high levels of service.

Acknowledgements: Research reported in this publication was supported by the National Institutes of Health and the National Science Foundation (Award ID: 98771, grant numbers R43MH102900 and R43MH099826).

References: [1] The IRIS™ Registry (Intelligent Research in Sight), retrieved from http://www.aao.org/iris-registry/index.cfm, [2] NCDR® (National Cardiovascular Data Registry), retrieved from https://www.ncdr.com/webncdr/

Maintenance of a Standard Pharmacy Terminology within a Controlled Medical Vocabulary Server

Michael Totzke, Jon Herbert RN, R. Dean Woolstenhulme MS, LAc
3M Health Information Systems, Murray, Utah

Introduction: The 3M Healthcare Data Dictionary (HDD) functions as a master reference terminology, which integrates and supplements multiple other standard vocabularies.[i] By mapping concepts to a central terminology, a terminology server like the HDD allows translation of data between standard terminologies, between legacy, or site-specific terminologies, and between standard and legacy terminologies. One of the several drug terminologies that the HDD incorporates is RxNorm. RxNorm is a standardized drug nomenclature developed and maintained by the National Library of Medicine (NLM) designed to help translate among multiple other standard drug terminologies. With the selection of RxNorm as the drug terminology standard required to meet Meaningful Use criteria, it has become necessary for the HDD to maintain RxNorm's drug data from a more longitudinal perspective. Our former process for maintaining RxNorm dealt solely with mapping of the current version, without regard to managing changes over time.

Methods: The concept identifier in the HDD is the Numeric Concept IDentifier (NCID) and we map the RxNorm RXCUI to our code based on concept meaning. The NCID denotes a unique concept in the HDD. The RXCUI denotes a unique concept in RxNorm, whereas the STR is the human readable description, and the RXAUI is the description numeric identifier. When an RxNorm concept is mapped to the equivalent HDD concept, the RXCUI/RXAUI/STR triplet is loaded as descriptions on the NCID. There were three basic steps at the beginning this project. To begin with, we needed to ensure that the current RxNorm triplets mapped in the HDD were in a state to be compared with the native RxNorm triplets (from the source terminology). We had to remove old RxNorm data that we hadn't maintained, and so we inactivated any RXAUI or STR value that was not associated with an active RXCUI on an NCID. This is because, of the three types of values, the RXCUI was the only code that we had been tracking on a monthly basis. Next, there were active RXCUIs mapped to NCIDs that didn't have an RXAUI and STR, or had only one or the other, that needed to be added to the HDD. Once we had triplets on every NCID where an active RXCUI was mapped, we could make our comparison. Our approach to this was to create a series of queries, as database views, in our ORACLE database, that would do all the comparison work of the two sources of data, RxNorm and the HDD.

Results: There were various types of outcomes in which the comparisons passed or failed. The first type of failure was a complete triplet missing from the HDD. The second type of failure to match was when there was an active RXCUI mapped to an NCID, but the RXAUI and the STR were inactive in the RxNorm source table. This occurs when the STR, or term, for the RxNorm concept was altered which means that the RXAUI also changes, but the RXCUI was conceptually the same. This type of change most often was the result of some type of cosmetic change, such as capitalization of a word. But there were also many of this type which resulted in a change of concept meaning. One example of this is RXCUI 80163, which changed from *Alanine 2.07 MG/ML* to *Alanine 20.7 MG/ML.* The next two types of failure were the result of either the RXAUI *or* the STR being changed, and the other not changing

Discussion: Mapping maintenance of a standard pharmacy terminology is extremely important to data quality. The impetus for changing our RxNorm mapping process came from several places, the most important of which was the need to improve the quality and consistency of the mapping. We realized that simply adding new RXCUIs when they were new, or simply removing RXCUIs when they were inactivated in RxNorm, perpetuated potentially serious data problems. Through this new process, we were able to find errors and inconsistencies in our own data, including identifying duplicate concepts (e.g. RXCUI 1309219 *poppyseed oil* and RXCUI 1432299 *poppy seed oil*), and errors in RxNorm data as well.

[i] Lau, L. M., and S. H. Lam. 1999. "Applying the desiderata for controlled medical vocabularies to drug information databases." *Proc AMIA Symp*:97-101. doi: D005742 [pii].

Semantic Role Labeling for Modeling Surgical Procedures in Operative Notes

Yan Wang, MS[1], Serguei Pakhomov, PhD[1, 2], James O. Ryan, MS[4], Genevieve B. Melton, MD, MA[1, 3]

[1]Institute for Health Informatics, [2]College of Pharmacy,[3] Department of Surgery, University of Minnesota, Minneapolis, MN; [4]Department of Computer Science, University of California Santa Cruz, Santa Cruz, CA

Introduction

The application of NLP systems to process clinical text continues to attract researchers who aim to extract valuable information from large numbers of clinical notes and to automatically provide necessary information to users in a timely and scalable fashion. In analyzing operative notes, linguistic features including the presence of extensive predication often with action verbs were previously found to be key to utilizing these notes[1]. We previously created PropBank style PAS frames for a set of top action verbs from operative notes[2], which were used as to as a framework for operative note SRL. In this study, we adapted and customized a SRL system for modeling surgical procedures in operative notes. Our adapted SRL system will be used in downstream tasks such as modeling of procedure phases, summarizing and presentation to provide important surgery information in a succinct and easily comprehensible fashion for clinicians.

Method

A Maximum Entropy (ME)[3]-based SRL model was trained on an annotated corpus collected randomly from 3,000 Laparoscopic Cholecystectomy and 6,481 notes from six other gastrointestinal surgical procedures for 20 top surgical action verbs. The Predicate Argument Structure (PAS) of the action verb in each sentence was annotated using the surgical PAS frames created previously[2]. A combination of additional features were use to augment the performance of the system, as follows. Extra syntactic phrases were also collected from syntactic output produced by MetaMap[4] for detecting noun phrases in biomedical text by using SPECIALIST minimal commitment parser[5]. The ME SRL model was trained using the Unified Medical Language System (UMLS) semantic types[6] of the head word as a new feature based on a set of heuristic rules for semantics type groupings. Finally, an adapted unlexicalized PCFG syntactic parser was used to obtain better syntactic parse trees for operative notes, which provides necessary syntactic information for SRL model training.

Result

Our SRL system along with two existing SRL systems (ASSERT[7]- a constituency-based SRL system trained on Wall Streat Journal corpus and - ClearNLP[8] - a dependency-based SRL system trained on a PropBank annotated clinical corpus including MiPACQ[9]) were evaluated. ASSERT had the lowest performance with a precision of 63%, recall of 54%, and F-measure of 58%. In comparison, ClearNLP had better performance with a precision of 69%, recall of 59%, and F-measure of 64%. Our ME SRL system trained on operative notes showed improved performance with a precision of 69%, recall of 82% and F-measure 74%. This was consistent with microaveraging and macroaveraging analyses, along with analysis by semantic role type (Tables 1 and 2). Table 2 shows the SRL system performance on different semantic role types such as Arg0 - a core argument usually specifies the agent of an action, ArgM-CAU - which indicates the reason of an action and ArgM-DIR – a semantic role shows the path of the motion. As shown in table 2, the SRL system performs very well on detecting core arguments of surgical actions, which contain the most important details about surgical actions.

Discussion

Our study demonstrates that a ME SRL system trained on annotated operative notes has improved performance over existing SRL tools. The corpus specifically annotated with domain PAS used in this study provided a robust training set for the ME engine to derive accurate model parameters. Future work will include scalable automatic acquisition of PAS frames for more predicates to obtain wider coverage of most surgical actions and to extend PAS frames to include nominal action predicates.

Acknowledgement

The authors would like to thank Fairview Health Services and grant support from the University of Minnesota Clinical and Translational Science Award 8UL1TR000114-02 and American Surgical Association Foundation Fellowship Award.

Table 1. Performance Evaluation With Different Aggregate Methods

Aggregate Method	Recall	Precision	F Measure
Microaveraging	66.7%	80.5%	73.0%
Macroaveraging	69.2%	82.1%	74.1%
Sentence labeling accuracy		50.7%	

*****Sentence labeling accuracy - the rate of correctly labeling all arguments of each sentence.

Table 2: Performance Evaluation By Semantic Role Types

Role Type	Recall	Precision	F Measure
Arg0	62.4%	88.0%	73.0%
Arg1	82.2%	98.0%	89.4%
Arg2	78.4%	96.7%	86.6%
Arg3	70.5%	95.2%	81.0%
Arg4	78.6%	84.6%	81.5%
ArgM-ADV (Adverbials)	67.4%	84.9%	75.2%
ArgM-CAU (Cause clauses)	0	0	-
ArgM-DIR (Directionals)	50.9%	77.8%	61.5%
ArgM-DIS (Discourse)	0	0	-
ArgM-EXT (Extent)	37.5%	60.0%	46.2%
ArgM-LOC (Locatives)	21.4%	50.0%	30.0%
ArgM-MNR (Manner)	46.8%	65.9%	54.7%
ArgM-MOD (Modals)	0	0	-
ArgM-NEG (Negation)	0	0	-
ArgM-PNC (Purpose)	32.0%	66.7%	43.2%
ArgM-PRD (Secondary prediction)	53.2%	75.0%	62.3%
ArgM-REC (Reciprocals)	0	0	-
ArgM-TMP (Temporal)	58.8%	90.5%	71.3%

References

1. Wang Y, Pakhomov S, Burkart NE, Ryan JO, Melton GB. A study of actions in operative notes. AMIA Annu Symp Proc. 2012;2012:1431-40.
2. Wang Y, Pakhomov S, Melton GB. Predicate Argument Structure Frames for Modeling Information in Operative Notes. Medinfo. 2013.
3. Park K-M, Rim H-C. Maximum Entropy based Semantic Role Labeling. CONLL '05 Proceedings of the Ninth Conference on Computational Natural Language Learning. 2005:209-12.
4. Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc. 2010;17:229-36.
5. McCray AT, Srinivasan S, Browne AC. Lexical methods for managing variation in biomedical terminologies. Proc Annu Symp Comput Appl Med Care. 1994:235-9.
6. Unified Medical Language System (UMLS). Available from: http://www.nlm.nih.gov/research/umls/.
7. Pradhan S, Ward W, Hacioglu K, Martin JH, editors. Shallow semantic parsing using Support Vector Machines. the Human Language Technology Conference/North American chapter of the Association for Computational Linguistics annual meeting (HLT/NAACL-2004); 2004.
8. ClearNLP. Available from: http://clearnlp.wikispaces.com.
9. Albright D, Lanfranchi A, Fredriksen A, Styler WF, Warner C, Hwang JD, et al. Towards comprehensive syntactic and semantic annotations of the clinical narrative. Journal of the American Medical Informatics Association. 2013 January 25, 2013.

Evaluation of Diagnosis Codes, Clinical Notes, and Medications on Identifying Subjects with a Specific Disease Phenotype

Wei-Qi Wei[1] MD PhD, Pedro L. Teixeira[1] BS, Huan Mo[1] MD, Robert M. Cronin[1, 2] MD, Jeremy Warner[1, 2] MD, Joshua C. Denny[1, 2] MD MS

Departments of [1]Biomedical Informatics and [2]Medicine, Vanderbilt University, Nashville, TN

Introduction: The process of finding accurate phenotypes (e.g., congestive heart failure) from heavy volumes of imperfect practice-based electronic health record (EHR) data is a crucial enabling step for clinical and genetic research using EHRs. Billing codes often have been used in traditional phenotyping since most patients with the disease should be assigned a relevant code for the billing purpose. However, due to their inaccuracy or incompleteness, using billing codes alone may result in low specificity or sensitivity. Work from collaborative phenotyping groups such as the Electronic Medical Records and Genomics (eMERGE) Network suggest that a combination of diagnosis codes with other EHR components, e.g. medication and clinical notes, shows an improved performance on multiple disease phenotypes. In this study, we tested a logical hypothesis that algorithmic combinations data from multiple components of EHR may improve the accuracy, and possibly recall. We evaluate the phenotyping performance of three major components of EHR – billing codes, keywords in clinical notes, and specific medications -- on a broad spectrum of pre-selected diseases.

Methods: This evaluation was conducted by using de-identified EHR data at Vanderbilt. ICD9 codes were retrieved from administrative claims data and aggregated with phenome-wide association codes (PheWAS) codes. We defined primary notes as problem lists, discharge summaries, or history and physical notes (H&P). Keyword searches and simple rules were performed on primary notes to determinate if a patient has a positive mention of the disease. The list of specific medications was obtained through MEDI, is a freely-available, computable medication-indication resource. Medication data in our EHRs are often embedded in clinical narratives and were obtained through employing MedEx in addition to electronic prescribing data. Ten diseases were selected for this evaluation study: atrial fibrillation (AFIB), Alzheimer's disease, breast cancer, gout, and human immunodeficiency virus (HIV) infection, multiple sclerosis (MS), Parkinson's disease, rheumatoid arthritis (RA), type 1 diabetes (T1D), and type 2 diabetes (T2D). For each disease, subjects were classified into one of the seven categories based on the evidence presented in their EHRs, i.e. 1. ICD only, 2. primary notes (PN) only, 3. medications only, 4. ICD and PN, 5. ICD and medications, 6. PN and medications, and 7. ICD, PN, and medications. A group of 25 patients per disease category (a total number of 175 patients for each disease, 1750 patients for the ten diseases) were randomly selected for physician chart review. Review results were used to estimate precision, recall, and F-score of each component. Kappa scores between reviewers were between 0.68-0.90.

Results: The precisions for single components alone were inconsistent and inadequate for accurately phenotyping (0.06-0.71). Limiting to patients with ≥2 diagnosis codes improved the average precision to 0.84. We observed a more stable and higher accuracy when using ≥2 components (mean±standard deviation: 0.91±0.08). Primary notes offered the best recall (0.77). The recall of diagnosis codes was 0.67. Again, two or more components provided a reasonably high and stable recall (0.59±0.16). Overall, the best performance (F score: 0.70±0.12) was achieved by using ≥2 components. Although the overall performance of using diagnosis codes (0.67±0.14) was only a slightly lower than using ≥2 components, its precision (0.71±0.13) is substantially worse than using ≥2 components (0.91±0.08).

Discussion and Conclusion: Our results show that use of multiple modalities of EHR data derived from available computable resources improved upon use of a single resource for precision and F-score. Using multiple ICD9 codes substantially improved ICD code precision to 0.84±0.12. Medications alone were not sufficient to diagnose disease, even for highly specific meds such as HIV drugs. We suggest that multiple EHR components should be considered for future phenotyping efforts. These methods may extend to many other diseases. More work is needed to eliminate false positives for some diseases, especially in PN.

Table 1. Precisions for various categories based on chart review results. The "only" categories refer to that category without other mentions; e.g., "ICD only" refers to patients with ICD codes without a Med or PN mention; "ICD" refers to patients with any ICD (they may also have Med or a PN mention. PN=Primary Notes; SD=Standard Deviation

Disease	ICD Only	PN Only	Meds Only	ICD+Meds	ICD+ PN	Meds+ PN	ICD+both	ICD	Meds	PN	≥2 ICDs	≥2 Components
AFIB	0.52	0.72	0.08	0.72	1.00	1.00	1.00	0.72	0.35	0.96	0.88	0.84
Alzheimer's	0.28	0.20	0.00	0.80	0.88	0.92	0.88	0.69	0.40	0.32	0.74	0.88
breast CA	0.12	0.72	0.04	0.88	0.96	1.00	1.00	0.45	0.81	0.84	1.00	0.97
Gout	0.56	0.84	0.00	0.92	1.00	1.00	1.00	0.81	0.69	0.91	0.93	1.00
HIV	0.52	0.00	0.00	0.92	0.84	0.88	1.00	0.81	0.69	0.20	0.89	0.95
MS	0.20	0.08	0.12	0.88	0.88	0.88	1.00	0.78	0.93	0.41	0.86	0.94
Parkinson	0.48	0.16	0.04	0.84	1.00	0.88	0.96	0.89	0.87	0.33	0.94	0.98
RA	0.36	0.20	0.00	0.64	0.76	0.88	0.84	0.68	0.73	0.27	0.77	0.78
T1D	0.28	0.12	0.04	0.16	0.92	0.84	0.76	0.59	0.49	0.45	0.62	0.91
T2D	0.36	0.68	0.24	0.60	0.80	1.00	0.84	0.65	0.65	0.80	0.73	0.81
Mean	***0.37***	***0.37***	***0.06***	***0.74***	***0.90***	***0.93***	***0.93***	***0.71***	***0.66***	***0.55***	***0.84***	***0.91***
SD	***0.15***	***0.32***	***0.08***	***0.23***	***0.09***	***0.06***	***0.09***	***0.13***	***0.20***	***0.29***	***0.12***	***0.08***

Table 2. Recalls for various categories based on chart review results.

Disease	ICD Only	PN Only	Meds Only	ICD+Meds	ICD+ PN	Meds+ PN	ICD+both	ICD	Meds	PN	≥2 ICDs	≥2 Components
AFIB	0.24	0.03	0.09	0.31	0.08	0.02	0.23	0.85	0.65	0.36	0.64	0.63
Alzheimer's	0.05	0.47	0.00	0.04	0.11	0.15	0.18	0.38	0.38	0.91	0.24	0.49
breast CA	0.09	0.42	0.00	0.02	0.29	0.05	0.14	0.53	0.21	0.89	0.55	0.49
Gout	0.18	0.40	0.00	0.01	0.37	0.01	0.03	0.58	0.05	0.82	0.34	0.42
HIV	0.21	0.00	0.00	0.40	0.03	0.04	0.33	0.96	0.76	0.39	0.82	0.79
MS	0.05	0.11	0.00	0.03	0.32	0.03	0.46	0.85	0.52	0.92	0.63	0.83
Parkinson	0.06	0.36	0.00	0.01	0.36	0.03	0.19	0.61	0.23	0.93	0.42	0.58
RA	0.05	0.60	0.00	0.00	0.25	0.01	0.08	0.38	0.10	0.95	0.31	0.35
T1D	0.21	0.12	0.00	0.00	0.65	0.00	0.01	0.88	0.02	0.79	0.60	0.67
T2D	0.10	0.21	0.06	0.10	0.12	0.12	0.29	0.61	0.56	0.74	0.42	0.63
Mean	***0.12***	***0.27***	***0.02***	***0.09***	***0.26***	***0.05***	***0.19***	***0.67***	***0.35***	***0.77***	***0.50***	***0.59***
SD	***0.08***	***0.20***	***0.03***	***0.14***	***0.19***	***0.05***	***0.14***	***0.21***	***0.27***	***0.22***	***0.18***	***0.16***

Table 3. F-scores for various categories based on chart review results.

Disease	ICD Only	PN Only	Meds Only	ICD+Meds	ICD+ PN	Meds+ PN	ICD+both	ICD	Meds	PN	≥2 ICDs	≥2 Components
AFIB	0.33	0.07	0.09	0.43	0.14	0.04	0.37	0.78	0.45	0.53	0.74	0.72
Alzheimer's	0.08	0.28	0.00	0.08	0.20	0.26	0.30	0.49	0.39	0.47	0.36	0.63
breast CA	0.10	0.53	0.00	0.05	0.44	0.09	0.24	0.49	0.33	0.86	0.71	0.65
Gout	0.27	0.54	0.00	0.01	0.54	0.03	0.05	0.68	0.09	0.86	0.49	0.59
HIV	0.29	0.00	0.00	0.56	0.06	0.07	0.49	0.88	0.72	0.27	0.85	0.86
MS	0.08	0.09	0.01	0.05	0.47	0.06	0.63	0.81	0.67	0.56	0.73	0.89
Parkinson	0.11	0.22	0.00	0.01	0.52	0.06	0.32	0.73	0.36	0.49	0.58	0.73
RA	0.08	0.30	0.00	0.01	0.38	0.03	0.15	0.49	0.17	0.43	0.44	0.48
T1D	0.24	0.12	0.00	0.00	0.76	0.00	0.03	0.71	0.03	0.58	0.61	0.77
T2D	0.16	0.32	0.09	0.17	0.21	0.21	0.43	0.63	0.60	0.77	0.53	0.70
Mean	***0.17***	***0.25***	***0.02***	***0.14***	***0.37***	***0.08***	***0.30***	***0.67***	***0.38***	***0.58***	***0.60***	***0.70***
SD	***0.10***	***0.19***	***0.04***	***0.20***	***0.22***	***0.09***	***0.19***	***0.14***	***0.24***	***0.19***	***0.15***	***0.12***

CONSIDERATIONS of DUAL PROCESS THEORIES for EHR DESIGN

Charlene Weir, PhD[1,3] Bryan Gibson, PhD[1], Alan Morris, MD[3], Jorie Butler,PhD[2], PhD,[1] Matt Samore[1] and Jonathan Nebeker, MD[1,3]

[1]IDEAS Center of Innovation, SLC VA, [2] GRECC, SLC VA [3]Dept. of Biomedical Informatics, University of Utah,

Introduction

In order to maximize the promise of EHRs, increasing attention is being focused on the redesign of the electronic health record.(1-3) One major goal is to enhance the cognitive support of clinicians and workflow in order to improve the quality of care, enhance safety, and expand the usefulness of the EHR.

Dual Process theories.(4, 5) posit two memory systems: 1) a network of associations that function through spreading activation to support rapid pattern-matching (System 1); and 2) a slow, rule-based, conscious system that functions through active reasoning (System 2). Both systems are continuous and simultaneous processes with humans generally motivated to preserve cognitive resources by avoiding System 2 processing, unless necessary.(4, 6, 7) Our goal is to expand the discussion of EHR design by reviewing the empirical psychological literature to incorporate a dual process a "matching hypothesis" for the design of decision support. We propose that maximizing rapid automatic processing is only the first step and that the next two priorities are to support higher levels of reasoning and the self-regulation of attention. Each issue is discussed below.

Support for Adaptive Control of Information Environment. As noted above, System 1 and System 2 processing occurs simultaneously and continuously. Managing attention resources in relation to the uncertainty of the information space in order to reach the epistemological goals of accuracy and speed is complex when there are multiple clinical priorities. Literature on adaptive control will be reviewed.(8)

Rich Contextual Displays. Expert decision-making uses pattern-matching mechanisms characteristics of the powerful associative memory system. However, if the information-space is impoverished or displays too rigid, clinicians engage in active search processes, which take substantial effort. Literature related to "feeling of knowing" integrated displays, and situational mental models will be presented.(9, 10)

Goal-based processing. Human information processing is contextual, goal-based, and linked to action. Simulation tools are one aspect of goal-based processing that is neglected in EHR design, despite the importance of simulation in human thought. Literature on simulation will be presented and linked to issues of clinical decision-making.(11-13)

Planning and Action (Decision-Making). Care planning software should support both planning and action. Research in decision-making have identified that information is handled differently at different times in the decision-making sequence. Literature on action support will be reviewed.(14, 15)

Support for Establishing Common Ground Across the Team. Common ground refers to the shared understanding between co-workers regarding their knowledge of the goals, intentions and responsibilities of each other. Many clinical work-arounds are instantiations of processes to establish common ground (16). Literature related to dual process impacts on establishing common ground will be reviewed in order to show how to improve clinical coordination.(17)

REFERENCES

1. Chaudhry B, Wang J, Wu S, Maglione M, Mojica W, Roth E, et al. Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. Ann Intern Med. 2006;144(10):742-52.

2. Gans D, Kralewski J, Hammons T, Dowd B. Medical groups' adoption of electronic health records and information systems. Health Affairs. (September/October 2005);24(5):1323-33.

3. Middleton B, Bloomrosen M, Dente M, Hash B, Koppel R, Overhage J, et al. Enhancing patient safety and quality of care by improving the usability of electronic health record systems: recommendations from AMIA. JAMIA. 2013.

4. Smith ER, DeCoster J. Dual-Process Models in Social and Cognitive Psychology: Conceptual Integration and Links to Underlying Memory Systems. Personality and Social Psychology Review. 2000;4(2):108-31.

5. Croskerry P. Clinical cognition and diagnostic error: applications of a dual process model of reasoning. Advances in health sciences education : theory and practice. 2009;14 Suppl 1:27-35.

6. Evans J, Frankish K, (Eds) In Two Minds: Dual Processes and Beyond Oxford Press; 2009.

7. Evans JS. Dual-processing accounts of reasoning, judgment, and social cognition. Annu Rev Psychol. 2008;59:255-78.

8. Gigerenzer G. Adaptive Thinking. Oxford University Press. 2000.

9. Wyer R. Social Comprehension and Judgement: The Role of Situation Models, Narratives and Implicit Theories. Lawrence Erlbaum Associates: London. 2004:s.

10. Kruglanski A, Gigerenzer G. Intuitive and Deliberate Judgments Are Based on Common Principles. Psychological Review. 2011;118(1):97-109.

11. Baumeister R, Masicampo E. Conscious Thought is for Facilitating Social and Cultural Interactions: How Mental Simlations Serve the Animal-Cultural Interface. Psychol Rev. 2010;117(3):945-97.

12. Kahol K, Vankipuram M, Smith ML. Cognitive simulators for medical education and training. Journal of biomedical informatics. 2009;42(4):593-604.

13. Schacter D, Addis D, Buckner R. Episodic Simulation of Future Events. Ann NY Acad Sci. 2008(1124):39-60.

14. Gollwitzer P. Action phases and mind-sets. In: Higgins ET, Sorrentino R, (Eds). Handbook of motivation and cognition: Foundations of social behavior (Vol 2). 2. New York: Guilford Press; 1990.

15. Gollwitzer P, Sheeran P. Implementation Intentions and Goal Achievement: A Meta-analysis of Effects and Processes Advances in Experimental Social Psychology. 2006;38:69-119.

16. Croskerry P. Clinical cognition and diagnostic error: applications of a dual process model of reasoning. Advances in health sciences education : theory and practice. 2009;14 Suppl 1:27-35.

17. Keysar B, Barr DJ. Approaches to Studying World Situated Language Use: Bridging the Language-as-Product and Language-as-Action. Cambridge, MA: MIT Press. In J C Trueswell & M K Tanenhaus (Eds), Approaches to Studying World Situated Language Use: Bridging the Language-as-Product and Language-as-Action. Cambridge, MA: MIT Press; 2005.

A Framework for Analyzing Inpatient Nursing Costs

John M. Welton, PhD, RN[1]
[1]University of Colorado College of Nursing, Aurora, CO

Introduction

Traditional methods of measuring nursing time (intensity) and costs do not address patient level nursing care. Most hospital cost accounting procedures use department or cost center averages known as nursing hours and nursing costs per patient day (NHPPD, NCPPD). This method is problematic for several reasons, first, there is no way to allocate nursing time or costs to individual patients. Second, the use of NHPPD/NCPPD do not measure individual variances in nursing care for patients within a medical diagnosis or Diagnosis Related Group (DRG). Essentially all patients are allocated the same "average" nursing care for each day of stay within the cost center. Because hospitals are reimbursed based on these diagnoses, there is no link between how nursing resources are expended for different patients, their related costs, the billing for inpatient care and payment to the hospital. Third, there is a compelling need to better identify cost drivers within the healthcare system and registered nurses make up over 25% of hospital expenditures[1]. Lastly, there is an emergent need to devise new methods to extract nursing related data from existing electronic health records (EHR) and combine these data with other operational, clinical, financial, and quality data systems to dovetail with other emerging operational informatics techniques and develop nursing business intelligence and analytic tools.

Methods

The University of Colorado College of Nursing formed a consortium of hospitals and healthcare systems in 2013 named the Colorado Collaborative for Nursing Research (CCNR). This brings together nurses and other clinicians and leaders from various healthcare settings to devise methods to share data to address common problems. One primary deliverable is to devise a framework to identify nursing direct care time and costs and allocate these to each individual patient to overcome the issues identified above. The value of this work is the ability to link several key measures of nursing care to each individual patient: the actual hours of care provided; the cost of that care using either a standard cost or the actual wage of the nurse and aides; the overall assignment of nurses to patients, and the distribution of the assignment to the scheduled nursing staff and aides for a particular shift.[2]

The cost model is based on original work conducted by the Yale team that developed the DRG to separate nursing costs into a nursing cost center based on patient level nursing intensity,[3] however this feature of the DRG was never implemented. More recent work by Kaplan and Porter out of the Harvard School of Business present a framework for patient level costing models in healthcare.[4] We draw from these collective work to guide the development of a patient level nurse costing model.

In two pilot studies conducted at separate intuitions, data were extracted from several existing EHR components: 1. The Clairvia (Cerner Corp) nurse staffing and scheduling software; 2. Existing human resources data to identify nurse characteristics and wages; 3. Discharge summaries to identify patient characteristics and diagnoses; and 4. Admission/Discharge/Transfer (ADT) information. Using Thompson's approach, the ability to identify the actual hours of care is based on the assignment of each nurse to a patient from which a cost using actual or standard wage can be applied. This method is more robust and explains a higher portion of variance of actual nursing care hours delivered that other methods such as nurse to patient ratios.[2]

Results

The key nurse and patient specific variables are detailed in Table 1. A few key metrics derived from these data are shown in Table 2 and include both nursing time and cost as well as other useful metrics about nurses and the care they deliver to patients. These new data derived from the nursing assignments can provide a rich analytic environment as well as near real-time data about internal operations and nursing resource management. In one of the pilot studies, nursing costs per patient day were found to vary between \$132 - \$1,455.[5] In the second pilot, 20% of the patients consumed 50.4% of all nursing hours and costs.[6] Also, for adult patients on the medical/surgical floors, nursing hours and costs per day declined by hospital length of stay from 6.7/\$334 for 1 day of stay, 5.2/\$260 for 7 days, 4.9/\$\$244 at 14 days, and 4.8/\$242 at 21 days.

Further development of both the data collection methods as well as analytics is anticipated. This includes adding nursing quality and performance metrics. For example, future models will be able to measure the demand for nursing

care and the actual care (time and costs) delivered using nursing terminologies in the Nursing Outcomes Classification tool mapped to nursing assessment data.[7] Metrics will identify varying patterns of care, for example more care than is needed is inefficient and extra time and costs can be calculated to aid resource decision making to provide the most efficient level of staffing based on patient needs rather than nurse to patient ratios. Also, nursing care time and costs will be benchmarked across different care settings aligned with key outcome goals such as length of stay and quality measures tied to value-based purchasing.

Table 1. Core Data from Nursing Data Extraction

Entity	Label	Comment
Nurse Characteristics	StaffID (PK); Role (RN, LPN, Aide); Unit experience; Wage; Hospital experience; Nursing experience; Assigned unit; Work category (FT/PT, per diem, traveler);	Main table for nurse characteristics
Patient Characteristics	PatientID (PK) Age, Sex, MSDRG, Hospital length of stay,	Main table for patient characteristics
Nurse-Patient Assignment	StaffID; PatientID; UnitID; Shift ID; Patient acuity; Demand nursing hours required; Nursing hours delivered; Nursing costs	Association table for nurse:patient assignment Nursing costs = time * wage

Note, a full list of all variables can be found at: http://links.lww.com/JONA/A303

Table 2. Nursing Care Metrics and Analytics

Patient Level	Unit Level	Hospital Level
• Direct nursing care hours and costs per day and per case. • Nursing acuity level (demand for nursing care) • Sum of nursing direct care time and costs per case • Mean experience level of provided nursing care.	• Mean assignment (# patients per nurse); direct care hours and costs all nurses and aides • Mean nursing experience level and sum and percent of hours by nursing degree (e.g. BSN) • Sum and percent of hours and costs for float or per diem.	• Mean direct nursing care time and costs per DRG/Diagnosis • Mean nursing hours and costs by day of stay. • Nursing time and cost variability over time and comparison of demand vs. actual care hours delivered

Conclusion

A method to extract nursing direct care time and costs from the nurse to patient assignment is proposed. Initial pilot study results confirm both the suitability of these methods to measure patient level nursing time and costs and extract data from existing EHRs. This method is vendor agnostic and can provide a more robust data environment to examine nursing care at the individual nurse-patient level of analysis. Further discussion is needed to identify ways to widely disseminate and test this model in a variety of patient care settings to identify and benchmark nursing care.

References

1. Welton JM. Hospital nursing workforce costs, wages, occupational mix, and resource utilization. *J Nurs Adm.* 2011;41(7-8):309-314.
2. Welton JM, Zone-Smith L, Bandyopadhyay D. Estimating nursing intensity and direct cost using the nurse-patient assignment. *J Nurs Admin.* 2009;39(6):276-284.
3. Thompson JD, Averill RF, Fetter RB. Planning, budgeting, and controlling--one look at the future: case-mix cost accounting. *Health Services Research.* 1979 Summer 1979;14(2):111-125.
4. Kaplan RS, Porter ME. How to solve the cost crisis in health care. *Harv Bus Rev.* 2011;89(9):3-18.
5. Jenkins P, Welton J. Measuring direct nursing cost per patient in the acute care setting. *J Nurs Adm.* 2014;44(5):257-262.
6. Welton JM, Caspers B, Sanford K. Inpatient Nursing Hours and Cost Outcomes Within a Health Care System. *American Organization of Nurse Executives 46th Annual Conference*; March 20-22, 2013; Denver, CO.
7. Caspers BA, Pickard B. Value-based resource management: a model for best value nursing care. *Nurs Adm Q.* 2013;37(2):95-104.

Exploring Bayesian Network Development Using Unsupervised Machine Learning for Patient Risk Assessment

Bruce Wilson, MBA[1], Christine Tsien Silvers, MD, PhD[1,2], Cindy Crump, MA[1], Loretta Schlachta-Fairchild, RN, PhD, FACHE[3], COL Jeffrey S. Ashley, AN, PhD[4]
[1]AFrame Digital, Inc., Reston, VA; [2]Children's Hospital Informatics Program, Boston, MA; [3]iTelehealth, Inc., Cocoa Beach, FL; [4]Center for Nursing Science and Clinical Inquiry, Walter Reed National Military Medical Center, Bethesda, MD

Introduction. Unsupervised machine learning methods were used to explore predictive Bayesian networks[1] in a research program to develop a mobile real-time patient risk management tool. Scoring systems or models to predict mortality have been described but typically only use data from the first 24 hours in an intensive care unit (ICU).[2-4] This work explores ICU outcome prediction using data from both the first 24 hours and beyond, as a stepping stone toward ultimately predicting need for early intervention in those with chronic diseases, which accounted for 84% of all health care spending in 2006 and cost $315.4 billion for just heart disease and stroke in 2010.[5]

Methods. Time-series data for 29 patient measurements—including physiological vital sign, laboratory values, demographic data, work history, and outcomes—were retrospectively collected for each patient in an IRB-approved protocol. Patients had been admitted for burn, infection, hypovolemia, and traumatic brain injury during a five-year period ending in October 2007. Bayesian networks were developed using a constraint-based unsupervised machine learning development package (FasterAnalytics, DecisionQ Corp.). Prior to training networks, the time-series data were pre-processed by (1) removal of invalid values, (2) transformation of each patient time series into a single unified dataset, and (3) definition of temporal "observation windows." Cross-validation and area under the receiver operating characteristic (ROC) curve (AUC) analyses were used to identify which set of time windows and data transforms produced the most predictive model. Bayesian networks were validated using a leave-one-out cross-validation methodology. Each test set's predictions were then used to calculate an ROC curve by feature of interest.

Results. Data were collected from 650 patients throughout their ICU stays. Models using hospital admission data alone showed that all patients appeared to have an equal chance for a positive outcome; models developed from chosen time frames prior to the known outcome, however, showed that electronically-collectible patient parameters, such as glucose, respiratory rate, and blood pressure, can be used to differentiate risk and estimate patient outcomes. Cross-validation analysis yielded a mean AUC of 80.4% for predicting outcome in this preliminary work.

Discussion: The initial results are promising; however, the modeling work has identified some limitations and additional requirements. First, the current models required substantial optimization in order to arrive at the best results. Additional research will involve testing additional reference time windows, data transforms especially temporal abstractions, model pruning techniques, and further experimentation with feature selection. Identification of decision thresholds that take prevalence into account and optimization of predictive values should also be explored. Second, clinical validation is necessary through review of clinical practice, along with a prospective, randomized, controlled clinical trial of the best candidate prediction models. While the study datasets are from an ICU, Bayesian network models have the potential to shed light on which easy-to-collect measurements might be highly correlated to identifying changing patient risk stratification using lightweight mobile monitoring tools.

Acknowledgements: The authors thank the ICU staff and patients, Phil Kalina and John Eberhardt (DecisionQ), and DARPA (W31P4Q-06-C-0016). The views expressed herein are those of the authors and do not reflect the official policy or position of Brooke Army Medical Center, the U.S. Army Medical Department, the U.S. Army Office of the Surgeon General, the Department of the Army, the Department of Defense, or the U.S. Government.

References

1. Pearl J, Russell S. Bayesian networks. In Arbib MA. Handbook of Brain Theory and Neural Networks. Cambridge, MA: MIT Press; 2002:157-160.
2. de Jongh MA, Verhofstad MH, Leenen LP. Accuracy of different survival prediction models in a trauma population. Br J Surg. 2010 Dec;97(12):1805-13. Epub 2010 Aug 19.
3. Millham FH, LaMorte WW. Factors associated with mortality in trauma. J Trauma. 2004;56:1090-1096.
4. Vassar MJ, et al. Prediction of outcome in intensive care unit trauma patients. J Trauma. 1999;47(2):324-9.
5. (2014) Retrieved July 17, 2014, from http://www.cdc.gov/chronicdisease/overview/index.htm.

Evaluating Living Situation, Occupation, and Hobby/Activity Information in the Electronic Health Record

Tamara J. Winden, MBA[1,5], Elizabeth S. Chen, PhD[3,4], Elizabeth Lindemann[1], Yan Wang, MS[1], Elizabeth W. Carter, MS[3], Genevieve B. Melton, MD, MA[1,2]
[1]Institute for Health Informatics, [2]Surgery, Univ. of Minnesota, Minneapolis, MN,
[3]Ctr for Clinical & Translational Science, [4]Medicine, Univ. of Vermont, Burlington, VT
[5]Division of Applied Research, Allina Health, Minneapolis, MN

Introduction: Social and individual behavioral factors play an important role in diagnosis, prevention, health outcomes, and quality of life[1,2]. As defined by the World Health Organization, "social determinants of health are the conditions in which people are born, grow, live, work and age"[3]. Living situation information, such as residence type, with whom the patient lives, housing density, physical living conditions, and social support, all have been shown to have significant impact on a patient's health outcomes[4,5]. In addition, occupation and hobbies/activities can adversely impact health[6,7]. Within electronic health record (EHR) systems, social determinant documentation may be entered as structured data or unstructured text (e.g., in clinical notes or free-text data entry fields). Building upon previous work[8], the goal of this study was to evaluate the documentation of living situation, occupation, and hobbies through analysis of specific ancillary notes as well as other structured and unstructured sections of the EHR.

Methods: Progress and consult notes from a publicly accessible data source, MTSamples.com (MTS), and the University of Pittsburgh Medical Center (UPMC) NLP Repository were annotated to extract general social history sections and sentences. Each sentence was then annotated using an annotation schema developed specifically for living situation, occupation, and hobbies/activities (Table 1). These "training" data served to validate the annotation schema, which was then used to annotate inpatient social work notes from the Epic® Systems Corporation EHR in 2013 at Fairview Health Systems (FHS) which included patients who consented to have their medical records used for research. Specifically, 100 social work inpatient notes were randomly extracted from the EHR and annotated to classify living situation, occupation, and hobbies/activities information. Lastly, the EHR was queried for structured occupation information documented within the social history module for all inpatients for 2013.

Results: A total of 691 progress and consult notes were evaluated from MTS (n=491) and UPMC (n=200). A total of 558 social history annotations were then further classified, by consensus of two reviewers, into 10 topic areas covering living situation, occupation, and hobbies/activities as well as related exposures. This analysis identified 546 unique sentences related to the topic areas. Of the total notes evaluated 10.4% had at least one sentence related to residence, 20.2% had references to living density or cohabitation, and approximately 24% had sentences related to occupation (Table 2). Evaluation of the 100 social worker notes showed 40.0% had at least one sentence related to residence, 39.0% had social support, and 34.0% had living density. Lastly, the structured field for occupation from the FHS EHR was found to be populated in only 0.03% of inpatients in 2013.

Discussion: Our evaluation of the MTS and UPMC progress notes demonstrates that the topic areas of living situation, occupation, and hobby/activities are being documented in the EHR with occupation-related sentences found in 24% and living density in 20.2% of notes. Exposure-related sentences are found much less frequently. The social work notes showed higher levels of documentation especially related to Social Support. Evaluation of additional notes and note types including physical and occupational therapy as well as other structured, semi-structured, and unstructured sections of the chart is in progress.

Conclusion: Social determinants are important considerations in the provision of care and are also becoming more important as in terms of population health management. The contributions of this work represent a first step towards further informing biomedical standards for the representation of social determinants in the EHR, as well as the design of clinical documentation tools.

Acknowledgments: This work was supported by the National Library of Medicine of the National Institutes of Health (R01LM011364) and University of Minnesota Clinical and Translational Science Award 8UL1TR000114-02. The content is solely the responsibility of the authors and does not represent the official views of the National Institutes of Health.

Table 1: Annotation Schema for Living Situation, Occupation, and Hobbies/Activities

Topic	*Definition (Example Sentence)*
Residence	Dwelling types, physical residence. (*"The patient lives at a skilled nursing facility."*)
Living Conditions	Sanitation, safety. (*"The patient was living in a home which was reported to us to be without any heat or light and a good deal of excrement and urine in the building.")*
Living Density	With whom the patient lives. (*"The patient lives with her daughter.")*
Social Support	Assistance in support of care or activities of daily living. (*"The patient lives with her family being helped by the family and friends.")*
Living Situation Exposure	Exposures related to living situation. (*"He lives in a home where there are smokers."*)
Animals	Presence of animals in residence, domesticated animals as well as livestock. (*"At this time, there is also exposure to indoor cats and dogs."*)
Occupation	A regular activity performed for payment, career, profession, vocation, employment, or non-paid vocations such as full-time student, stay-at-home mom, retired, unemployed. (*"He was an IT software developer, but he has been out of work."*)
Occupation Exposure	Exposures related to occupation. (*"This is a 91-year-old male with a previous history of working in the coalmine and significant exposure to silica…"*)
Hobby / Activity	A hobby or activity, paid or non-paid, engaged in as a means of passing time; an avocation. (*"Hobbies: Computers, hiking, camping, fishing."*)
Hobby / Activity Exposure	Exposures related to hobbies or activities. (*"He likes to refinish old furniture so he works extensively with lead paint."*)

Table 2: Percent of total notes and number of sentences. [Percent of total unique notes (# individual sentences)]

	MTS	*UPMC*	*TOTAL "Training"*	*FHS*
Residence	6.3% (35)	20.5% (48)	10.4% (83)	40.0% (54)
Living Conditions	0% (0)	1.5% (6)	0.4% (6)	0% (0)
Living Density	17.1% (85)	28.0% (58)	20.2% (143)	34.0% (34)
Social Support	2.2% (12)	5.5% (14)	3.2% (26)	39.0%(39)
Living Situation Exposure	3.3% (17)	1.5% (3)	2.8% (20)	0% (0)
Animals	3.1% (15)	1.5% (4)	2.6% (19)	1.0% (1)
Occupation	26.5% (161)	18.0% (40)	24.0% (201)	27% (27)
Occupation Exposure	0.6% (3)	1.0% (2)	0.7% (5)	0% (0)
Hobby / Activity	6.9% (41)	1.0% (2)	5.2% (43)	15% (15)
Hobby / Activity Exposure	0% (0)	0% (0)	0% (0)	0% (0)
Total Sentences	369	177	546	113

References

1. HealthyPeople.Gov: U.S. Department of Health & Human services; 2012 [cited 2014 1/7/2014]. Available from: http://www.healthypeople.gov/2020/about/DOHAbout.aspx#socialfactors.
2. National Research Council. Genes, Behavior, and the Social Environment: Moving Beyond the Nature/Nurture Debate. . Washington DC: National Research Council, 2006.
3. Social Determinants of Health: World Health Organization; 2014 [cited 2014 1/7/2014]. Available from: http://www.who.int/social_determinants/sdh_definition/en/index.html.
4. Thomson H, Thomas S, Sellstrom E, Petticrew M. Housing improvements for health and associated socio-economic outcomes. Cochrane Database Syst Rev. 2013;2:CD008657. doi: 10.1002/14651858.CD008657.pub2. PubMed PMID: 23450585.
5. Shaw M. Housing and public health. Annu Rev Public Health. 2004;25:397-418. doi: 10.1146/annurev.publhealth.25.101802.123036. PubMed PMID: 15015927.
6. Ho LA, Kuschner WG. Respiratory health in home and leisure pursuits. Clin Chest Med. 2012;33(4):715-29. doi: 10.1016/j.ccm.2012.08.001. PubMed PMID: 23153611.
7. Smith TD, DeJoy DM. Occupational injury in America: An analysis of risk factors using data from the General Social Survey (GSS). J Safety Res. 2012;43(1):67-74. doi: 10.1016/j.jsr.2011.12.002. PubMed PMID: 22385742.
8. Chen ES, Manaktala S, Sarkar IN, Melton GB. A multi-site content analysis of social history information in clinical notes. AMIA Annu Symp Proc. 2011;2011:227-36. PubMed PMID: 22195074; PubMed Central PMCID: PMCPMC3243209.

Automated Identification of Unsuspected Lung Nodule Findings in Radiology Reports with Natural Language Processing and Text Classification

Ryan Wise MSc, Jonathan Duckart MS, Jianji Yang PhD
Portland VA Medical Center, Portland, Oregon, USA

Abstract

Detection of lung nodules on radiology images is the critical first step in timely follow-up and cancer diagnosis. Detection is hindered by the difficulty of identifying findings buried in radiology reports. An automated text classification algorithm is developed to detect these cases. The algorithm achieves a 94% recall in recognizing lung nodules. Future work is needed to integrate the algorithm into the workflow of the care coordinator for patient follow-up.

Introduction

Lung nodules on radiographic images are important indicators for potential malignancies and can lead to early detection and treatment of lung cancer. The Portland VA Medical Center (PVAMC) implemented an Unsuspected Radiologic Findings (URFs) Pathway in 1996 to ensure the timely communication, evaluation and diagnosis of lung nodules[1]. In 2013, a web-based lung-nodule-tracking registry was implemented to facilitate population-level management of URFs. In order to expand the URF program to facilities where no URF diagnostic code is used, an automated algorithm is needed to identify URF cases from free-text reports. In addition, an algorithm can further improve PVAMC's detection of URFs by identifying cases with missing URF codes.

Past work on automated identification of lung nodules in radiology reports include keyword-based[2,3] and rule-based[4] natural language processing (NLP) systems. In this work, however, we addressed the problem using text classification with machine learning algorithms[5,6]. The machine-learning approach allows greater flexibility in training the system over time to be sensitive to inter-facility differences in report and dictation styles, helping the system to achieve a degree of portability not possible in hard-coded rules and keywords.

Methods

The URFs Pathway program at the PVAMC has created an annotated corpus of reports, consisting of image studies where any part of the chest can be seen. Roughly 5% of those are coded to reflect a positive URF. We used this corpus and the Scikit-learn module[7] to build a bag-of-words model to classify image reports as either positively or negatively indicating URFs. Optimizing recall of positive URF findings was our criterion for model selection.

Benchmarking to identify the best algorithms. A random selection of 10,000 cases from the corpus were used to train and test 11 different classifiers, and the results were compared to choose models for further tuning.

Final Model. The linear-kernel support vector machine (SVM) was selected for further optimization. The development and testing process pipeline are depicted in Figure 1. The development phase used 10,000 cases divided into 6,500 for 3-fold cross validation and 3,500 for testing. Preprocessing steps included regular-expression matching to collapse related concepts into single features in order to achieve greater statistical efficiency (Table 1). For example, measurements in the text notes, such as "3.5mm by 2.01mm" or "sub-1mm," would each be replaced by the single "MEASURE" feature. Two feature-selection models (a linear SVM with L1 penalty and a chi-squared feature selector) were used to obtain the most discriminative features for the classifiers. An inverted class-rate sampling technique was used to balance the two classes. The grid search program in Scikit-learn[7] was used to choose the regularization parameter that optimized recall. Finally, an SVM with an L2 penalty was trained and evaluated using a different set of 50,000 cases, with 32,500 used in training and 17,500 in testing.

Results

Table 2 shows the top six performers from the 11 benchmarking algorithms. The SVM with L1 penalty had the highest recall and F1 scores for URF-Positive cases. The performance for the final model is presented in Table 3. There was a 15% improvement in URF-Positive recall from the base model, accompanied by a drop in precision.

Discussion

While the performance of the final model is promising, further improvement on the model can be made in several areas, such as including relevant patient information into the feature set, e.g. smoking status, age, and gender. In addition, a user-feedback mechanism can be built into the registry to allow care coordinator to flag the false-positive results for routine model retraining and improvement. The trade-off made in focusing on higher recall at the expense of precision means that the care coordinator will need to review more false positives at first, but by improving the model using user feedback, a better precision could be achieved. Finally, exporting the model to other facilities with no codes for URF will require further evaluation of performance, and the model may need to be retrained and retuned after further user feedback. In conclusion, this automated text classification algorithm could be used to augment the current URF process at PVAMC and start URF monitoring processes at other facilities.

Acknowledgements
The authors would like to thank Drs. Aaron Cohen and Steve Bedrick for the valuable suggestions and comments to help improve the performance of the classifiers. The authors would also like to thank Judy McConnachie for her input and support.

Figure 1. Processing pipeline.

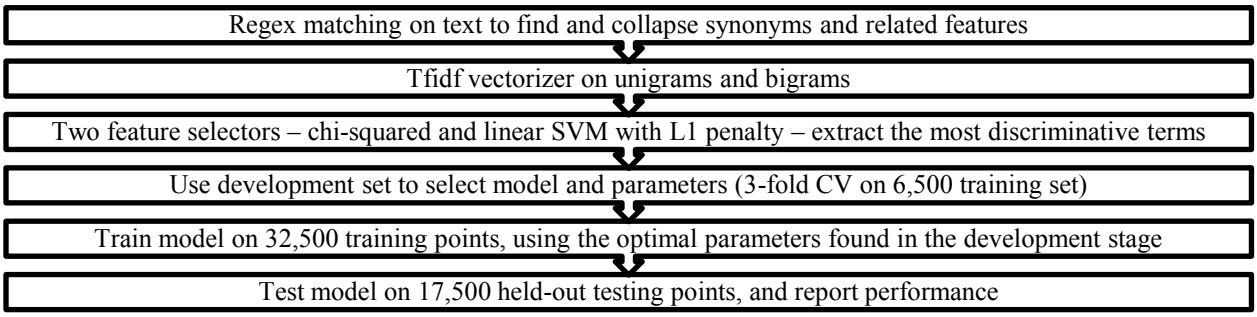


Table 1. Examples of term variations that were collapsed to related concepts.

Collapsed Concept	Regular Expression (examples)
MEASURE	(2.4mm, 2 by 5.5cm, 1.0 x 2.4cm, sub-1mm)
NODULE	nodul\w*("nodule," "nodular," etc.)
DATE	[0-9]+?[-/][0-9]+?[-/][0-9]+
LUNG	(lobe, lung, pleural, lingular, pulmonary, word beginning with pneumo-, etc.)
CANCER	(?:cancer|carcin|metastat)\w*

Table 2. Benchmarking algorithms. Table only displays the top six performing classifier results. Linear SVM with L1 penalty is the best performer on recall of URF-positive cases (underlined). SGD: Stochastic Gradient Descent, SVM: Linear Support Vector Classification.

Classifiers	L1-penalty SVC	L1-penalty SGD	L2-penalty SVC	L2-penalty SGD	Elastic-net penalty SGD	Naïve Bayes
URF-Positive Recall	**0.82**	0.79	0.78	0.78	0.78	0.77
URF-Positive precision	0.97	0.99	0.97	0.97	0.99	0.97
URF-Positive F1-score	0.89	0.88	0.87	0.87	0.87	0.86

Table 3. Performance of our tuned linear SVM with L2 penalty and regularization parameter C=0.9. Recall is the target for optimization. The final model reached a recall of 0.94 (underlined), with precision at 0.54. Performance of model on classifying both URF positive and negative cases, and the weighted average is reported.

	Precision	Recall	F1-score	Support
URF-Positive	0.54	**0.94**	0.69	772
URF-Negative	1.00	0.96	0.98	16,728
Weighted Average	0.98	0.96	0.97	17,500

References

1. Holden WE, Lewinson DM, Osborne ML, Griffin C, Spencer A, Duncan C, Deffebach ME. Use of a Clinical Pathway to Manage Unsuspected Radiologic Findings. *CHEST 2004: 125(4): 1753-1760.*
2. Danforth KN, Early MI and et al. Automated Identification of Patients with Pulmonary Nodules in an Integrated Health System Using Administrative Health Plan Data, Radiology Reports, and Natural Language Processing. *J. Thorac Oncol. 2012;7: 1257-1262*
3. Dutta S, Long WJ, Automated Detection Using Natural Language Processing of Radiologists recommendations for Additional Imaging of Incidental Findings. *Ann. Emerg Med.2013 Vol. 62;2:162-169*
4. Personal communication with New Haven Veterans Affair Medical Center Cancer Tracking Program.
5. Sebastiani, F. Machine learning in automated text categorization. *ACM CSUR. 2002 Vol. 34; 1: 1-47.*
6. Cohen AM. Five-way Smoking Status Classification Using Text Hot-Spot Identification and Error-correcting Output Codes. *J Am Med Inform Assoc 2008;1532-35*
7. Scikit-learn Machine Learning in Python. [cited on March 11, 2014]; Available from: http://scikit-learn.org/

Understanding Local Drivers of Health Outcomes: The North Carolina Community Health Information Portal

Charlene A. Wong, MPH[1], Deb Aldridge, MSN, RN-BC[2]
Annette Dubard, MD, MPH[1], Chase Haddix, MHA[1]

[1]Community Care of North Carolina, Raleigh, NC;
[2]Community Care of Southern Piedmont, Concord, NC

Background

Healthcare providers, public health workers, and community planners increasingly recognize the need to combine, analyze, and display health information in ways that promote understanding of clinical, social, and economic determinants of health. Geographic information system applications have proven to be highly impactful in targeting ways to improve health outcomes.

Methods

With federal funding from the Office of National Coordinator Beacon Award, Community Care of North Carolina, Southern Piedmont Beacon Community, NC Institute for Public Health, American Academy of Family Physicians' Robert Graham Center, and HealthLandscapes, Inc. collaborated to design a web-based geospatial visualization tool that integrates several public and clinical health indicators sourced from numerous databases including the US Census Bureau, Centers for Disease Control and Prevention, NC State Center for Health Statistics, Centers for Medicare and Medicaid Services Chronic Condition Warehouse.

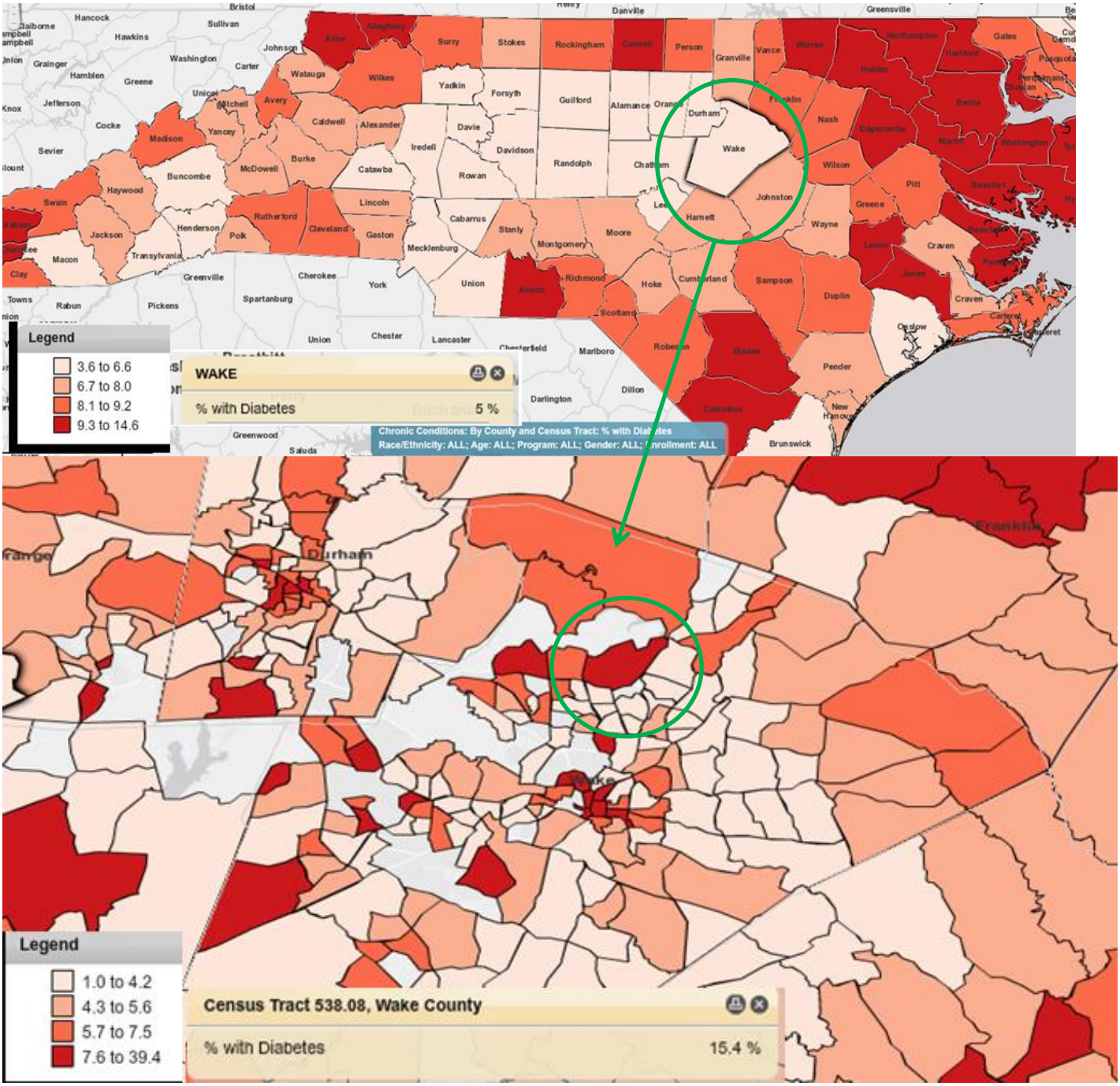
Results

The North Carolina Community Health Information Portal (NC-HIP, https://nchip.n3cn.org/) contains over 230 public and clinical health indicators including Medicaid and Medicare demographics, disease prevalence, cost, utilization and quality indicators; and general population indicators of socioeconomic and public health status. NC-HIP can geographically display disease burden at the health referral region, county, or census-tract level within North Carolina (Figure 1). Additionally, users can upload their own health data and compare their data with pre-loaded health indicators. Unique to NC-HIP, users can filter select Medicaid indicators based on gender, age, and race/ethnicity to examine health disparities at the local level.

Conclusion

NC-HIP equips healthcare providers, public health workers, policy-makers and other stakeholders with extensive health-related information to effectively respond to community needs through innovative data visualization. By revealing correlations between disease, social determinants, and access to care, NC-HIP serves as a resource for compiling a Community Health Needs Assessment (CHNA), guides optimal allocation of healthcare resources, and catalyzes targeted quality improvement activities at the local level.

Figure 1. Diabetes prevalence among Medicaid beneficiaries seen at the county level (top image) and at the census-tract level (bottom image, census tracts displayed for Wake County, NC only). Darker colors indicate greater diabetes prevalence. (Data source: Medicaid Paid Claims)



Identifying Clinical Decision Support Failures using Change-point Detection

Adam Wright, PhD[1,2,3], Francine L. Maloney, MPH[3], Rachel Ramoni, DMD, ScD[2,4], Milos Hauskrecht, PhD[5], Peter J. Embi, MD, MS[6], Pamela Neri, MS[3], Dean F. Sittig, PhD[7], David W. Bates, MD, MSc[1,2,3]

[1]Brigham and Women's Hospital, Boston, MA [2]Harvard Medical School, Boston, MA
[3]Partners HealthCare, Boston, MA [4]Harvard School of Dental Medicine, Boston, MA
[5]University of Pittsburgh, Pittsburgh, PA [6]The Ohio State University, Columbus, OH
[7]University of Texas Health Science Center, Houston, TX

Introduction: Evidence suggests that CDS can improve health care quality, safety, and effectiveness (1). However, unintended consequences and safety issues around CDS have also been reported (2). In prior qualitative studies, we identified several cases where CDS systems failed, in many cases without being noticed. Some of these failures led to patient harm. We sought to develop a method for detecting CDS failures.

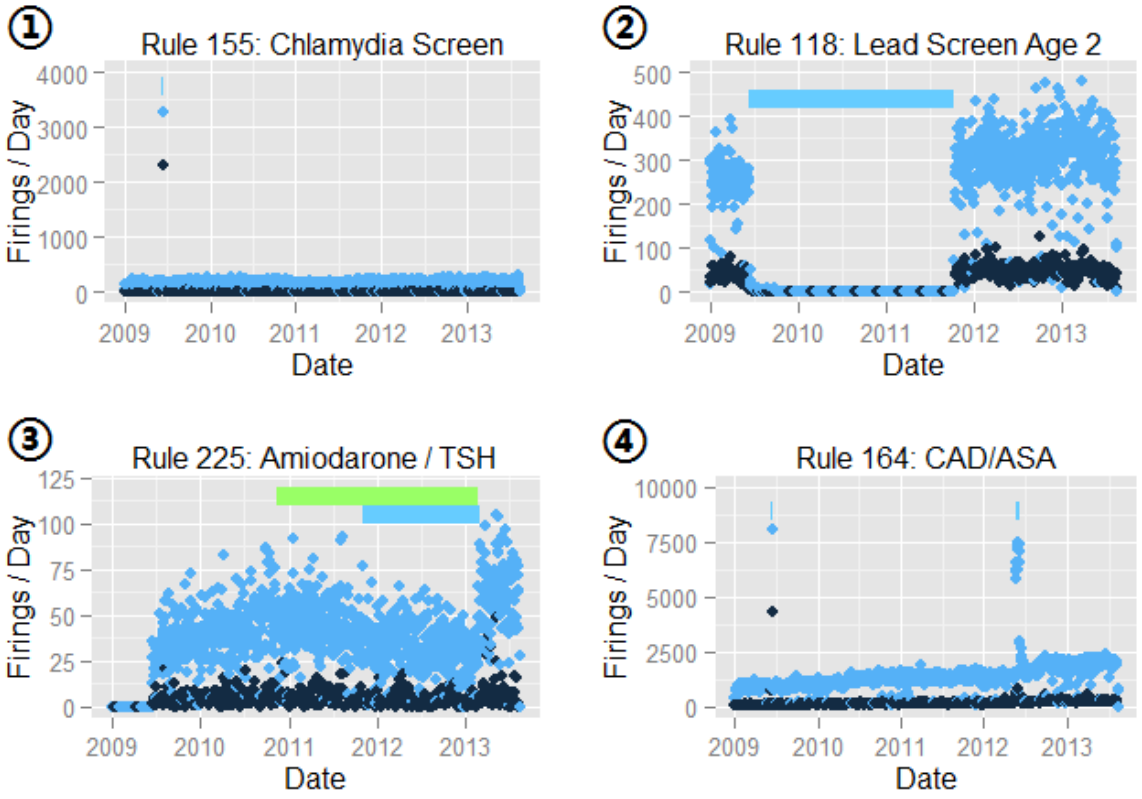
Methods: Change-point detection is a simple form of anomaly detection aimed at identifying changes in sequential and time series data. We used five years of data on alert firing from Brigham and Women's Hospital (BWH)'s outpatient electronic medical record system. We built Poisson change-point detectors for each alert separately. Change-point detectors assume that data is generated by a consistent process and then look for points where it is likely that the generating process underwent a change in expected rate (3).

Results: Figure 1 shows the firing rates for four alerts from BWH. The blue bars show anomalies detected using change-point analysis. Of the anomalies, one was previously known to both clinical leadership and the EHR development team, one was known to the EHR team only, and two were previously unknown. We investigated each of these four anomalies, and found:

1. System-wide spike in alerts: Panel 1 shows a representative alert for chlamydia screening which reveals a large spike in the total number of alerts in a two-day period on June 15 and 16, 2009. The same spike is seen in all alerts at BWH. We found that a bug was introduced after a software update, causing incorrect and repeated alerts to appear – sometimes dozens per patient. The EHR team received several support calls and, after two days, resolved the bug by reverting to an earlier version of the code. Although the EHR team resolved the bug, they did not alert users or clinical leadership about the issue after resolution.
2. Lead screenings for 2 year-olds: At BWH, clinicians are alerted to screen children for lead at ages 1, 2, 3 and 4. Panel 2 shows the firing rate for the lead screening alert for 2 year olds. We found a large anomaly from 2009 to 2011, where the alert slowed abruptly, stopped and then abruptly resumed. Interviews with clinical, risk management and technical staff revealed that no one was aware of the issue before we identified it. Manually maintained change logs show no edits to the alert since 2006, and the root cause has not been identified.
3. Amiodarone/thyroid stimulating hormone (TSH) testing: BWH uses an alert to remind clinicians to monitor TSH in patients receiving amiodarone. In November, 2010 BWH changed the internal code for amiodarone, but only for new orders. The logic of the rule was not updated when the code was changed, so the alert continued to fire for patients already on amiodarone, but not patients newly started on the drug and the rule started slowly failing in November, 2010, represented by the slow decline in the graph. In February, 2013, we discovered the error coincidentally during a demonstration of the system and reported it to the Partners knowledge management team, who fixed it immediately, leading to the step-change in firing rate.
4. Aspirin and coronary artery disease: This alert suggests starting aspirin in patients with coronary artery disease who were not already receiving antiplatelet therapy. Upon investigation, it appears that this spike was due to a malfunction in the drug classification service, which caused the alert to fire for patients with CAD, even if they were already taking aspirin. The problem resolved when the classification services were fixed in response to a non-alert-related issue. No BWH staff were aware of the issue previously.

Discussion CDS malfunctions represent a real threat to patient safety. These relatively simple change-point detection algorithms we developed identified four actual CDS anomalies, only one of which was previously known. In future work, we plan to expand our system to include more sophisticated models for anomaly detection, to run it on additional types of CDS and to design and develop a real-time dashboard for detecting anomalies before they lead to widespread patient harm.

Figure 1: Firing rate of four alerts at BWH over a five-year period (weekend days in black, weekdays in blue) with anomalies indicated (horizontal blue bars show detected anomalies, green bar shows known anomaly).



References

1. Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. BMJ. 2005;330(7494):765.
2. Ash JS, Sittig DF, Campbell EM, Guappone KP, Dykstra RH. Some unintended consequences of clinical decision support systems. Proc AMIA Ann Symp. 2007:26-30. Epub 2008/08/13.
3. Page E. Continuous inspection schemes. Biometrika. 1954;41(1/2):100-15.

Layered Spaces for Clinical Information Retrieval

Stephen Wu, PhD, Dingcheng Li, PhD, James Masanz, MS, Hongfang Liu, PhD
Mayo Clinic, Rochester, MN

Introduction

In clinical natural language processing (NLP), significant attenion has been directed toward and information extraction (IE) techniques; systems such as cTAKES[1] typically require some tailoring to specific use cases. In contrast, clinical information retrieval (IR) techniques intentionally generalize across use cases – in the form of queries. We hypothesize that clinical IR techniques can benefit by the inclusion of existing NLP techniques. To test this hypothesis, we introduce several NLP-based "layers" for IR in a language modeling framework, evaluating on a cohort identification task as introduced by the TREC Medical Records Track[2].

Methods

Our retrieval models use a language modeling approach, which embodies the intuition that a user will imagine a document that he/she wants, then write a query to match it. After preprocessing as in Figure 1, we include the following layers via interpolation (with example queries):

- Text: A standard Query Likelihood model[3]. For example, "patients who have a carotid endarterectomy".
- Concepts: UMLS CUIs produced in cTAKES. For example, "C0014099 C0014098 C0014098" where cTAKES yields both unique CUIs on "carotid endarterectomy", plus the latter one on "endarterectomy".
- Dependencies: Dependency parse nodes and arcs produced in cTAKES. For example, the dependency arcs "have↷endarterectomy carotid↶endarterectomy".
- Topics: Inferred from an LDA model with 800 topics.

Documents were retrieved from the TREC-med data set[2].

Results

Table 1 shows retrieval performance in MAP (mean average precision) from the inclusion of various layers. It is clear that the basic Text layer is the most useful of all tested layers, followed by Concepts, Dependencies, and Topics. The concept space yields the largest gain in performance over a baseline model, as it appears to complement the Text space well. Surprisingly, Topics (latent semantics), Dependencies (dependency arcs), and an overall combined model do not yield significant improvements on performance. It is important to note that the implementation details and data representation have a significant effect on system performance. We optimized for the Dirichlet smoothing parameter, but not for the number of LDA topics, the dependency parse representation, or linear interpolation coefficients.

Discussion

Our results imply that alternative representations of documents, no matter how linguistically sound or medically salient, have yet to completely replace text search. State-of-the-art language modeling systems[4] achieve excellent performance without the use of other layers (MAP=0.412). While work on alternative layers has shown promise[5, 6], they are rarely as optimized as the text layer, and further work needs to be done to capture their full potential.

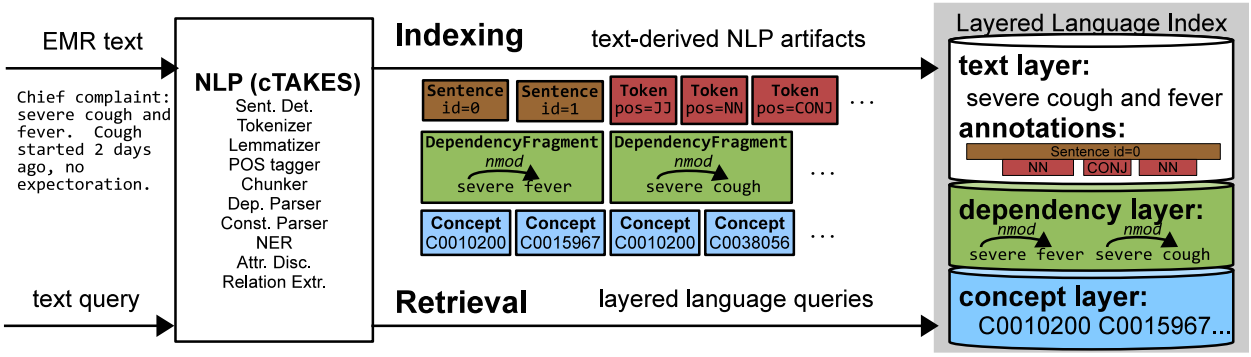


Figure 1: Indexing and retrieval pipelines incorporating several NLP techniques for IR

Model	**+Text**	**+Concepts**	**+Dependencies**	**+Topics**
Text	0.3261	0.3308	0.3266	0.3261
Concepts		0.2998	0.2894	0.2922
Dependencies			0.0889	0.0864
Topics				0.0807

Table 1: Retrieval performance (Mean Average Precision) for various combinations of language layers

References

1. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010 Sep-Oct;17(5):507-13.
2. Voorhees E, Hersh W, editors. Overview of the TREC 2012 medical records track. The Twenty-first Text REtrieval Conference Proceedings TREC; 2012; Gaithersburg, MD: National Institute of Standards and Technology.
3. Manning CD, Raghavan P, Schütze H. Introduction to information retrieval: Cambridge University Press Cambridge; 2008.
4. Zhu D, Carterette B, editors. Combining multi-level evidence for medical record retrieval. Proceedings of the 2012 international workshop on Smart health and wellbeing; 2012: ACM.
5. Wei X, Croft WB, editors. LDA-based document models for ad-hoc retrieval. Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval; 2006: ACM.
6. Zuccon G, Koopman B, Nguyen A, Vickers D, Butt L, editors. Exploiting medical hierarchies for concept-based information retrieval. Proceedings of the Seventeenth Australasian Document Computing Symposium; 2012: ACM.

Mining electronic health record data to detect drug-repurposing signals for cancers

Hua Xu Ph.D.[1], Qingxia Chen, Ph.D.[2], Jeremy Warner M.D., M.S. [3,4], Xue Han, M.S.[2], Min Jiang M.S.[1], Anushi Shah, M.S.[4], Melinda C. Aldrich, M.P.H., Ph.D. [5,6], Qi Dai, M.D.[6], Joshua C. Denny M.D., M.S. [3,4]

[1]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX; [2]Department of Biostatistics, [3]Department of Medicine, [4]Department of Biomedical Informatics, [5]Department of Thoracic Surgery, [6]Division of Epidemiology, School of Medicine, Vanderbilt University, Nashville, TN, USA;

Introduction: Rapid growth in the implementation of large electronic health records (EHRs) has led to an unprecedented expansion in the availability of longitudinal datasets for epidemiologic and genomic research.[1] An exciting, but largely unexplored application of EHRs is to use them for drug repurposing studies, which seek new indications for existing drugs.[2] For example, recent evidence suggests metformin, a first-line therapy for type 2 diabetes, improves cancer survival improves cancer survival [3,4] and decreases cancer risk [5,6] versus other glucose-lowering therapies, indicating its clinical promise as a cancer therapeutic and an antineoplastic agent. We hypothesize that potential new indications of existing drugs, if real, can be quickly detected using EHRs, with the help of informatics tools and methods.

Methods: We linked a de-identified version of the Vanderbilt Tumor Registry to the Synthetic Derivative (SD) database, a de-identified copy of Vanderbilt's EHR. We identified a cancer cohort (excluding non-melanoma skin cancer) of 65,869 individuals. We limited drugs that were associated with more than 5,000 subjects in the SD, resulting in 301 generic drug ingredients. We further removed over-the-counter drugs, antineoplastics, and drugs used in supportive care of cancer patients, resulting in 226 drugs. For each drug, we identified cancer subjects who were exposed vs. unexposed to the drug using MedEx, a medication information extraction system for clinical text. Other covariates and outcomes included patient demographics (age, sex, race), tumor information (e.g., tumor type and stage), and time from cancer diagnosis to death or last medical record date (censored). To adjust for co-morbidity, we also included ICD 9 codes (mapped to disease groups) associated with each patient. We then examined each drug for its association with cancer mortality using Cox proportional hazards models controlling for the above covariates. We ranked all 226 drugs based on the size of the hazard ratio (HR) from Cox models. For the top 20 ranked drugs associated with reduced mortality, we manually searched Google or PubMed for prior evidence of the association.

Results: A total of 43,310 individuals were exposed to the 226 drugs included in this analysis. Table 1 shows the top 20 drugs that were predicted to be associated with improved cancer survival in this study. We found strong associations for a number of statins, angiotensin receptor blockers (but not angiotensin converting enzyme inhibitors), beta-blockers, and a few other medications. Metformin was ranked 18th. Among these 20 drugs, we found supporting evidence for 13 drugs. Some evidence was strong, e.g., the National Surgical Adjuvant Breast and Bowel Project is conducting a phase 3 study of rosuvastatin (ranked 1st in our study) versus placebo in patients with surgically removed stage I or II colorectal cancer to determine whether recurrence can be prevented. Much of the evidence is relatively weak. For example, sildenafil, ranked 7th in our list, is associated with a publication suggesting that it enhances the sensitivity of cancers to standard chemotherapeutic drugs.[7] If we further filter the drug list by p-value, thirteen of them have a statistically significant p-value ($<2.2x10^{-4}$, using Bonferroni correction for multiple comparisons with type-I error rate of 0.05). Among them, nine drugs have strong or weak evidence from literature, showing they could be related to cancer survival.

Conclusion: By linking a tumor registry to a large EHR database, we conducted a large scale screening of 226 drugs for their potentials for cancer therapy. Using the Cox model, we ranked drugs according to their potential to reduce mortality. Manual review showed supporting evidence in the literature for some top-ranked drugs, indicating the application of EHRs for drug repositioning. More data is needed to validate these associations for potential clinical use. With broader application of informatics methods in EHR data extraction and analysis, we envision rapid generation of large datasets for drug repurposing discovery and validation.

Table 1. Top 20 drugs that were associated with improved cancer survival, as predicted by this study. Drugs were ranked based on Hazard Ratio. Insignificant p-values ($<2.2x10^{-4}$) were highlighted in grey.

Rank	Drug Name	Literature Evidence	Hazard Ratio (HR)	HR 95% CI		P-value
1	rosuvastatin	Yes – strong	0.6014	0.5158	0.7013	8.66E-11
2	olmesartan	No	0.624	0.5136	0.7581	2.05E-06
3	ibandronate	Yes – weak	0.6419	0.4995	0.8249	5.31E-04
4	aripiprazole	No	0.6703	0.4449	1.0098	5.57E-02
5	carvedilol	Yes – weak	0.6773	0.5964	0.7691	1.89E-09
6	ropinirole	No	0.6837	0.5265	0.8879	4.34E-03
7	sildenafil	Yes – weak	0.6854	0.6109	0.7689	1.20E-10
8	insulin.lispro	No	0.7004	0.5899	0.8317	4.84E-05
9	desloratadine	Yes – weak	0.7053	0.5761	0.8634	7.18E-04
10	ergocalciferol	Yes – strong	0.7086	0.5435	0.9238	1.09E-02
11	pravastatin	Yes – strong	0.7179	0.6307	0.8172	5.32E-07
12	rosiglitazone	Yes – weak	0.72	0.6170	0.8403	3.08E-05
13	Metoprolol tartrate	No	0.7374	0.6280	0.8658	2.00E-04
14	thyroxine	No	0.746	0.6965	0.7992	< e-16
15	simvastatin	Yes – strong	0.7537	0.7080	0.8024	< e-16
16	ezetimibe	Yes – strong	0.754	0.6537	0.8698	1.07E-04
17	niacin	Yes – strong	0.7554	0.6544	0.8720	1.28E-04
18	metformin	Yes – strong	0.7571	0.6956	0.8241	1.22E-10
19	fenofibrate	Yes – weak	0.7577	0.6415	0.8948	1.08E-03
20	irbesartan	No	0.7668	0.6480	0.9074	1.99E-03

References

1. Kohane IS. Using electronic health records to drive discovery in disease genomics. *Nature Reviews Genetics.* 2011;12(6):417-428.

2. Tatonetti NP, Patrick PY, Daneshjou R, Altman RB. Data-driven prediction of drug effects and interactions. *Science translational medicine.* 2012;4(125):125ra131-125ra131.

3. Landman GW, Kleefstra N, van Hateren KJ, Groenier KH, Gans RO, Bilo HJ. Metformin associated with lower cancer mortality in type 2 diabetes: ZODIAC-16. *Diabetes Care.* Feb 2010;33(2):322-326.

4. Currie CJ, Poole CD, Jenkins-Jones S, Gale EA, Johnson JA, Morgan CL. Mortality after incident cancer in people with and without type 2 diabetes: impact of metformin on survival. *Diabetes Care.* Feb 2012;35(2):299-304.

5. Evans JM, Donnelly LA, Emslie-Smith AM, Alessi DR, Morris AD. Metformin and reduced risk of cancer in diabetic patients. *Bmj.* Jun 4 2005;330(7503):1304-1305.

6. Ruiter R, Visser LE, van Herk-Sukel MP, et al. Lower risk of cancer in patients on metformin in comparison with those on sulfonylurea derivatives: results from a large population-based follow-up study. *Diabetes Care.* Jan 2012;35(1):119-124.

7. Shi Z, Tiwari AK, Patel AS, Fu LW, Chen ZS. Roles of sildenafil in enhancing drug sensitivity in cancer. *Cancer Res.* Jun 1 2011;71(11):3735-3738.

Operationalizing patient-generated health data: Home blood pressure monitoring as an example

Shipeng Yu, PhD[1], J. Marc Overhage, MD, PhD[1], Paul Tang, MD[2]
[1] Siemens Health Services, Malvern, PA [2]Palo Alto Medical Foundation, Palo Alto, CA

Introduction

Patient-generated health data (PGHD) are health-related data that are created, recorded, gathered, or inferred by from patients or their designees to help address a health concern. It is widely believed that engaging patients to play a more active role in their health and to collect PGHD can lead to better outcomes and more efficient care, but in order to achieve those benefits we need to better understand the workflows that will allow health care providers' electronically accepting and incorporating patient-generated health information into their clinical workflows. Home blood pressure monitoring is a good example of PGHD: the data are high volume, frequent and use in practice is not well understood and for which automation may play an important assistive role.[1] The objective of the current study is to explore the effectiveness of some automated trending algorithms on HBPM data, and to retrospectively compare with the intervention logs from clinicians acting as dedicated care managers who monitored these patients without the automated tool in order to shed some light on future research of operationalizing PGHD and seamlessly integrating PGHD and data generated through clinical encounters.

Methods

We used data drawn from EMPOWER-H, a trial of patient centric care management of hypertension, for this study. The 142 patients participating in this study were to record their BP measurements twice daily and share them with care managers who reviewed that data and used their clinical judgment to intervene with patients and their PCPs. The time frame for this study was from March 2012 to January 2013. We explored three trending algorithms to fit the HBP data from each patient in the study period, separately for systolic BP (SBP) and diastolic BP (DBP): the *Moving Average* method computes the mean BP measurements over the past 14 days at every time point; the *Polynomial Fit* method conducts a 5-degree polynomial fit to the BP measurement data; and the *Gaussian Process Fit* method does a customized, non-parametric functional fitting to the BP measurement data. We then designed an automated alerting system where an alert was generated if either SBP or DBP 14-day moving average is above a predefined threshold (135 for SBP and 85 for DBP) for 14 continuous days. Once an alert was generated, the algorithm waited for 14 days to be activated again for further alert generation. We then compared the alerting algorithms with care managers' interventions that involved a medication change or suggested clinic visit or consultation. For each alert that was generated, we looked at whether there was an intervention within 7 days before or after the alerting date. We can then compute the accuracy of the alerting system treating the clinician's interventions as the gold standard.

Results

The 14-day moving average method generated 477 alerts in total, about 3.5 alerts per patient for the 10 month period. 56 patients (39%) did not have any alert, and 86 patients (61%) had at least one alert, out of which 19 patients (13%) had more than 10 alerts. Out of the 477 alerts, 113 (24%) had an intervention within 7 days before the alert, 88 (19%) had an intervention within 7 days after the alert, and 168 (35%) had an intervention within 7 days before or after the alert. Out of the 86 patients that had at least one alert, 28 (33%) had no intervention in the 7-day radius of any alert, 26 (30%) had interventions for at least 50% of the alerts, and 2 (2%) had interventions for 100% of the alerts (Figure 1).

Discussion

The initial results revealed that the algorithm did not generate any alerts for about 40% of the patients suggesting that their hypertension was under control. The concordance between the alerts generated by our algorithm and the patient states that triggered timely medical intervention was limited (around 35%). A careful review of all alerts for a selected subset of 10 patients indicated that often times the clinician elected not to make any changes in medical treatment at the time because the clinicians were aware of other considerations (e.g., patient preference not to change medications, desire to try lifestyle modification, patient undergoing stressful circumstances). We would expect to generate better alerts if the algorithm incorporated knowledge of these factors.

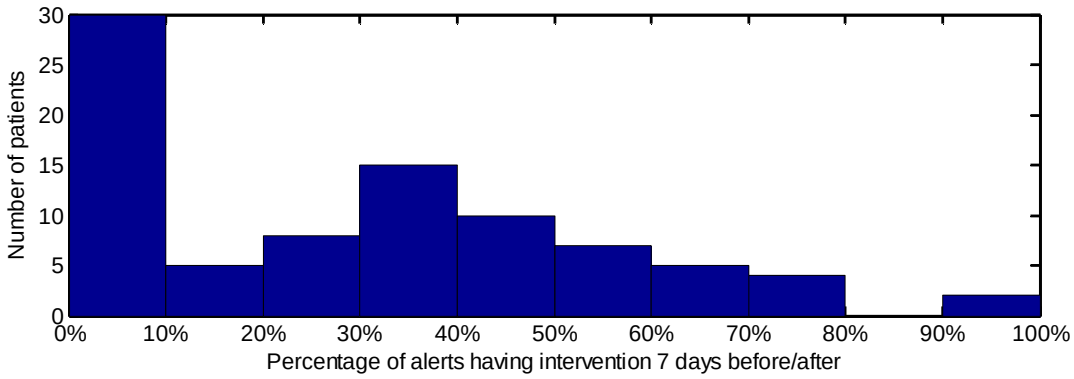


Figure 1 -- Clinicians did not intervene at all for 28 patients in whom the algorithm generated alerts and for two patients clinicians had intervened in all cases in which the algorithm generated an alert. Overall, alerts corresponded to clinician's interventions in about a third of cases.

References

[1] Shapiro M, Johnston D, Wald J and Mon D. Patient-Generated Health Data: White Paper Prepared for the Office of the National Coordinator for Health IT by RTI International. April 2012.

Domain Adaptation for Semantic Role Labeling of Clinical Text

Yaoyun Zhang, Ph.D., Buzhou Tang, Ph.D., Min Jiang, MS, Jingqi Wang, MS, Yonghui Wu Ph.D., Hua Xu Ph.D.
School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA

Introduction

Semantic role labeling (SRL), which extracts a shallow semantic relation representation from different surface textual forms of free text sentences (e.g., in "*liver biopsy*", "*liver*" is the source of the "*biopsy*" event), is important for understanding natural language.[1] While SRL approaches and corpora in open English domains are plentiful, few SRL studies have been conducted in the medical domain. This is probably due to the lack of annotated clinical SRL corpora, which are time-consuming and costly to build in the medical domain. Our study investigates domain adaptation techniques for clinical SRL. We want to leverage SRL resources from newswire and biomedical literature to improve performance and decrease annotation cost for clinical SRL. To the best of our knowledge, this is the first work introducing domain adaptation algorithms to address the task of clinical SRL.

Methods

Datasets: The PropBank corpus generated using newswire of financial domain and the BioProp corpus from biomedical literature specific to proteins were the source domain datasets. MiPACQ,[2] a clinical corpus with manually annotated SRL information, was the target domain dataset. **Algorithms:** Three state-of-the-art domain adaptation algorithms were employed: (1) transfer self-training (TransferSelf), utilizing top n-source domain instances with highest similarity to the target domain for enrichment of the target training set;[3] (2) instance pruning (InstancePrune), merging the source domain dataset with the target training set after removing the top n-most different instances from the target domain;[4] (3) feature augmentation (FeatureAug), amplifying target training features with weight-adapted source domain features.[5] **Evaluation:** The SRL performance using different domain adaptation algorithms with state-of-the-art open domain SRL features[6,7] was evaluated using 10-fold cross validation on the MiPACQ corpus by precision, recall and F-measure. The performance of the two subtasks argument identification (AI) and classification (AC), together with the combined task is reported. Three baselines; using the source domain only, target domain alone and direct merging the source and target domain datasets for training were produced and compared. To assess the effect of sample size, we also generated learning curves (plots between training sample size and F-measure) for different methods.

Results

As displayed in Table 1, domain adaptation using both the outside resources, PropBank and BioProp improved the SRL performance on MiPACQ. The FeatureAug algorithm with PropBank as the source domain dataset achieved a statistically significant higher F-measure as compared to the baseline using MiPACQ dataset only (F-measures of 82.83% & 81.53% respectively), indicating that domain adaptation algorithms could improve SRL performance for clinical text. Use of the FeatureAug algorithm leveraging the PropBank corpus with MiPACQ required lower amounts of the dataset for training (<50%) to perform at a comparable level as the baseline of MiPACQ alone (90%) (Figure 1). This demonstrates that annotation cost of clinical SRL can be reduced significantly by leveraging existing SRL resources from other domains.

Discussion

In this study, we address SRL of clinical narratives as a domain adaptation problem. We used two existing SRL corpora outside of the clinical domain and evaluated three state-of-the-art domain adaptation algorithms for the task of SRL of clinical text. Our results showed that domain adaptation strategies not only improved the performance of SRL on clinical text but also reduced annotation costs.

Further, different domain adaptation algorithms may be effective for different source domain datasets. FeatureAug achieved the best performance on PropBank, while InstancePrune achieved the best performance on BioProp. PropBank is assumed to be more diverse from the clinical domain as compared to BioProp. PropBank also has a larger dataset than MiPACQ (~9:1), thus increasing contribution of more useful features as well as noise distinct from the features of MiPACQ. As illustrated in Figure 2, increased sample size of PropBank decreased SRL performance without domain adaptation. Nonetheless, the feature weighting mechanism of FeatureAug facilitated the choice of effective features from PropBank, demonstrating a larger tolerance for domain gap. The data size of BioProp is smaller than MiPACQ (~1:6.4), with a lower feature distribution when merged with MiPACQ. InstancePrune achieved better performance by removing dissimilar instances along with potentially noisy features from BioProp.

Acknowledgement: This study was supported in part by National Institute of General Medical Sciences grant 1R01GM102282, National Library of Medicine grant R01LM010681, National Cancer Institute grant R01CA141307, Cancer Prevention & Research Institute of Texas grant R1307.

Table 1 Performance with and without domain adaptation using PropBank/BioProp (%).

		AI			AC	AI+AC		
		P	R	F_1	Acc	P	R	F_1
MiPACQ only		93.24	92.81	93.02	87.65	81.72	81.34	81.53
PropBank + MiPACQ	Source only	86.54	76.36	81.13	72.57	62.80	55.41	58.87
	Source&Target	93.99	90.36	92.14	86.60	81.39	78.25	79.79
	FeatureAug	94.08	93.82	93.95	88.17	82.95	82.71	**82.83***
	TransferSelf	93.23	92.77	93.00	87.68	81.74	81.33	81.54
	InstancePrune	94.22	92.70	93.45	86.95	81.93	80.60	81.26
BioProp + Mi-PACQ	Source only	53.48	30.62	38.95	53.06	28.38	16.25	20.67
	Source&Target	93.43	92.07	92.74	88.07	82.28	81.08	81.68
	FeatureAug	93.13	92.83	92.98	87.79	81.76	81.50	81.63
	TransferSelf	93.41	92.77	93.09	87.82	82.04	81.48	81.75
	InstancePrune	93.43	92.68	93.05	88.07	82.28	81.62	**81.95**

AI: Argument Identification AC: Argument Classification *: statistically significant by the Wilcoxon signed-rank test

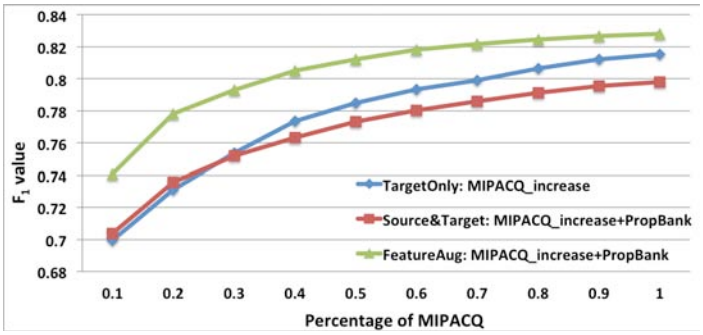


Figure 1 Learning curve with increasing the percentage of MiPACQ available for training. For MiPACQ_increase+PropBank, the whole source dataset of PropBank is used for training. The y-axis denotes the averaged F_1 value, using 10-fold cross-validation. The three learning curves are for using the target dataset only, directly merging the source and target datasets and domain adaptation using the FeatureAug algorithm, respectively.

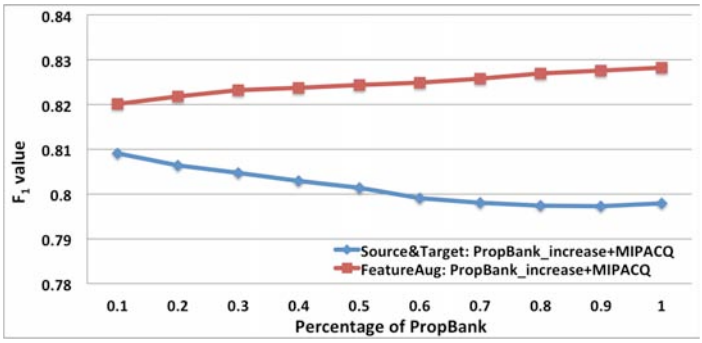


Figure 2 Learning curve with increasing the percentage of PropBank available for training, together with 9-fold of the target dataset MiPACQ. The y-axis denotes the averaged F_1 value, using 10-fold cross-validation. The two learning curves are for directly merging the source and target datasets and domain adaptation using the FeatureAug algorithm, respectively.

Reference

1. Pradhan, S. S., Ward, W. & Martin, J. H. Towards robust semantic role labeling. *Computational Linguistics.* 2008;34:289–310.
2. Albright, D., et al. Towards comprehensive syntactic and semantic annotations of the clinical narrative. *Journal of the American Medical Informatics Association*. 2013;0:1–9.
3. Xu, R., Xu, J. & Wang, X. Instance level transfer learning for cross lingual opinion analysis. *Proc of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis.* 2011;182–188.
4. Jiang, J., Zhai, C. Instance weighting for domain adaptation in NLP. *Proc of ACL*. 2007;264-271.
5. Daumé III, H. Frustratingly easy domain adaptation. *Proc of ACL*. 2007;256-263.
6. Pradhan S, Hacioglu K, Krugler V, *et al*. Support Vector Learning for Semantic Argument Classification. *Machine Learning* 2005;60(1-3):11–39.
7. Gildea D, Jurafsky D. Automatic Labeling of Semantic Roles. *Computational Linguistics.* 2002; 28(3):245–88.

Qualitative evaluation of three phenotype information models to find methotrexate liver injury

Qian Zhu, PhD[1], Huan Mo, MD, MS[2], Luke V. Rasmussen[3], Andrew Post, MD, PhD[4], Jennifer A. Pacheco[3], Jie Xu, MS[3], Richard C. Kiefer[1], Peter Speltz[2], Enid Montague, PhD[3], William K. Thompson, PhD[3], Joshua C. Denny, MD, MS[2], Jyotishman Pathak, PhD[1]

[1]Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA
[2]Department of Biomedical Informatics, Vanderbilt University, Nashville, TN
[3]Department of Preventive Medicine, Northwestern University, Chicago, IL, USA
[4]Department of Biomedical Informatics, Emory University, Atlanta, GA, USA

Introduction The fundamental basis for EHR driven phenotyping is to represent a query or algorithm, such that it may be executed against a repository of clinical data. Hence, the overarching goal of this project is to formally evaluate existing phenotype algorithm representation models, identify gaps, and where applicable, propose and implement extensions. To achieve this objective, we evaluated the National Quality Forum (NQF) Quality Data Model (QDM)[1] and HL7 Health Quality Measure Format (HQMF)[2] for the Measure Authoring Tool (MAT)[3], star schema and ontology for Informatics for Integrating Biology and the Bedside (i2b2)[4], and temporal abstraction ontology (TAO)[5] for Eureka! Clinical Analytics (Eureka)[6].

Materials and Methods

Phenotype algorithm: The "Methotrexate (MTX) drug-induced liver injury" phenotype algorithm shown in Figure 1 was selected for this evaluation task because it was a relatively simple algorithm that includes the complexity common to many more complicated phenotypes, such as, 1) multiple criteria (inclusion and exclusion), 2) multiple data types (diagnosis, medication, laboratory, etc.), and 3) temporal relationships.

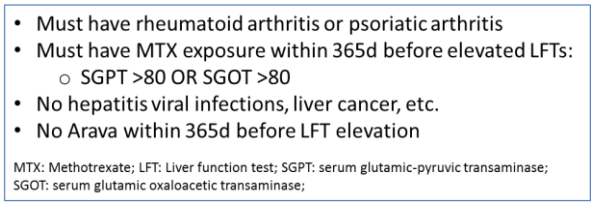


Figure 1. MTX drug-induced liver injury algorithm

Information model comparison: Three well-used phenotype authoring tools (MAT, i2b2 and Eureka) were explored in this study to identify gaps among the information models behind them by representing the MTX algorithm in each. QDM (used by MAT) is an information model describing clinical concepts in a standardized format. The star schema (used by i2b2) is designed to capture arbitrary data types and allow for rapidly developed analysis queries by incorporating with the i2b2 ontology (also called metadata) that is integral for querying the data. TAO (used by Eureka) provides a model for specifying phenotypes in terms of categories of codes, thresholds in value and slope, and both frequencies and sequences of both events and observations. During the implementation process, a list of major features for capturing semantics of phenotype algorithm has been identified and compared these three information models. The feature matrix for comparison is shown in Table 1.

Discussion In our evaluation, we have identified several strengths and limitations for the information models using in the three authoring tools. All of them have the capability to represent phenotyping algorithms for both machine and human consumption, and have their own advantages compared to the other two. For MAT, representation of terminology value sets within the QDM as code lists external to the measure definition, which are assigned a globally unique identifier, allows for their re-use. In addition, QDM requires strict definitions of value sets – code sets, code set version, a code and description for each entry. For i2b2, the query can be easily transformed and implemented against the backend clinical relational database. For Eureka, it includes well-defined temporal pattern types that cover a majority of temporal events occurring in phenotype algorithms. Phenotypes are composed from discrete building blocks (e.g., rheumatoid arthritis diagnosis codes, LFT value thresholds) that may be reused in other phenotypes. Additionally, we identified several limitations, including lack of support for specifying natural language processing (NLP) constructs and significant challenges in the ability to transform HQMF elements and components into executable queries (e.g., into SQL). These will be extensions proposed for future studies.

Acknowledgment

This work has been supported in part by funding from the NIH (R01-GM105688), and in part by PHS Grant UL1TR000454 from the CTSA Program, NIH, NCATS; and Grant Number R24HL085343 from the NHLBI.

Table 1. Feature Matrix of information models comparison

	QDM + HQMF (MAT)	Star Schema + Ontology (i2b2)	TAO (Eureka)
Data type (Diagnosis, Medication, Lab)	Yes	Yes	Yes
Comparison operators (equal to, greater than, less than, greater than or equal to, less than or equal to)	Yes	Yes	Yes
Logic operators (And, Or, Not)	Yes	Yes	Yes (*And* and *Or* supported in phenotype editor, *Not* supported by querying phenotypes in i2b2)
Arithmetic operators (Addition, Subtraction, Multiplication, Division, Modulo Reduction)	No	Yes	No
Aggregate Functions (MIN , MAX , SUM , AVG , COUNT)	Yes	No	No
Counting rules (e.g. multiple diagnosis codes or criteria to specify one phenotype)	Yes, but without clearly defined rules	Yes, but not for multiple criteria	Yes
Temporal Constraint	Yes	Yes (Temporal sequence of events)	Yes (Four types of temporal relations included, Category, Sequence, Frequency, Value threshold)
Complex Projection	No	No	No
Standardized terminologies	Yes Integrating VSAC	Yes (Standard terms being used)	Incomplete (Local terms being used)
Compatible to other formats	Yes (Translator being used from HQMF to i2b2[7])	Yes (Translator being used from HQMF to i2b2[7])	Yes
Logic sharing (whether partial components or entire algorithm) between algorithms	Yes	Yes	Yes
NLP support (ad-hoc or real-time)	No	No	No
Translatable into SQL	Yes (translated from HQMF -> i2b2 -> SQL	Yes	Yes

Reference

1. *NQF: Quality Data Model*. [cited 2014 March 1]; Available from: www.qualityforum.org/QualityDataModel.aspx.
2. *HL7 Health Quality Measure Format* [cited 2014 March 1]; Available from: http://www.hl7standards.com/blog/2009/09/17/what-is-hqmf-health-quality-measures-format/.
3. *NQF: Measure Authoring Tool (MAT)*. [cited 2014 March 1]; Available from: www.qualityforum.org/MAT/.
4. Murphy, S.N., et al., *Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2).* Journal of the American Medical Informatics Association, 2010. **17**(2): p. 124-130.
5. Post, A.R., et al., *Temporal Abstraction-based Clinical Phenotyping with Eureka!* AMIA Annual Symposium Proceedings, 2013: p. 1160-1169.
6. *Eureka! Clinical Analytics*. [cited 2014 March 1]; Available from: https://eureka.cci.emory.edu/about.jsp.
7. Klann, J.G. and S.N. Murphy, *Computing Health Quality Measures Using Informatics for Integrating Biology and the Bedside.* Journal of medical Internet research, 2013. **15**(4).

Allergies and Intolerances – Standards for Interoperability

Elaine J. Ayres, MS, RD[1], Russell Leftwich, MD[2], Lisa R. Nelson, MSc, MBA[3]
Laboratory for Informatics Development, NIH Clinical Center, National Institutes of Health, Bethesda, MD[1]; State of Tennessee, Office of eHealth Initiatives, Nashville, TN[2]; Life Over Time Solutions, Westerly, RI[3]

Abstract

The management and exchange of drug, food and environmental allergies and intolerances is essential for patient safety. Over the past several years, the international standards community has developed new models for the documentation and interoperability of allergies and intolerances. This panel will address new HL7 standards for adverse reactions, allergies and intolerances including V3 clinical models, FHIR resources and Consolidated CDA templates. Case studies will provide attendees with examples of how to represent key allergy and intolerance concepts including undifferentiated adverse reactions, reaction severity and condition criticality. The use of these concepts for quality measure reporting will also be discussed.

Learning Objective 1: Attendees will be able to describe the standards, data model attributes and terminology for documentation and interoperability of allergies and intolerances.
Learning Objective 2: Attendees will be able to compare and contrast Meaningful Use Stage 2 and 3 requirements for allergies and intolerances versus current HL7 documentation and transmissions standards including V3, FHIR and C-CDA models.

Introduction

Various information models exist for the representation of allergies and intolerances within electronic health records. HL7[1] has completed an extensive comparison of international models for the representation of adverse reactions, allergies and intolerances to create new international standards. This body of work has been balloted and published as a Domain Analysis Model along with Clinical Models for adverse reactions, allergies and intolerances. The HL7 Consolidated CDA (C-CDA) R1.1 standard includes section and entry level templates for the inclusion of allergy and intolerance data elements. The HL7 Fast Healthcare Interoperability Resources (FHIR) standard has been developed based on the V3 models. These standards now provide the EHR community with a variety of methods to document and exchange information on allergies and intolerances.

Standards for Stage 2 Meaningful Use and proposed standards for Stage 3 include the following core measure for medication allergies (§170.314(a)(7): *Medication allergy list - Enable a user to electronically record, change, and access a patient's active medication allergy list as well as medication allergy history: (i) in the Ambulatory setting over multiple encounters; or Inpatient setting for the duration of an entire hospitalization*[2]. Certified EHR systems use various formats for recording adverse reactions, allergies and intolerances. The "allergy list" may be a combination of observed or reported reactions (hives), or it may represent a condition such as "sulfa allergy". Lists may include both allergies and intolerances to medications, food, devices or environmental triggers. The list may represent a mix of concepts e.g. medications (aspirin), medication classes (analgesics), foods (peanuts), food groups (milk products), organisms (elm tree) or allergens such as elm pollen. Variations in concept representation and terminology binding have hindered interoperability and pose patient safety risks.

The following topical presentations will demonstrate available standards for the representation and interoperability of adverse reactions, allergies and intolerances. Audience discussion regarding current standards and issues will be sought following the presentations.

Elaine Ayres – HL7 Information Models for Adverse Reactions, Allergies and Intolerances Including FHIR Resources

In 2011, the HL7 Patient Care Work Group began a review of logical models related to allergies and intolerances. This review included models from the US Veterans Administration, Canada, Australia, the UK and the Netherlands. The development of a Domain Analysis Model with 15 use cases and functional models was balloted through HL7

in 2013 and published as an informative standard in 2014[3]. Clinical Models detailing the relationship of concepts were also developed and balloted in 2013 and published in 2014 as a Draft Standard for Trial Use (DSTU)[3]. This work represents the first universal standard for managing and exchanging adverse reactions, allergies and intolerances designed as well as support clinical decision support and quality reporting.

The HL7 V3 model for allergies and intolerances has several key concepts. In a clinical care episode when a patient presents with a symptom such as urticaria, the causative agent must be determined. Pending a clinical history, the classification of this event is an "adverse reaction". Therefore, the HL7 model uses the "adverse reaction" as the point of entry into documentation on the "list". Further clinical evaluation allows the reaction to be classified as an allergy or intolerance. The model proposes the use of substances to describe the causative agent of the adverse reaction. Based on current work of HL7, IHTSDO[4] and ONC[5], there is much discussion related to terminology bindings. At present the HL7 model is not prescriptive in this regard. Additions to the "list" may be represented as "reactions", the episodes reported by a patient, or as witnessed by the practitioners, or they may be represented as "conditions" such as a peanut allergy. Reactions may further be described with a severity classification. Conditions may include a clinical assessment of criticality. The model also describes the reconciliation of allergy and intolerance lists and the concept representation of "no known allergies or intolerances" or "unable to determine triggering agent".

FHIR is a next generation standards framework created by HL7[6]. FHIR modules (called resources) combine features of V2, V3 and CDA standards, with a strong focus on ease of implementation. A key component of a FHIR resource is the dependence on terminology binding. Resources for "adverse reaction" and "allergy and intolerance" are based on the constructs of the V3 model. The adverse reaction resource may be used as a singular concept e.g. hives with no known cause, or together with the allergy and intolerance resource. Data fields within the "list" also use the concepts of "substance" (with evolving terminology bindings), severity and criticality.

Russell Leftwich – Case study examples of adverse reactions, severity and criticality using HL7 models.

The historical model for the list of allergies and intolerances is the handwritten list on the front of a paper chart. This "list" might be a list of medications, occasionally a food or environmental substance and sometimes an annotation with an attribute, such as "severe". The interoperable exchange of this clinical data for the purposes of patient safety, clinical decision support, and quality metrics requires the use of a formal domain model that represents the essential information about reaction type, substance, and attributes that convey the criticality of the condition. This model is the basis for development of standards for exchange of clinical data[7].

Development of the HL7 standard is guided by use cases that portray the clinical scenarios, associated workflows, data requirements of recipients of information, and the associated semantic and process interoperability requirements. These use cases also serve as examples to inform adopters of the standard of its value and as guidance for implementation and integration into clinical workflows. Selected examples of use cases for adverse reactions, allergies and intolerances in this presentation illustrate these concepts[3]. A representation of these concepts within V3 messages and FHIR resources will demonstrate the utility to the EHR and transitions of care communities.

Lisa Nelson – Allergy and intolerance representation in standard digital documents

Exchange of standards-based interoperable data is the key to the successful communication of patients' allergy and intolerance information. It also is needed to automate the computation of quality measures designed to assess progress toward goals to improve availability of allergy and intolerance information in electronic health record (EHR) systems.

This presentation explains how allergy and intolerance information is represented in digital documents that conform to the HL7 Consolidated CDA (C-CDA) standard. It also shows how allergy and intolerance information in C-CDA format can be used to generate digital patient-level reports using the HL7 Quality Reporting Document Architecture (QRDA) standard—a standard used to supply data for the computation of performance toward quality measures. Participants will gain technical insight needed to understand and guide the adoption of systems that create and consume allergy and intolerance data, and measure its availability in EHR systems.

HL7 Clinical Document architecture, Release 2 (CDA R2) specifies a standard way to create digital documents for use in healthcare. CDA R2 was released in 2005. Worldwide, it is the most widely adopted application HL7 V3[8]. HL7 V3 applications all utilize the V3 Reference Information Model (RIM). The RIM is a data model which acts as a foundation upon which to build other models and standards so that implementations will produce data that is more interoperable. CDA documents offer a form of information exchange used in Meaningful Use. The presentation provides a rudimentary overview of the CDA R2 standard and familiarizes participants with the core principles and key characteristics of this standard.

Consolidated CDA Release 1.1 (C-CDA R1.1) is a set of 9 document templates and the associated section and entry templates used in the documents. The templates harmonize work done across HL7, HITSP[9], and IHE[10] to clarify previous designs that included inconsistencies. C-CDA provides a single set of templates which replace the prior work and support greater interoperability. Meaningful Use Stage 2 requires CDA documents conforming to the templates established by C-CDA R1.1. Several templates are included for representing allergy and intolerance information. The presentation reviews the available templates in C-CDA R1.1 for allergy and intolerance and changes proposed for C-CDA R2.0. Participants will see how current modeling compares to the actual templates defined in C-CDA and will discuss options for addressing gaps. A comparison of the V3 and FHIR templates versus the C-CDA models will also be included.

A Quality Reporting Document Architecture (QRDA) document is a special type of CDA document. QRDA is a document format that provides a standard structure with which to report quality measure data to organizations that will analyze and interpret the data. This standard includes a specific QRDA Category I DSTU designed to carry data based on Meaningful Use Stage 2 quality measures expressed in HQMF format. Health Quality Measure Form (HQMF) is a standard developed to express quality measures as "e-Measures". The presentation will familiarize participants with the 3 types of QRDA documents and describe their use in computing measures relevant to assessing available allergy and intolerance information.

Discussion Questions:

1. How are audience members currently recording allergy and intolerance data in their EHR's?
2. How do you define the term "allergy" when it is used as the header of an allergy list?
3. How should allergy and intolerance lists represent non-drug substances?
4. What are the preferred terminology bindings?
5. How do participants send allergy and intolerance data from provider to provider, or to the patient?
6. What enhancements should be made to these HL7 standards?

Disclaimer: All participants have agreed to take part in this panel.

References

1. Health Level 7 (HL7) http://www.hl7.org Ann Arbor, MI
2. CMS EHR Incentives for Meaningful Use http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/Stage_2.html Issued. Accessed March 11, 2014
3. HL7 Publications – Clinical Domains: http://www.hl7.org/implement/standards/product_brief.cfm?product_id=308 Accessed March 12, 2014
4. International Health Terminology Standards Organization (IHTSDO) http://www.ihtsdo.org/ Accessed March 13, 2014
5. Office if the National Coordinator for Health IT (ONC) http://www.healthit.gov/ Accessed March 13, 2014
6. HL7 FHIR (DSTU Version) http://www.hl7.org/implement/standards/FHIR-Develop/summary.html Accessed March 13, 2014
7. Introduction to HL7 Standards: https://www.hl7.org/implement/standards/index.cfm?ref=quicklinks Accessed March 12, 2014
8. Benson, T., Principles of Health Interoperability HL7 and SNOMED, Springer, New York, 2010.
9. Healthcare Information Technology Standards Panel (HITSP) http://www.hitsp.org/ Accessed March 13, 2014
10. Integrating the Healthcare Enterprise (IHE) http://www.ihe.net/ Accessed March 13, 2014

There is Nothing as Practical as a Good Theory: Building the PCORNet Clinical Data Research Network

Charles Borromeo, MS[1], Bari Dzomba, MS[2], Mark G. Weiner, MD[3], Harold Lehman, MD[4]
[1]University of Pittsburgh School of Medicine, [2]Penn State Hershey College of Medicine
[3]Temple University School of Medicine, [4]Johns Hopkins University School of Medicine

Abstract : The goal of the Patient-Centered Outcomes Research Institute's (PCORI) National Patient-Centered Clinical Data Research Network (CDRN) Program is to improve the nation's capacity to conduct comparative effectiveness research (CER) efficiently, by creating a large, highly representative electronic data infrastructure for conducting clinical outcomes research. The creation of a sustainable network requires investment in information technology, adherence to data standards, development of sensible regulatory processes, and alignment with the clinical operations of the participating health systems. This Panel represents the members of the PaTH CDRN. Each member will describe their institution's baseline readiness and their development since the award. We will also describe key issues and insights we've experienced in creating our CDRN including choosing and using common data models, practical issues in applying data standards, and implementing inter-institutional query tools like i2b2, and addressing the regulatory and research informatics issues inherent to a multi-institutional research network

Learning Objectives :
Through participation in this panel, audience members will be better equipped to participate in a CDRN through an understanding of :

1. The impact of decisions regarding common data models on the types of research questions that can be addressed
2. Practical issues in mapping medication, diagnostic, and procedural terminologies to national standards
3. Implementation and use of i2b2 and PopMedNet for sharing queries and returning results across a network.
4. Regulatory and Research Informatics issues and solutions related to the development and application of clinical data research networks

Fundamental to the notion of a learning healthcare system is the ability to integrate comprehensive clinical data to continually generate and test evidence the helps support more informed decisions on the parts of patients and providers. To ensure generalizability of the findings, the clinical data needs to come from a broad range of providers, spanning different geography and covering patients of all ages, genders, races and socioeconomic status. The purpose of the PCORI CDRN program is to facilitate evidence generation through creation of a set of individual CDRNs that individually focus on a core set of disease areas, but can collectively work together to provide insight on a national scale into burden of disease, its management, and comparative effectiveness of therapies.

The PaTH CDRN is a collaboration between University of Pittsburgh/UPMC, Penn State, Temple University and the Johns Hopkins University, and their associated Health Systems. The collaborating institutions all have a history of clinical research experience and use of electronic health records, though variable experience with clinical research data infrastructure. Our success in developing the building blocks of a Clinical Data Research Network reflects our pragmatic approach to challenging issues in data integration, data standards, and implementation of networking, hardware and applications. Through a combination of didactic presentations and interactive discussion, this panel will help audience members understand the technical, administrative and financial issues in developing or participating in a CDRN.

University of Pittsburgh
Creating a *de novo* clinical research network presents a myriad of technical and logistical issues. Identifying appropriate software, developing a project management process, data ownership, and navigating the IRB issues are all required before the inter-site network can start to take shape. Complicating matters more, the PaTH Network faced a number of external networking requirements as a member of the PCORnet ecosystem. In response to these issues our network standardized on the i2b2 software platform and exchanged information across a VPN using SHRINE. We provide access to the PaTH data through the University of Pittsburgh's Comparative Effectiveness

Research Core Data Center (CERC-DC). The CERC-DC provides PaTH researchers with secure data storage, high-throughput computing allowing direct data analysis in a secure environment. The PCORnet data queries are conducted within the sites and the results are sent to the CERC-DC. Through this approach, we met the expectations of both our internal PaTH researchers and the needs outlined by the PCORI leadership. This discussion will illustrate the technical decisions and challenges faced when building a new research network.

Charles Borromeo, MS is a programmer at the Department of Biomedical Informatics (DBMI) at the University of Pittsburgh. He has led several academic development projects including FaceBase and the Biositemaps Resource Discovery System. Prior to his work at the University of Pittsburgh, he has spent over fifteen years working in informatics at several organizations like Pfizer and the Immune Tolerance Network. He holds a Master of Science degree in Biomedical Informatics from the University of Pittsburgh.

Penn State University
How does Big Data meet the Electronic Medical Record (EMR)? According to the American Society for Information Science and Technology, biological and health-based data are complicated, and require novel ways to acquire, process and disseminate the data efficiently. There can be barriers to successful implementations that need to be overcome before these data can be effectively and routinely managed. Healthcare research involving Big Data can pose a challenging set of problems that require the synergistic insights of cross functional teams, which include clinical, research, and information technology.

We understand the complex nature of acquiring, processing, and maintaining the health-based data from our EMR, billing, and specialty applications. Our team was ready to look at different data standards, understand the advantages and disadvantages of each, and be ready with contingency plans to overcome any risk associated with choosing one standard over another. As part of this collaborative effort, Penn State University had a unique challenge, in that our EMR system was different than all of our partner EMR systems. We successfully managed to overcome what could have been a barrier, and work through our own internal processes that were unique to us at a technology and organizational cultural level. From a commonality perspective, mapping the terminologies was a process the entire collaboration needed to do in sync with each other, and Penn State was able to make the modifications needed where we were unique, such as with the Cerner to NDF-RT to RxNorm mapping method. This discussion will focus on data standards and methods of mapping terminologies.

Bari Dzomba, Research Data Warehouse Program Manager at Pennsylvania State University College of Medicine, and Adjunct Faculty Pennsylvania State University Berks Division of Engineering, Business and Computing, has been in the Information Technology field for twenty years, and has twenty five years in the fields of education, healthcare, and information. Bari is currently the principal investigator on a mixed methods research study using a health-based social media dataset focusing on predicting leadership in online health and wellness communities as part of the dissertation requirements for her PhD at Alvernia University. She hold a Master of Science in Information Science from Pennsylvania State University Great Valley, and has held various leadership positions in the fields of healthcare, health insurance, and education.

Temple University
Agreement on a common data model (CDM) is often considered fundamental to the functioning of a research network. Consistent formatting of data across network members ensures that queries designed at one institution will run, at least syntactically, at other institutions. However, even with application of standardized vocabularies, there can be nuances in the recording of data that make interpretation and integration of query results from different institutions more challenging than it may seem. Furthermore, since institutions may have existing data warehouses in different formats, there are logistical issues in maintaining multiple versions of the data in different models. Lastly, as with many information standards, there are many CDMs to choose from, each one with purported advantages and disadvantages over others. This discussion will highlight several options in existing common data models, demonstrate their similarities and differences and provide a rationale for making an optimal decision about which one(s) to choose.

Mark Weiner, MD is Professor of Clinical Sciences, Assistant Dean for Informatics at the Temple University School of Medicine and the Chief Medical Information Officer for the Temple University Health System. He has a track record of work in clinical research databases through his earlier roles in the development of the PICARD

(Pennsylvania Integrated Clinical and Administrative Research Database) system at the University of Pennsylvania that integrates clinical and administrative databases in support of the clinical research enterprise; co-Chief of the Biostatistics and Informatics Core of the VA Center for Health Equity Research and Promotion,; and Co-chair of the Data Core of the FDA mini-Sentinel Initiative.

Johns Hopkins University
The recurrent question that arises in any multi-system work is, how do we maintain the meaning of the data as it is mapped and merged with data from other sources? We take the perspective that EHR data is bad research data, so we must start with understanding the sharing of research data. A recent IOM report lays out many of the issues, half of them involving protecting the research participant (in our case, patients in our centers) and the other half, data issues. CDISC and other organizations have been concerned with transmitting data with fidelity, while studies have shown that even mapping to standards can lead to incompatible results.

We addressed these issues in several ways. We will discuss how our process for defining common data elements balanced available metadata standards against the practicalities of extracting, storing, and sharing data of similar natures but different details, and balanced maintaining meta data, while not burgeoning the system either. All this, within an i2b2 framework.

Dr. Harold P. Lehmann, Professor of Health Sciences Informatics, Pediatrics, and Health Policy and Management, is a board-certified general pediatrician from Columbia University and Babies Hospital, with general pediatric fellowship training from Johns Hopkins and doctoral informatics training from Stanford. His research concerns evidence-based medicine (EBM), ranging from authoring reports to researching novel methods of delivering research results to opinion leaders and practitioners. He has served as a methodologist on a number of professional-society guidelines and has demonstrated the use of advanced decision-analytic methods in guideline development. His current work focuses on the informatics infrastructure of research, including research informatics support, ontologies for human studies and for appropriate inference from electronic health records, on evidence resources for community health workers, and on human (social) services informatics. In addition, he leads the Johns Hopkins efforts in informatics training across all three schools of health sciences, and is the faculty lead for the Informatics Core for the Institute for Clinical and Translational Research at Johns Hopkins. He has served as an Associate Editor for the Journal of the American Medical Informatics Association since 2011. He was elected a Fellow of the American College of Medical Informatics in 2006.

Discussion Questions :

We are interested in participating in a CDRN. What are the key justifications, assurances and processes you developed to gain the essential support of organizational leadership and Institutional Review Boards for CDRN participation?

The Meaningful Use 2 standards incentivize hospitals to standardize on terminologies like SNOMED, LOINC, and RxNORM. Conversely the current Medicare standard billing procedure uses NDC and ICD9 codes. Based on your experience, how should CDRNs reconcile multiple standards and what role can informatics play to help bridge this gap?

There are a number of PCORI CDRNs and related networks for patients called Patient-Powered Research Networks (PPRN). How do you ensure the ability to link your CDRN with these other centers?

All participants have agreed to take part in the panel.

AMIA POI-WG and Eval-WG sponsored panel

Imaginary and real costs of implementing HIT

Panelists

Mark Dente, MD[1], Andrea Gelzer, MD[2], MS, FACP, Ross Koppel, PhD, FACMI[3], Jos Aarts, PhD, FACMI[4]

Moderator

Catherine K. Craven, MLS, MA[5]

[1]Department of Veterans Affairs and University of Utah, Salt Lake City, UT; [2]AmeriHealth Caritas, Philadelphia, PA; [3]University of Pennsylvania, Philadelphia, PA; [4]Erasmus University Rotterdam, The Netherlands; [5]University of Missouri, Columbia, MO

Abstract

The costs of implementing HIT have always been shrouded in clouds. Cost estimates were always a kind of a back of the envelope calculations, with apparent clear figures for contracting such as license and consulting fees and hardware purchase and very unclear figures for the costs of involving the organization, such as releasing health care professionals from their day-to-day duties, production loss, etc. This panel aims to discern myth and reality, to summarize what little is known, to present hands-on experience to deal with costs and propose a multidisciplinary research agenda, involving health informaticists, health care professionals, social scientists and economists.

Introduction

The costs of implementing HIT have always been shrouded in clouds. Even though many remain optimistic about HIT's ultimate value, a recent publication cites a study that despite a "bribe of nearly $27 billion to digitize patient records, nearly 70% of physicians say electronic health record systems have not been worth it [1]." Poor functionality and high costs are quoted as the main reasons. Cost overruns are cited as a major reason for failed implementations [2]. A Gartner study reports that the actual lifetime cost of a system is estimated at 6.7 times the initial costs and consists of the components of enhancing the application (adding new features and new business needs), maintaining the application (bug fixes and software adaptation) and keeping the application operational (help desk, upgrading of computer systems and expanding storage capacity). On the other hand studies show that professionals spend more time managing their data on a computer system. It means that time spent at computer cannot be spent on other activities, such as caring for patients, a phenomenon that economists have termed the opportunity cost of choice. Poor HIT usability is blamed, but also the healthcare system requires professionals to account for their activities. The AMIA implementation discussion list testifies to the frustration of health care professionals spending so much time on what they call 'busy work.' In this way implementation of HIT can add to the cost that is not mentioned in the Gartner report. Proponents argue that the benefits lie in the increased quality, safety, efficacy and efficiency of health care delivery, but the verdict about the influence of HIT is still out [3]. The question however remains who are bearing the costs of HIT implementation and to whom accrue the benefits [4]. This panel seeks to discuss the issue, to separate myth and reality, to articulate the insights of administrators, who are increasingly cognizant of theses issues, and propose an agenda, how costs and benefits can be studied. The discussion will include following questions.

- What are the real and hidden costs of HIT implementation?
- Should contracting HIT be seen as purchasing a product or service?
- What additional costs can be identified in a contract?
- As systems mature, what is the role of performance guarantees?
- How does innovation related to HIT enter into the cost equation?

- What are the costs of poor usability?
- What benefits do the purchasers expect?
- To whom should cost and benefit accrue?

Panel members

- **Mark Dente, MD** is Deputy Chief Medical Informatics Officer at the Veterans Health Administration for Community IT Engagement and Staff Specialist at the Department of Biomedical Informatics of the University of Utah. Previously he was Chief Medical Officer for GE Healthcare Information Technology. He has been active in many areas, including interoperability, business development and intellectual property. He currently has a leadership role in the VA's VistA evolution. In Utah he works with faculty and students to innovate research and education.
- **Andrea Gelzer, MD**, **MS, FACP** is the Senior Vice President and Corporate Chief Medical Officer for AmeriHealth Caritas. She is responsible for medical management strategy, medical policy development, quality management, corporate provider network strategy and medical informatics for all AmeriHealth health plans.
- **Ross Koppel, PhD, FACMI** is a professor of sociology at the University of Pennsylvania. His field of expertise is sociology of work and research methods. In the former capacity he has studied extensively the impact of health information technology on healthcare professionals and addressed problems of HIT usability and safety. In his studies he found that doctors and nurses were not only frustrated by poor functioning IT, but spend a lot of time on entering data for clerical purposes. And that is adding to the costs of implementation.
- **Jos Aarts, PhD, FACMI**, Chair, POI WG, is assistant professor of biomedical informatics at Erasmus University Rotterdam. He has studied the implementation of HIT extensively and found that cost overruns and delays are quite common. Though sociotechnical perspectives inform his research, he thinks that the economics of health informatics is a research area that needs to be developed further.
- **Catherine K. Craven, MLS, MA (Moderator)**, Chair, EVAL WG, is a doctoral candidate in Health (Clinical) Informatics at the University of Missouri, Columbia, Missouri, USA. Her research covers HIT implementation processes in small, rural Critical Access Hospitals; technology impact on ICU clinical team communication; and she is a team member of the University of Missouri's CMS Health Care Innovations Challenge award program to leverage HIT to improve primary care coordination and patient engagement for better quality, care and cost outcomes.

Brief description of panelist presentations

Dr. Dente made the transition from a corporate vendor to the Veterans Health Administration that developed its own HIT software. He will address the business model of a vendor and compare with the HIT open source approach that is dominant in the Veterans Health Administration. Especially he will address the question how innovation enters into the cost equation [5].

Dr. Gelzer has implemented a population-based care management platform across her organization. As a payer, her organization is often called upon to financially support and play an important role in setting financial conditions for HIT adoption. The implementation costs to the payer to deliver population based clinical information across the care delivery continuum must also be considered.. AmeriHealth Caritas is involved in supporting Health Information Exchanges (HIEs). Her presentation will center on the role payers play, who accrues the costs and benefits of HIT and the expectation of a return on investment (ROI).

Professor Koppel's portion focused on implementation costs that are often unrecognized by hospitals and doctors because of complex bookkeeping (a polite term when referring to hospitals), because of the complexity of payment schedules to vendors, and, primarily because most of the costs are obscured in everyday operations of clinicians and the IT personnel. In addition, it is usually not in the interest of the C-suite leaders to display the full cost of implementation to either boards of directors or many staff. Regarding a monumental purchase—of often many hundreds of millions of dollars in software; and 3 to 5 times that in implementation costs—few want their purchase decisions to be show to be unwise or of uncertain ROI.

Dr. Aarts will pull together the insights of the panelists and will present a model developed by one of his students to determine the integral costs of implementation and validated in a case study of an EHR implementation in an outpatient clinic in Montréal, Quebec [6]. Her model identifies costs associated with developing organizational

capabilities, the pre-implementation phase, the implementation of a functional EHR, its optimization and finally post-implementation and operations.

Learning objectives

After participating in the session, the attendant should be able to:

- Formulate an approach to identify cost elements and cost allocations of HIT implementation in a healthcare organization.
- Evaluate cost elements of HIT implementation.
- Articulate a research agenda for biomedical informatics related to HIT implementation cost and benefits realization.

Conflict of interest

The participants have no conflict of interest to declare.

References

[1] Verdon DR. EHRs: the real story -- Why a national outcry from physicians will shake the health information technology sector. Med Econ. 2014 February 10;91(3):18-20, 6-7.

[2] Aarts J, Doorewaard H, Berg M. Understanding implementation: the case of a computerized physician order entry system in a large Dutch university medical center. J Am Med Inform Assoc. 2004 May-Jun;11(3):207-16.

[3] Landrigan CP, Parry GJ, Bones CB, Hackbarth AD, Goldmann DA, Sharek PJ. Temporal trends in rates of patient harm resulting from medical care. N Engl J Med. 2010 Nov 25;363(22):2124-34.

[4] Jones SS, Heaton PS, Rudin RS, Schneider EC. Unraveling the IT productivity paradox - lessons for health care. N Engl J Med. 2012 Jun 14;366(24):2243-5.

[5] Saleem JJ, Flanagan ME, Wilck NR, Demetriades J, Doebbeling BN. The next-generation electronic health record: perspectives of key leaders from the US Department of Veterans Affairs. J Am Med Inform Assoc. 2013 Jun;20(e1):e175-7.

[6] Creton de Limerville L. What are the cost components associated with the implementation of electronic medical records in health care organizations? MSc Thesis, Erasmus University Rotterdam, June 2014.

Public and Global Health Informatics Year in Review

Brian E. Dixon, MPA, PhD[a,b,c], Jamie Pina, PhD, MSPH[d,e], Janise Richards, PhD, MPH, MS[f], Hadi Kharrazi, MD[g], Anne Turner, PhD, MLIS, MPH[h]

[a] Indiana University School of Informatics and Computing, Department of BioHealth Informatics, Indianapolis, IN
[b] Regenstrief Institute, Center for Biomedical Informatics, Indianapolis, IN
[c] Center for Health Information and Communication, Department of Veterans Affairs, Veterans Health Administration, Health Services Research and Development Service CIN 13-416, Indianapolis, IN
[d] RTI International, Waltham, MA
[e] Emory University, Atlanta, GA
[f] Centers for Disease Control and Prevention, Center for Global Health, Atlanta GA
[g] Department of Health Policy and Management, Johns Hopkins School of Public Health, Baltimore, MD
[h] Department of Biomedical Informatics and Medical Education, University of Washington School of Public Health, Seattle, WA

Abstract

The disciplines of public health and global health informatics are rapidly expanding within the field of biomedical informatics. Increased attention and activity by the Centers for Disease Control and Prevention in the U.S. as well as health ministries, the World Health Organization, and non-governmental organizations are generating new knowledge and lessons regarding the development, implementation, and use of information systems in health care delivery around the globe. Thus, a growing body of literature now contains important insights and lessons from international informatics activities, stimulating the need to synthesize the knowledge for the field. In this panel, a review of recent literature in the areas of public health and global health informatics will be presented. Key articles revealing trends, methods, and lessons will be summarized to bring attendees up-to-date on the use of informatics in low resource settings.

Introduction

At the AMIA 2013 Annual Symposium, Drs. Brian Dixon and Jamie Pina presented an inaugural AMIA panel presentation focused on summarizing recent literature in Public Health Informatics (PHI) and Global Health Informatics (GHI). The presentation was well-attended, and it broadened the topics covered by "Year in Review" sessions at the AMIA 2013 Annual Symposium. We seek to present a similar panel at the 2014 symposium. Our goal is to provide a concise overview of landmark papers in PHI and GHI for the year prior to the 2014 symposium using the methods described below. Furthermore, like last year, we will provide the output of our efforts as working documents available on the MyAMIA community site enabling attendees, as well as those who cannot physically attend, to leverage our work throughout the year.

Overview

The fields of public and global health informatics are expanding. In 2012, the PHI Fellowship Program administered by the U.S. Centers for Disease Control and Prevention (CDC) became a Department of Labor Registered Apprenticeship. PHI was also one of the six sub-disciplines of informatics specified in the ONC certificate programs funded between 2010 and 2013. As a result, publications in PHI have dramatically increased. This is evidenced by an increase in *Journal of the American Medical Informatics Association (JAMIA)* PHI articles, as well as the emergence of a journal dedicated completed to the study and practice of PHI, the *Online Journal of Public Health Informatics* (www.ojphi.org). Similar increases in programs and publications have been observed in the field of GHI. For example, Fogerty International fellowships now bring scholars from low-income counties to the U.S. for biomedical informatics training. The CDC also promotes a Global Public Health Informatics Program (http://www.cdc.gov/globalhealth/dphswd/gphip/what/objectives.htm) to build capacity and sustainability of health information systems world-wide. The impact factor and quality of publications in the *International Journal of*

Medical Informatics (IJMI) has also increased, leading to a significantly larger number of important papers published by non-U.S. scholars in informatics (1).

Given recent and continuing advances in PHI and GHI, an annual review of landmark publications in these sub-disciplines within the field of biomedical informatics is appropriate for the AMIA Annual Symposium. Two panelists (BED and JP) will each present a concise bibliography of recent landmark publications in either PHI or GHI. The session will be divided equally between PHI and GHI. A summary examining the convergence, synergies and possible future cooperation will be presented by the third panelist (JR). Time will be reserved at the end for discussion, question, and comments with the audience.

Methods

The methods used to select the literature in each part of the session are adapted from the work of previous AMIA Year in Review sessions. Our search methodology to identify potentially high impact literature in PHI and GHI is outlined below. The methods will be executed using MEDLINE and selected resources not currently indexed such as the conference proceedings of the International Society for Disease Surveillance. Recommendations will also be solicited from the AMIA PHI and GHI working group membership.

Candidate articles will be reviewed on a quarterly basis by a committee of experts in PHI and GHI, including Drs. Dixon, Turner, Kharrazi, Pina, and Richards, the Review's core team members. Additional input from the AMIA PHI and GHI working groups will be solicited. Selected articles from the list of candidates will be determined through consensus discussions on quarterly conference calls. Final lists of selected articles will be made available online via the MyAMIA site for reference following the symposium.

Public Health Informatics

To identify articles in PHI, we will use the following query (customized for MEDLINE):

"Public Health Informatics"[mh] OR ("exp Public Health"[mh] and "exp Informatics"[mh]) OR ("public health"[mp] and "informatics"[mp])

This query, when limited to publications between 1/1/2013 and 12/31/2013 in MEDLINE, returned 219 results. This is a reasonable set for the committee to review and discuss within a short timeframe.

In the presentation, Dr. Dixon will present the final selection of papers from the set of PHI candidate papers. Papers may be further categorized into themes, such as disease surveillance, population health, public health decision support, or chronic disease management depending on the list of candidate papers and the perspectives of the review committee. When reviewing the papers, Dr. Dixon will briefly describe their aim, methods, and key findings. Paper results will also be discussed in the context of larger PHI developments, such as the release (or study) of a specific system (e.g., Biosense 2.0) or change in public health infrastructure (e.g., re-organization of the CDC's division responsible for PHI). Context will help audience members connect specific study findings with larger trends within the biomedical informatics field of study and practice.

Global Health Informatics

To identify articles in GHI, we will use the following query (customized for MEDLINE):

("exp Informatics"[mh] OR "exp Telemedicine"[mh] OR "information system"[mp]) AND ("Developing Countries"[mh] OR global OR ministry OR "low resource" or "resource-limited"[mp])

This query, when limited to publications between 1/1/2013 and 3/13/2014 in MEDLINE, returned 345 results and was a reasonable set for the committee to review, analyze and discuss. Due a slightly shorter time-frame for publication of GHI-related research and activities, we plan to examine the "grey literature" and other non-traditional research sources, to acquire the depth and breadth of informatics activities globally in the past year.

In the presentation, Dr, Pina will present the final selection of papers from the set of GHI candidate papers. Papers may be further categorized into themes, such as mHealth, infrastructure, evaluation, capacity development, system implementation, etc. depending on candidate papers located and the review committee's assessments. The reviewers will briefly describe the paper's aim, methods, and key findings. Findings from the trends in the papers reviewed will also be discussed in context of global health or global informatics policy related activities, such as World Health Organization (WHO) guidelines, the US Congress reauthorization of the President's Emergency Plan For AIDs Relief (PEPFAR) and the increasing number of individual country eHealth Strategy Policy documents.

Following the PHI and GHI presentations, Dr. Richards will provide a summary to examine findings to provide context for possible PHI and GHI priority activities for the coming year. The session will end with discussion. Our experience from last year's presentation demonstrated that these related but unique sub-fields within biomedical informatics generate many interesting questions and comments from the audience and facilitate a dialogue between AMIA's international and domestic membership.

Discussion Questions

We will draw from the following discussion questions to engage audience participation:

- We found several interesting trends in knowledge generation within PHI & GHI over the past year. Which trend(s) were surprising to you?
- Projecting on what we have found, what do you think will be the biggest trend for 2015?
- What trends do you know of in GHI and PHI for the past year that may not have made it into the literature, yet?
- What initiatives, organizations and/or funding may exist to drive improvements in PHI & GHI research methods as well as dissemination and translation into practice?
- How can AMIA, especially the PHI & GHI working groups, better support the generation of new knowledge for research and practice in our respective disciplines?

References

1. Talmon J, Safran C, de Fatima Marin H, Degoulet P, Geissbuhler A. IJMI impact. International Journal of Medical Informatics. 2010;79(11):733-5.

Informatics without Borders: International Outreach of US-based Training Programs

Biomedical research and health services research have been increasingly relying on international collaboration. Team science has expanded beyond regional and national collaborations into an ecosystem of collaborators that are united by common goals through an electronic infrastructure for research. It has thus become important that all collaborators have adequate human and material resources to participate in international research networks. We will discuss how some biomedical informatics training programs based in the USA have been partnering with institutions from around the world to build capacity in informatics and have been empowering researchers with infrastructure and tools to participate in globalized biomedical research. We will cover (a) cultural and economic barriers for implementation of certain programs, (b) programs we developed to respond to the short- and medium-term needs of our international collaborators, (c) how we adapted existing research-enabling tools to facilitate research in developed and developing countries, and (d) how we adapted distance-learning and related technologies to facilitate this work. We will also discuss the importance of embedding international students in our training programs in the US, and the advantages that this type of diversity brings in terms of promoting broader cultural understanding that benefits all our faculty, students, and staff.

Panelists

1. Cynthia S. Gadd, PhD, MBA, MS, Professor and Vice-Chair for Educational Affairs, Department of Biomedical Informatics, School of Medicine, Vanderbilt University, Nashville, Tennessee

Title: Lessons Learned From an Indo-US Collaboration in Clinical Research Informatics Training

Vanderbilt University Department of Biomedical Informatics (VU-DBMI) and collaborators at the National Institute for Epidemiology (NIE) in Chennai, India envisioned the development of a clinical research informatics (CRI) program at NIE to include: (1) launching of a CRI certificate program at NIE utilizing best-in-class resources, including VU-DBMI and Indian faculty, with increasing responsibility for teaching by NIE scientists as their expertise grows; (2) establishment of an Indian CRI-REDCap Consortium Center to promote and support the use of next generation CRI in the service of clinical and translational research in India; and (3) development, through initial intense training and continuing education, of local CRI leaders at NIE, to fulfill the CRI-REDCap Consortium Center's education and research consultation missions. We will report and discuss the opportunities and challenges presented by this collaboration.

2. William Hersh, MD, FACP, FACMI, Professor and Chair, Department of Medical Informatics & Clinical Epidemiology, School of Medicine, Oregon Health & Science University, Portland, Oregon

Title: Adapting US-Centric Informatics Training to Expand Capacity in Low and Middle Income Countries

As emerging economies move from developing to developed status, the priorities for basic health care and public health take on additional burdens of behavioral/lifestyle aspects of health and chronic disease management. This management can be augmented by informatics. The

overall goal of our collaboration is to develop an informatics training program in global health that combines the existing strengths of the Department of Medical Informatics of Hospital Italiano de Buenos Aires (HIBA), Argentina and its established collaboration with the Department of Medical Informatics & Clinical Epidemiology (DMICE) of Oregon Health & Science University (OHSU) to develop a new focus in clinical and translational research informatics. We build upon the mutual strengths of both programs in the use of distance learning technologies to train researchers and informaticians, from single courses to entire fellowships. Attendees of the courses have gained knowledge to improve their research capabilities, while fellows have successfully started research careers in informatics.

3. Rebecca Crowley, MD, MS, Director, Biomedical Informatics Graduate Training Program, Associate Professor of Biomedical Informatics, University of Pittsburgh, Pittsburgh, Pennsylvania

Title: Enhancing Research and Informatics Capacity for Health Information in Colombia (ENRICH)

The collaboration between Javeriana University in Bogota, Colombia and the University of Pittsburgh has enabled the development of (1) on-site educational programs in Colombia and (2) certificate and master's training for Colombian clinicians, scientists and engineers in the Biomedical Informatics and Clinical Research training programs at the University of Pittsburgh. On-site activities have ranged from large multi-day courses, including a general review of Biomedical Informatics taught by a group of Pittsburgh and Javeriana faculty, to smaller focused workshops, including a short course in telemedicine. Colombian trainees in our program have worked on a variety of research projects in global health informatics, biosurveillance, and ontology development. A major focus of our partnership has been to help develop informatics capacity that builds on both technical and clinical expertise in Colombia. Cross fertilization with other global health informatics initiatives has provided unique opportunities for both Colombian trainees and other trainees of our training program.

4. Lucila Ohno-Machado, MD, MBA, PhD, Associate Dean for Informatics & Technology, Professor of Medicine and Chief, Division of Biomedical Informatics, School of Medicine, University of California San Diego, La Jolla, CA

Title: Brazil-Mozambique-US Collaboration in Biomedical Informatics and Bioinformatics

In partnership with the Federal University of Sao Paulo in Brazil and the Eduardo Mondlane University in Maputo, Mozambique, the University of California San Diego developed a series of programs to enhance local informatics training in areas in which gaps were noted and the application of the knowledge acquired in training was most feasible. In Brazil, a strong program in health informatics was extended to remote regions of the country via a tele-education program. In Mozambique, a series of short courses and symposia were developed to raise awareness of informatics in healthcare and biology. The courses in Mozambique were delivered in Portuguese by Brazilian collaborators. The symposia were offered with simultaneous translation into Portuguese. We will discuss lessons learned from these unique experiences as well as from training international students and postdoctoral fellows at UCSD.

Safety-Enhanced Design as a Meaningful Use Objective: Evaluating and Advancing the Usability of Electronic Health Records

Jan Horsky, PhD[1,3], Blaine Y. Takesue, MD[2,4], Rebecca Farr[5], Janet Campbell[6]

[1]Brigham & Women's Hospital, Boston, MA; [2]Regenstrief Institute, Indianapolis, IN, [3]Harvard Medical School, Boston, MA, [4]Indiana University School of Medicine, Indianapolis, IN, [5]Intermountain Healthcare, Salt Lake City, UT, [6]Epic Systems Corporation, Verona, WI

Abstract

Effective, safe and routine use of HIT by clinicians is predicated on the availability of well-designed systems that have excellent usability characteristics and can be integrated into common clinical workflows. The ONC has required, for the first time, that institutions and vendors developing EHRs submit results of summative usability evaluations as part of their application for Meaningful Use Stage 2 certification. The intent was to let developers show evidence of usability of their product so consumers could make informed purchase decisions. The expected content of each usability report was specified by the NIST in the Common Industry Format Template. However, the methods and extent of each evaluation study likely varied at each institution certifying their product. Panelists will discuss test studies, results and lessons learned for three home-grown EHRs – The Medical Gopher, HELP, and the Longitudinal Medical Records systems, and present a vendor perspective from Epic Systems.

Introduction

Effective use of health information technology (HIT) and continuing progress in its adoption by clinicians are the objectives of national policy initiatives in many countries.[1] The rationale driving the implementation of electronic health record systems (EHR) with advanced decision support is their potential to improve and monitor care quality and increase patient safety.[2] There is an emerging consensus among leaders in the industry, academia and government that a sustained positive effect of HIT can only be achieved by using systems that are specifically designed for the complex and dynamic healthcare environment, provide cognitive and decision support, have high standards of usability, advanced design of the human interface and can be integrated well into clinical workflows.[3]

The Office of the National Coordinator for HIT (ONC) has underscored the significance of human interface design as a core attribute of high-performing and safe information systems in Stage 2 of its Meaningful Use requirements for 2014.[4] It has added a requirement to report the results of summative usability evaluation and an attestation that development was done in accordance with a user-centered design process in order to receive ONC certification and CMS financial incentives. The intent was to help vendors and developers at academic and public institutions demonstrate the evidence of their product's usability in a format allowing both independent evaluation and comparison.[5] Test results are made public on the ONC website to allow consumers and procurers at large institutions to review and assess basic usability characteristics of systems they may consider purchasing. Developers may also use the insights and lessons learned from the testing to further refine the design of their systems and to focus on problem areas that may not have been previously identified.

The format and required content of reported results was specified in the Customized Common Industry Format Template[6] by the National Institute of Standards and Technology (NIST) but methods and extent of evaluations were not and likely varied by system and institution. For example, in the publicly accessible reports, the number of clinicians participating in observational studies varied from 5 to 30 and test settings, scenario content, the interpretation of results such as path deviations, type and severity of errors and task success or failure criteria were markedly different from one report to another and sometimes not described or reported. Evaluators were either hired usability professionals or experts from each institution.

The panelists will summarize and present their methods of evaluation and the results of summative studies they conducted in the process of generating data for submission to their respective certifying bodies. Each presenter will conclude with the lessons learned from the testing and significance of the results for their organization, insights to patterns of use by clinicians that may have not been previously known and what actions can be taken to address limitations, problems or observed workarounds by future redesign of specific products.

Urgency, timeliness and significance of HIT usability evaluation

Usability test results and user-centered design (UCD) attestations were new requirements for Stage 2 this year and the topic is therefore timely and of interest to most organizations, private and commercial, that had to go through this process relatively recently and for the first time. Consumers have now a new source for examining the results of usability evaluations for systems they either own or plan to buy and can compare them across vendors. This practice is recommended and encouraged by an HIMSS usability task force[7] but has been relatively difficult to carry out. The publicly available results are now more likely to encourage prospective buyers to include usability criteria more prominently and with greater confidence as factors in purchasing decisions. Usability is frequently cited by clinicians as a highly desirable characteristic of an EHR, strongly affecting their productivity and time spent documenting care, but by many accounts often lacking in current EHRs. There is considerable research evidence of positive correlation between the quality human-computer interface design and patient safety[8, 9] and that poorly designed HIT may not only slow down clinical work but may increase the risk of medical error.[10]

Panelists

The proposed panel consists of 15-minute presentations by each participant, followed by a moderated discussion.

Jan Horsky, PhD (organizer)

Dr. Horsky leads research in human-computer interaction at Partners HealthCare in Boston and conducts periodic formative and summative usability evaluations of information systems that are developed and maintained internally, primarily the Longitudinal Medical Record (LMR) system. The LMR is used daily by over 5,000 physicians in diverse ambulatory settings ranging from small community practices to hospital-based ambulatory centers. His research is concerned with cognitive aspects of human-computer interaction in dynamic healthcare environments, usability evaluations of HIT, and with the effects of clinical information systems on the quality of care and medical error. Past projects include a cognitive usability study of clinicians interacting with a computer-based provider order entry system (CPOE), and the evaluation of a serious medication dosing error that was in part related to the use of a CPOE system. He led the Safety Enhanced Design effort for the certification of inpatient and outpatient systems at Partners. Dr. Horsky is Instructor in Medicine at Harvard Medical School and adjunct faculty at Usability and User Experience Program at Bentley University in Waltham, MA.

Blaine Y. Takesue, MD

Dr. Takesue is the clinical lead for the re-engineered version of the Regenstrief Gopher CPOE. It is currently deployed throughout the Eskenazi Health system, including its hospital and outpatient clinics and used by more than 800 providers to write over 7,000 notes and orders daily. Dr. Takesue also leads the Gopher Meaningful Use certification process and the Meaningful Use attestation/reporting effort for Eskenazi Health. His areas of interest include improving CPOE user experience and combining theory, technology and infrastructure to transform data into knowledge to address clinical concerns and to improve provider and patient health care experiences. Dr. Takesue is Assistant Professor of Clinical Medicine at the Indiana University School of Medicine.

Rebecca Farr

Ms. Farr has 35 years of healthcare experience, including administrative management, information technology, quality assurance and government regulations and compliance. Her expertise is in requirements and design analysis, along with operationalizing large, highly-integrated projects. Her recent focus has been on facilitating the enhancement of Intermountain Healthcare's clinical systems to meet ONC HITECH Certification requirements and to support the clinical workflows required to meet CMS Meaningful Use regulations. The inpatient clinical system (HELP1) and ambulatory clinical system (HELP2) were tested for certification. HELP was the first hospital information system to collect patient data needed for clinical decision-making and at the same time incorporate a medical knowledge base and inference engine to assist clinicians in decision making.

Janet Campbell

Ms. Campbell is Senior Software Architect at Epic Systems Corporation, an EHR vendor for mid-size and large medical groups, hospitals and integrated healthcare organizations. She has led over the last eleven years development of products such as the Stork Obstetrics module, Lucy personal health record, and MyChart Bedside shared record for hospitalized patients. Ms. Campbell represents Epic in national conversations on interoperability, usability, meaningful use, and patient engagement and is Vice Chair of the Electronic Health Record Association's Clinical Experience Workgroup.

Vimla L. Patel, PhD, DrSc (Moderator)

Dr. Patel is Senior Research Scientist and Director of *Center for Cognitive Studies in Medicine and Public Health* at the New York Academy of Medicine, with adjunct appointments at Columbia and Cornell Universities as Professor of Biomedical Informatics and Public Health, respectively. Trained as a cognitive scientist at McGill University where she served as Professor of Medicine and Psychology, her research focuses on modeling medical decision-making and on cognitive mechanisms underlying human performance in health care. Her recent studies focus on complexity of distributed cognitive systems that underlie critical care decisions, generation of medical errors and on the impact of technology on human cognition for competent performance.

An elected fellow of the Royal Society of Canada (Academy of Social Sciences), the American College of Medical Informatics, and the New York Academy of Medicine, she was a recipient of the annual Swedish "Woman of Science" award in 1999. She received an Honorary Doctor of Science degree from the University of Victoria in 1998, in recognition of her contributions through cognitive studies in the domain of health informatics. She is an associate editor of the *Journal of Biomedical Informatics* and sits on the editorial boards of *Artificial Intelligence in Medicine* and *Advances in Health Science Education*.

Discussion topics and questions

- This panel reviews three home-grown systems and a vendor system. Can methods, lessons learned and results be compared across vendor systems and commercial applications intended for small private clinics?
- Three systems are at institutions of different sizes: two large networks consisting of multiple hospital, clinics and private practices (Partners and Intermountain) and a smaller system consisting of the Eskenazi Hospital and over 20 mostly community based clinics (Regenstrief). How did the evaluations differ in scope, size of the test sample cohort and interpretation of the result? Are they directly comparable?
- How did institutions decide between evaluating systems internally and contracting usability professionals?
- What design changes are planned based on the findings from each study?
- What results were unexpected or surprising? Were any unintended effects observed?
- What new insights did he evaluators learn and how were they reformulated in to redesign projects?
- What was the time and effort required at each institutions to collect data and generate the report?

References

1. Simborg DW, Detmer DE, Berner ES. The wave has finally broken: Now what? Journal of the American Medical Informatics Association. 2013 June 1, 2013;20(e1):e21-e5.
2. Blumenthal D. Launching HITECH. New England Journal of Medicine. 2010 Feb 4;362(5):382-5. PubMed PMID: 20042745. Epub 2010/01/01. eng.
3. Sittig DF, Singh H. Electronic health records and national patient-safety goals. The New England journal of medicine. 2012 Nov 8;367(19):1854-60. PubMed PMID: 23134389. Epub 2012/11/09. eng.
4. Office of the Secretary. Health Information Technology: Standards, Implementation Specifications, and Certification Criteria for Electronic Health Record Technology, 2014 Edition. In: Health And Human Services, editor. March 7 - Proposed rules ed. Washington, D.C.2012. p. 13832-85.
5. Lowry SZ, Quinn MT, Ramaiah M, Schumacher RM, Patterson ES, North R, et al. Technical Evaluation, Testing and Validation of the Usability of Electronic Health Records. 2012 NISTIR 7804.
6. Schumacher RM, Lowry SZ. Customized Common Industry Format Template for Electronic Health Record Usability Testing. Washington, D.C.: National Institute of Standards and Technology, 2010 NISTIR 7742.
7. HIMSS. Selecting an EMR for Your Practice: Evaluating Usability. Healthcare Information and Management Systems Society: EHR Usability Task Force, 2010.
8. Chaffee BW, Zimmerman CR. Developing and implementing clinical decision support for use in a computerized prescriber-order-entry system. Am J Health Syst Pharm. 2010 Mar 1;67(5):391-400. PubMed PMID: 20172991. Epub 2010/02/23. eng.
9. Institute of Medicine. Health IT and Patient Safety: Building Safer Systems for Better Care. Washington, D.C.: The National Academies Press; 2011. 197 p.
10. Horsky J, Kuperman GJ, Patel VL. Comprehensive analysis of a medication dosing error related to CPOE. Journal of the American Medical Informatics Association. 2005 Jul-Aug;12(4):377-82. PubMed PMID: 15802485. English.

The Clinical Quality Framework Initiative: Harmonizing Clinical Decision Support and Clinical Quality Measurement Standards to Enable Interoperable Quality Improvement

Organizer: Kensaku Kawamoto, MD, PhD, MHS

Panelists
Kensaku Kawamoto, MD, PhD, MHS[a]
Marc J. Hadley, PhD[b]
Kate Goodrich, MD, MHS[c]
Jacob Reider, MD[d]

Panelist Affiliations
[a] Department of Biomedical Informatics, University of Utah, Salt Lake City, UT
[b] MITRE Corporation, Boston, MA
[c] Centers for Medicare & Medicaid Services, Washington, D.C.
[d] Office of the National Coordinator for Health IT, Washington, D.C.

Abstract
Clinical decision support (CDS) and electronic clinical quality measurement (eCQM) are closely related and share many common requirements. However, CDS and eCQM standards were developed in parallel and utilize different approaches for representing patient information and computable expression logic. This divergence imposes additional burdens on users of the standards and limits the ability to cross-leverage eCQM resources for CDS, and vice versa. To address this important problem, the Office of the National Coordinator for Health IT (ONC) and the Centers for Medicare & Medicaid Services (CMS) are sponsoring an open, public-private collaborative effort to establish a single, harmonized set of standards for CDS and eCQM. This effort, known as the Clinical Quality Framework (CQF) initiative, is developing and validating common standards for representing patient data, expression logic, and metadata for the purposes of quality improvement. Moreover, CQF is re-factoring existing CDS and eCQM standards to utilize these common components. In this panel, CQF leaders – including the Deputy National Coordinator of ONC and the Director of Quality Measurement at CMS – will provide an overview of CQF, its methodology, and its deliverables. Standards developed and validated through CQF will be available to policy makers for potential inclusion in future EHR certification requirements.

Description
The panel will be organized as follows:

Time	Speaker	Topic
10 min	Reider	ONC perspective on motivation and goals for CQF
10 min	Goodrich	CMS perspective on motivation and goals for CQF
20 min	Hadley	Overview of CQF and its methodology Harmonization of metadata Harmonization of expression logic

Time	Speaker	Topic
20 min	Kawamoto	Harmonization of patient information model Updates to existing CDS and eCQM standards to utilize harmonized components Pilot implementations Summary of current status and future directions
30 min	Reider	Panel discussion with audience

Dr. Reider will serve as the moderator and introduce each of the panel members and their organizations. He will also describe the motivation and goals underlying ONC's sponsorship of the CQF initiative. Dr. Goodrich will then describe the motivation and objectives of the effort from the perspective of CMS. Dr. Hadley will provide an overview of the CQF initiative and its methodology. He will then describe how metadata and expression logic for CDS and eCQM are being harmonized. Dr. Kawamoto will then describe how a common patient information model is being developed for CDS and eCQM, as well as how existing CDS and eCQM standards are being updated to utilize these harmonized component standards. He will then describe pilot implementations of the standards, summarize the current status of CQF, and outline planned future directions. These presentations will be followed by a panel discussion with the audience moderated by Dr. Reider.

Reider: Dr. Reider is the Deputy National Coordinator of ONC. Prior to his leadership role at ONC, Dr. Reider was the Chief Medical Informatics Officer of Allscripts. Dr. Reider will moderate the session and discuss the motivation behind the ONC initiating and sponsoring the CQF initiative.

Goodrich: Dr. Goodrich is the Director of the Quality Measurement and Health Assessment Group at CMS. In this role, Dr. Goodrich oversees the implementation of 12 quality measurement and public reporting programs and partners with CMS components on several other programs. She continues to practice clinical medicine as a hospitalist and Associate Professor of Medicine at George Washington University Hospital. Dr. Goodrich will discuss the motivation and objectives underlying CMS's sponsorship of the CQF initiative.

Hadley: Dr. Hadley is Principal Software Systems Engineer at MITRE Corporation, an expert on eCQM, and co-Initiative Coordinator of the CQF initiative. Dr. Hadley will provide an overview of the CQF initiative and its membership. Dr. Hadley will also describe the systematic methodology of the CQF initiative, which utilizes the ONC's Standards & Interoperability Framework for developing and validating interoperability standards. He will also describe core principles of the CQF initiative, which include collaboration, transparency, and pragmatism. He will then describe the CQF initiative's efforts to develop common standards for metadata and expression logic for CDS and eCQM.

Kawamoto: Dr. Kawamoto is Associate Chief Medical Information Officer, Assistant Professor of Biomedical Informatics, and Director of Knowledge Management and Mobilization at the University of Utah. Along with Dr. Hadley, Dr. Kawamoto is co-Initiative Coordinator of the CQF initiative. Previously, he served as the Initiative Coordinator of the ONC Health eDecisions initiative, which developed the CDS standards being harmonized with eCQM

standards in the CQF initiative. Dr. Kawamoto will describe the CQF initiative's efforts to develop a common patient information model standard for CDS and eCQM, which includes the development of a common logical model known as the Quality Information and Clinical Knowledge (QUICK) model, as well as a physical model based on the HL7 Fast Healthcare Interoperability Resources (FHIR) standard. He will then describe how these common standards for patient information, metadata, and expression logic are being incorporated into relevant CDS and eCQM standards, namely the HL7 CDS Knowledge Artifact Specification, the HL7 Decision Support Service Implementation Guide, and the HL7 Health Quality Measures Format (HQMF). Dr. Kawamoto will also describe experiences with pilot implementations of these standards and conclude with a summary of the current status and future directions of the CQF initiative. Interoperability specifications within the scope of the CQF initiative are planned for balloting in HL7 in September 2014, and the standards developed and validated through CQF will be available to policy makers for potential inclusion in future EHR certification requirements.

Reider: Dr. Reider will lead a moderated discussion with the audience. The objectives of this discussion will be to answer questions from the audience and to obtain feedback from the audience to guide the future directions of the CQF initiative and related efforts. The discussions questions listed below will be used to stimulate this discussion.

Significance of panel topic and anticipated audience

Current and proposed EHR certification requirements associated with the U.S. Meaningful Use program include requirements for divergent CDS and eCQM standards. The CQF initiative is seeking to expeditiously address this important challenge by developing and validating a unified set of standards for CDS and eCQM. Of note, the results of the CQF initiative will be available to federal policy makers for potential inclusion in future EHR certification requirements. Therefore, the panel topic is timely, urgent, needed, and of great significance to the clinical informatics community. The anticipated audience is AMIA attendees interested in CDS, eCQM, and/or Meaningful Use and associated EHR certification criteria, including members from academia, the public sector, and private industry.

Discussion questions

What suggestions do you have for the CQF initiative and related activities, such as EHR certification requirements for CDS and eCQM?

What gaps still exist in the available standards related to CDS and eCQM? How should we address those gaps?

What clinical domain areas can most benefit from standards-based, interoperable CDS and eCQM (e.g., immunizations, chronic disease management, chemotherapy)?

How should CDS and eCQM be leveraged in the short-term and long-term to improve clinical quality and health outcomes?

Participation statement

All proposed panelists have agreed to participate in the panel if the proposal is accepted.

Acknowledgements

The CQF initiative is supported by funding from the ONC and CMS, as well as by voluntary contributions of effort from various members of the CQF community.

Genomic dark matter, druggability and misunderstood targets

Subramani Mani, MBBS, PhD[1,4]

Joshua Swamidass, MD, PhD[2]

Noel Southall, PhD[3]

Tudor Oprea, MD, PhD[1]

[1] Division of Translational Informatics, Department of Internal Medicine, University of New Mexico Health Sciences Center, Albuquerque, NM

[2] Division of Laboratory and Genomic Medicine, Department of Pathology and Immunology, Washington University in St. Louis, St. Louis, MO

[3] National Center for Advancing Translational Sciences (NCATS), Rockville, MD

[4] Panel organizer

Researchers and pharmaceutical companies have typically focused on a small part of the genome-encoded proteome for drug targeting thus shedding light on these regions from a druggability perspective while leaving large tracts of the genome-encoded proteome or exome dark. This panel will introduce the concepts of genomic dark matter, druggability of genes/proteins and misunderstood targets. We will also introduce an in-silico framework for illuminating the druggable genome by discussing the approaches taken for the creation of the Illuminating the Druggable Genome Knowledge Management Center (IDG KMC) and the Target Central Resource Database (TCRD). TCRD will serve as a go to resource for developing a mechanistic understanding of a newly discovered set of biomarkers for a clinical condition from a druggability perspective and this will be illustrated in the panel presentation using the set of ranked biomarkers identified for early detection of neonatal sepsis. In short the proposed panel will discuss effective strategies for shedding light on the dark matter of the druggable genome to identify useful and effective drug targets.

Learning objectives:

1. Understand the concept of druggability
2. To learn how to define drug targets
3. Understand the concept of genomic “dark matter” from a druggability perspective
4. Understand how the genomic dark matter can be “illuminated” from a druggability perspective

From a druggability perspective large tracts of the exome, that is, the proteins encoded by the genes in the exome have been left unexplored and hence “dark”. Based on their review of literature Hopkins et al. report that 399 non-redundant molecular targets have been shown to bind efficaciously with small molecules[1] out of more than 10,000 likely such targets in the human genome based on projections of ligand binding domains[2]. The challenge is to identify the relevant subset of the potential druggable targets that are represented in disease-linked

proteins. Note that approximately ninety percent of the currently marketed drugs interact with the various proteins encoded by the genes represented in the exome[3].

Dr. Joshua Swamidass, MD, PhD: What is the meaning of druggability?

Modern genome-wide technologies—that study all the proteins in the genome—can identify genes mechanistically related to human disease. One hope in these studies is to identify new drug targets, genes that can be pharmaceutically modulated to treat disease. However, only a fraction of genes can be effectively targeted with a drug. In other words, only a fraction of genes are translated to proteins that are considered “druggable.” How is druggability determined and defined? There are, actually, several approaches to answering this question. Here, we will present several common approaches, and examine their implications for interpreting genome-wide analysis and drug discovery.

Dr. Noel Southall, PhD: Misunderstood targets

There is a high failure rate for drugs in clinical development, and an inability to adequately predict whether a drug will be efficacious. To better increase the chances of success in the clinic, one can choose to work solely on clinically validated targets. The other option is to choose to work on novel targets from those classes of proteins that have had success in the past, such as G-protein-coupled receptors (GPCRs), kinases, and ion channels. Will this strategy improve the chances of success? Different classes of targets have different liabilities, especially the problem (or benefit) of polypharmacology which is not easily addressed. The recent practice of developing allosteric compounds to avoid the problems of polypharmacology introduces new unknowns that can also send one right back to the beginning. And there are gaps in understanding how drugs work even after their marketing authorization. Here we consider some details of target validation and try to add another level of sophistication into the target druggability debate.

Dr. Tudor Oprea, MD, PhD: Illuminating the Druggable Genome Knowledge Management Center (IDG KMC) and the Target Central Resource Database (TCRD)

The goal of the Illuminating the Druggable Genome Knowledge Management Center (IDG KMC) is to evaluate, classify and rank all the disease-linked proteins based on their potential as druggable targets. The main focus will be on the four protein superfamilies: G-protein-coupled receptors (GPCRs), nuclear receptors (NRs), ion channels (ICs) and kinases. By integrating data extracted from multiple sources using algorithmic processing, prediction and human curation, the emerging knowledge base will hyperlink and annotate all the relevant proteins. The KMC will link disease, pathway, protein, gene, chemical, bioactivity, drug discovery and clinical information elements from databases, literature, patents and other documents in a "Target Central" Resource Database (TCRD). Based on this rich knowledge representation we will be able to categorize proteins into four classes (Tclin - clinical; Tchem - manipulated by chemicals; Tmacro - manipulated by macromolecules; and Tdark - the genomic "dark matter") and enable efficient and effective target prioritization for new therapeutics as the knowledge base grows and develops.

Dr. Subramani Mani, MBBS, PhD: Proteomic biomarker druggability profiling using the domain of neonatal sepsis

A set of predictive, diagnostic or prognostic biomarkers relevant to a specific clinical condition or disease can be identified using focused literature search or obtained from research studies designed specifically for biomarker discovery. While these biomarkers may play an effective role in early detection, diagnostic evaluation or for assessment of prognosis, a mechanistic understanding is needed to ascertain if any of these biomarkers can be targeted from a druggability perspective to eventually move the biomarker(s) to the realm of therapeutics. Using pathway analytics and the TCRD resource component of IDG IMC we will try to show which of the neonatal sepsis biomarkers identified from the bio-signature study of neonatal sepsis conducted at the University of New Mexico Children's Hospital can serve as potential therapeutic targets for further evaluation from a drug development perspective.

Researchers are evaluating the scope, promise and potential of genomic medicine from different perspectives such as genetic risk quantification for diseases, determination of drug efficacy and dosage based on genetic profiling and for predicting adverse drug reactions. However, the major challenge in moving towards genomic medicine is to identify new targets of therapeutic value so that newer drugs can be developed to treat various diseases promptly and effectively for which no effective pharmaceutical interventions are currently available. The proposed panel will shed light on this challenge of moving genomic medicine forward from a therapeutic perspective.

The target audience for this panel will be researchers in the domains of clinical informatics, translational bioinformatics (broadly defined), drug discovery and pharmacogenomics. The panel discussion would be of interest to researchers in academia and industry.

Discussion questions:

1, What is druggability and druggability profiling?

2. What exactly are drug targets?

3. How do you define drugs from a druggability perspective?

4. Why has the new drug development pipeline slowed down considerably?

5. How will the IDG KMC accelerate new drug development efforts?

The panel organizer has personally contacted all the other panel participants and they have agreed to take part in the panel.

References:

1. Hopkins AL, Groom CR. The druggable genome. Nature reviews Drug discovery 2002;**1**(9):727-30
2. Bailey D, Zanders E, Dean P. The end of the beginning for genomic medicine. Nature biotechnology 2001;**19**(3):207-08
3. Rask-Andersen M, Almén MS, Schiöth HB. Trends in the exploitation of novel drug targets. Nature reviews Drug discovery 2011;**10**(8):579-90

Handling Clinical and Next Generation Sequencing data: new strategies in i2b2 and tranSMART

Shawn Murphy[1,3], Lori Phillips[1], Paul Avillach[3], Matteo Gabetta[2], Riccardo Bellazzi[2]

[1]Massachusetts General Hospital, Bsoton, MA / Partners Healthcare, Charlestown MA
[2]Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Italy
[3]Center for Biomedical Informatics, Harvard Medical School, Boston MA

Abstract

Within the next one to two years, we can expect to have over 100,000 genomes sequenced. Many of these genomes will belong to patients for whom we have extensive medical records. This will present an opportunity to find rapid associations of genotype to phenotype and exercise precision medicine. It is sobering therefore to learn that although we are poised to take advantage of the explosions of genomic data as well as the data from our healthcare enterprise, that there are few approaches to data representation that allow us to query the two types of data in unison. Although genome browsers at UCSC, Ensembl, and the National Center for Biotechnology Information (NCBI), as well as the model organism databases (e.g., Wormbase, Flybase, Saccharomyces Genome Database (SGD), and Mouse Genome Database (MGD]), have become essential tools for the analysis of genomic, molecular biology data, the representation of phenotypic data is unable to accommodate the complex representations of human phenotypes as exists in our healthcare system [11]. In the same way, typical representations of healthcare data in both transaction-based electronic medical records (VISTA, OpenEMR), as well as analytic healthcare models (mini-sentinel, virtual data model, and observational medical outcomes partnership) do not accommodate genomic data with any facility [12].

In this panel we will discuss and present three approaches that have been taken within the i2b2 community to approach the implementation of a genomic-phenomic query capability. These three groups agreed to contrast their approaches using a standard dataset to illustrate their specific approaches and tradeoffs.

For a first benchmarking of genomic and clinical data integration on the i2b2 platform, we focused on common data types and their basic usage. The goal of the three teams was to integrate whole exome variation data with human phenotype data, to enable smooth, intuitive, and rapid querying of both phenotypic and genomic variables. Whole exome SNPs and indels were extracted from The 1000 Genomes Project's Phase 1 Integrated Release [1] using The Genome Analysis Toolkit [2] and annotated using ANNOVAR [3] Cellular phenotype data collected on the same individuals was obtained from Wu et al. [4]. The teams were then asked to apply their method to answer the following basic use cases:

1. Which individuals with a lower mode of HLA-DQB1 protein levels (i.e. HLA-DQB1 log protein ratio < 0) have missense or nonsense mutations in that gene?
2. How many individuals with log HLA-DQB1 protein ratio < 0 have probably damaging missense mutations in that gene, as determined by PolyPhen-2 [5]?
3. How do the above answers change if we select just YRI (Yoruban) or CEU (European individuals)?
4. How many individuals carry variants previously implicated in hypertrophic cardiomyopathy, but recently found to be quite prevalent in normal populations [6]?

The ability for the platforms to address these use cases helps to evaluate the real-world utility of the data representations and query software. Each approach is now described in detail:

Approach 1: classic i2b2 data representation – Lori Phillips

The classic i2b2 data representation uses the standard Sequence Ontology (http://www.sequenceontology.org/) represented within the format of the i2b2 ontology. The ontology organizes a set of concepts describing the structural change of the variant, such as a "Single-nucleotide polymorphism: SNP/SNV, or the insertion or the deletion of bases: Indel. Then the attributes of the structural variation are expressed as "modifiers" for that variant, such as numbers representing the location of the variant, the gene in which the variation is thought to occur, whether it is synonymous, and perhaps the PolyPhen-2 score of the variation which predicts the biological effect on the function of the gene protein.

The main features of this approach are:

1. Flexibility to create queries that mix phenotype and genotype information in the visual interface of i2b2 where Boolean queries are represented. These queries can be created in real time, and answers returned in seconds from tables with millions of structural variants.
2. Variant information that is loaded directly from an ANNOVAR output file, which fully represent detailed annotations on millions of variants.
3. Ability to represent fully annotated structural variants in a simple to use query interface.

Approach 2: tranSMART – Paul Avillach

TranSMART is an open source knowledge management platform that enables scientists to develop and refine research hypotheses by investigating correlations between disparate data sources [7]. This open source platform is an i2b2 spinoff created in 2008 by Johnson & Johnson in the context of clinical trials [8]. On top of i2b2 the tranSMART application adds lightweight analytic capabilities, an expandable analysis framework and many data management options. We developed an analytic and integration pipeline to load ANOVAR variant annotations in tranSMART

The main features of this approach are:

1. The ability to create fine-grained phenotypic queries against fine-grained genomic information
2. R and Bioconductor analytics extend the i2b2 functionalities.
3. The visual dashboards enables the generation of multiple hypotheses.
4. Using SQL*Loader permits all the genomic data to be loaded very quickly into the Oracle database.

Approach 3: NoSQL – Matteo Gabetta

The i2b2+NoSQL approach extends the i2b2 platform in order to query both phenotype and genotype data by using two databases: the standard i2b2 warehouse to store phenotypes and CouchDB [9], a NoSQL document store, for genetic variants.

The main features of this approach are:

1. The flexibility and scalability provided by NoSQL databases in general and by CouchDB in particular. CouchDB is based on a schema-less data model: it stores one JSON document for each patient's variant containing all the associated data. Moreover, it scales very well when the data volume increases [10]).
2. Very fast query times, provided by an indexing system built on the JSON files.
3. The automatic creation of JSON files from a set of multi-sample VCF files
4. The opportunity to separately tune the two databases, given the uneven nature of the data they are entrusted to manage.
5. The system is thus based on two parts: data annotation/upload and data query.

Discussion questions

- The panel will show the results of the three approaches on the case studies,
- It will discuss the main similarities/differences between the three approaches
- It will show the strong and weak points of the 3 approaches
- It will emphasize the costs, in terms of time and computational resources

Panel organizer and participants

The panel is organized by Shawn Murphy and Riccardo Bellazzi, and the panelists are Lori Phillips, Paul Avillach, Matteo Gabetta. The panel organizer declares that all participants have agreed to take part on the panel.

References

[1] 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012 Nov 1;491(7422):56–65.

[2] DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011 May;43(5):491–8.

[3] Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. Oxford University Press; 2010 Sep;38(16):e164–4.

[4] Wu L, Candille SI, Choi Y, Xie D, Jiang L, Li-Pook-Than J, et al. Variation and genetic control of protein abundance in humans. Nature. 2013 Jul 4;499(7456):79–82.

[5] Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010 Apr;7(4):248–9.

[6] Andreasen C, Nielsen JB, Refsgaard L, Holst AG, Christensen AH, Andreasen L, et al. New population-based exome data are questioning the pathogenicity of previously cardiomyopathy-associated genetic variants. Eur J Hum Genet. 2013 Sep;21(9):918–28.

[7] Szalma S, Koka V, Khasanova T, Perakslis ED. Effective knowledge management in translational medicine. J Transl Med. BioMed Central Ltd; 2010;8(1):68.

[8] Perakslis ED, Van Dam J, Szalma S. How informatics can potentiate precompetitive open-source collaboration to jump-start drug discovery and development. Clin Pharmacol Ther. 2010 May;87(5):614–6.

[9] Anderson, J. Chris, Jan Lehnardt, and Noah Slater. *CouchDB: the definitive guide*. O'Reilly Media, Inc., 2010.

[10] Cattell, Rick. "Scalable SQL and NoSQL data stores." *ACM SIGMOD Record* 39.4 (2011): 12-27.

[11] Schattner P (2007) Automated Querying of Genome Databases. PLoS Comput Biol 3(1): e1. doi:10.1371/journal.pcbi.0030001

[12] El-Sappagh, S.H. (2012) Electronic Health Record Data Model Optimized for Knowledge Discovery, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 1, September 2012 ISSN (Online): 1694-0814

Panel: Clinical Natural Language Processing in Languages Other Than English

Aurélie Névéol[1], PhD, Hercules Dalianis[2], PhD, Guergana Savova[3], PhD, Pierre Zweigenbaum[1], PhD

[1] CNRS, LIMSI – rue John von Neumann, 91400 Orsay, France
[2] DSV, Stockholm University, P.O Box 7003, 164 07 Kista, Sweden
[3]Children's Hospital Boston and Harvard Medical School, Boston, Massachusetts, USA

{neveol;pz}@limsi.fr,hercules@dsv.su.se,Guergana.Savova@childrens.harvard.edu

Abstract

Natural Language Processing (NLP) of clinical free-text has received a lot of attention from the scientific community. Clinical documents are routinely created across health care providing institutions and are generally written in the official language(s) of the country these institutions are located in. As a result, free-text clinical information is written in a large variety of languages. While most of the efforts for clinical NLP have focused on English, there is a strong need to extend this work to other languages, for instance in order to gain medical information about patient cohorts in geographical areas where English is not an official language. Furthermore, adapting current NLP methods developed for English to other languages may provide useful insight on the generalizability of algorithms and lead to increased robustness. This panel aims to provide an overview of clinical NLP for languages other than English, as for example French, Swedish and Bulgarian and discuss future methodological advances of clinical NLP in a context that encompasses English as well as other languages.

General Description of the Panel

The goal of this panel is to engage the medical informatics and clinical Natural Language Processing community in a discussion about ways to advance research through languages other than English. We will provide an overview the current state of clinical NLP in a variety of European and non-European languages as well as focused reports on French, Swedish and Bulgarian. We will motivate the need for developing clinical NLP in languages other than English by the potential for methodological and medical advances. Finally, we will propose strategies to contribute to advance work on languages other than English and integrate it in a state-of-the art platform.

Clinical NLP in languages other than English

Natural Language Processing (NLP) of clinical free-text has received a lot of attention from the scientific community, demonstrating its potential to provide the means to analyze large quantities of documents rapidly and accurately (Demner-Fushman et al. 2010). Prime clinical applications for NLP include assisting healthcare professionals with retrospective studies and clinical decision making. The ability to analyze clinical text in languages other than English opens access to important medical data concerning cohorts of patients who are treated in countries where English is not the official language. Recently, Kohane et al. (2012) also showed the impact of methods allowing an aggregated exploitation of clinical data. In this context, data extracted from clinical texts in languages other than English adds another dimension to data aggregation.

As the importance of clinical NLP gains recognition, clinical corpora become available to researchers in languages other than English, prompting work that sometimes builds on methods validated for English. Adapting systems that work well for English to another language is a difficult task that may be carried out with varying level of success depending on the task and language (Grouin et al., 2009; Velupillai et al. 2014; Täckström et al., 2012). For non-European languages, approaches that account for entirely different word and sentence structures sometimes need to be developed (Shinohara et al. 2013), and cultural differences between clinical narrative styles accounted for (Wu et al. 2013). Access to terminologies and corpora in languages other than English can also be challenging (Schulz et al. 2013; Xu et al. 2013). These experiments prompt a reflexion on how to carry out clinical NLP in a more global context: should methods be developed for one language and then ported to other languages? Can the source language method benefit from the porting? Can algorithms be more robust if they are designed with a multi-language perspective from the start?

French is widely spoken around the world and benefits from one of the largest coverage in the UMLS. Automatic de-identification is becoming quite advanced for French (Grouin & Névéol, 2013), leading to good results for targeted clinical information extraction tasks (Deléger et al. 2010; Grouin et al. 2011). Recent efforts from the French biomedical Informatics community have addressed rules and regulations to improve the access of NLP researchers to clinical corpus. Furthermore, the success of initiatives such as that reported by Grouin et al. (2011) increased the awareness of the potential implication of clinical NLP in clinical practice and contributed to making the timing ripe for making clinical corpus available for annotation and NLP tool development. On-going efforts currently address the annotation of clinical corpora for entity, modality and relations. Tools are being designed for information extraction as well as semantic indexing, information retrieval and clinical data visualization.

Much of the research in **Swedish** clinical NLP has used the Stockholm EPR Corpus, (Dalianis 2012), that contains more than one million patient records encompassing the years 2006-2010, from over 550 clinical units origin from Karolinska University Hospital. Part of this corpus has been manually annotated for Protected Health Information, negations, uncertainty levels, symptoms, diseases, drugs, body parts and abbreviations. The annotated corpora have been used both for training of machine learning systems and evaluation. Some applications are explorative as comorbidity networks, warning and reporting systems detecting hospital acquired infections or adverse drug events, but also work on text simplification of patient record content for the layman patient, (Dalianis 2012). Tools that have been developed for this is an adaptation of NegEx for Swedish (Skeppstedt 2012), a system for classifying terms into six levels of assertion levels pyConTextSwe, (Velupillai et al. 2014), abbreviation detection, (Isenius et al. 2012) and machine learning system based on CRF++ that recognizes named clinical entities as symptoms, diseases, drugs and body, (Skeppstedt et al. 2014).

Integrating languages other than English in Apache cTAKES

Apache cTAKES (ctakes.apache.org) has been quite successful in assembling and sustaining a global community of developers and users of state-of-the-art English language clinical NLP. Because these techniques involve computational machine learning methods, datasets from the targeted language are needed to train and evaluate the algorithms on. We will discuss what types and size of data were used to build the various cTAKES components – sentence boundary detector, tokenizer, part of speech tagger, syntactic parser, event and temporal expression detector, temporal relation modules, general relation module. We will also discuss what types of gold standard labels (and how much of each type) are needed to port cTAKES components to other language within the light of some use cases such as porting the temporal expression discovery and normalization module originally developed for English (Bethard, 2013) to Swedish. We will outline available resources in other languages such as Swedish, Finnish, Bulgarian. This is a step towards globalization of information extraction from the clinical narrative.

Panelists

Prof. Hercules Dalianis (Professor at Stockholm University, Sweden) will present on work building on the Stockholm EPR Corpus, a major resource for Swedish clinical NLP.
Dr. Aurélie Névéol (staff scientist at LIMSI-CNRS, France) will act as a moderator and will present the medical and methodological benefits of clinical NLP in languages other than English. Dr. Névéol has been leading a project addressing the automatic understanding of French clinical narratives for translational research.
Dr. Guergana Savova (Assistant Professor at Harvard Medical School) will talk about integrating clinical NLP in different languages. Dr. Savova has been leading the development of the core Clinical Text Analysis and Knowledge Extraction System, now part of the Apache Software Foundation (cTAKES; ctakes.apache.org).
Dr. Pierre Zweigenbaum (principal investigator at LIMSI-CNRS, France) will discuss research efforts for clinical NLP in French, many of which he has initiated and coordinated over the past decade.

List of Discussion Points

After the introductory presentations, the moderator will ask questions as well as solicit questions from the audience, to prompt discussion among the panelists. Potential topics and questions include:
- Describe problems and experiences regarding NLP work on clinical text. How language-dependant is the work you are familiar with? What are the specificities of each language?
- Describe instances of a successful NLP application in a language other than English that yielded interesting new medical knowledge.
- Discuss the methodological challenges to bringing NLP in languages other than English to the level of state-of-the-art for English. Are some specific languages riper than others? Which languages? Why?

- Describe the technical and organizational challenges that must be overcome to integrate several languages other than English in an NLP platform such as cTAKES

References

1. Bethard, Steven. 2013. A Synchronous Context Free Grammar for Time Normalization. Proc. Conference on Empirical Methods in Natural Language Processing. http://www.aclweb.org/anthology/D13-1078
2. Dalianis, H, M. Hassel, A. Henriksson and M. Skeppstedt. Stockholm EPR Corpus: A Clinical Database Used to Improve Health Care. Proc. 4th Swedish Language Technology Conference, (SLTC-2012), Lund, Sweden, October 25-26, 2012, pp. 17-18
3. Deléger L, Grouin C, Zweigenbaum P: Extracting medication information from French clinical texts. *Stud Health Technol Inform* 2010, 160:949–953.
4. Demner-Fushman D, Chapman WW, McDonald CJ: What can natural language processing do for clinical decision support? J Biomed Inform 2009, 42:760–772.
5. Grouin C, Rosier A, Dameron O, Zweigenbaum P. Testing tactics to localize de-identification. Stud Health Technol Inform. 2009;150:735-9.
6. Grouin C, Deléger L, Rosier A, Temal L, Dameron O, Van Hille P, Burgun A, Zweigenbaum P. Automatic computation of CHA2DS2-VASc score: information extraction from clinical texts for thromboembolism risk assessment. AMIA Annu Symp Proc. 2011;2011:501-10.
7. Grouin C, Névéol A. De-identification of clinical notes in French: towards a protocol for reference corpus development. J Biomed Inform. 2013 Dec 29. pii:S1532-0464(13)00205-0.
8. Isenius, N., Velupillai, S, and Kvist, M. Initial Results in the Development of SCAN: a Swedish Clinical Abbreviation Normalizer. In Proc. CLEF 2012 Workshop on Cross-Language Evaluation of Methods, Applications, and Resources for eHealth Document Analysis – CLEFeHealth2012, CLEF, Rome, Italy.
9. Kohane IS, McMurry A, Weber G, MacFadden D, Rappaport L, Kunkel L, Bickel J, Wattanasin N, Spence S, Murphy S, Churchill S. The co-morbidity burden of children and young adults with autism spectrum disorders. PLoS One. 2012;7(4):e33224.
10. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010 Sep-Oct;17(5):507-13.
11. Skeppstedt, M. Negation detection in Swedish clinical text: An adaption of NegEx to Swedish. Journal of Biomedical Semantics, 2011. 2(Suppl 3), S3.
12. Skeppstedt, M., M. Kvist, H. Dalianis and Nilsson, G.H.. Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study. Journal of Biomedical Informatics, 2014. DOI: 10.1016/j.jbi.2014.01.012
13. Täckström O, McDonald R, Uszkoreit J. Cross-lingual word clusters for direct transfer of linguistic structure. Proc NAACL-HLT 2012, 477–87, Stroudsburg, PA, USA.
14. Velupillai, S., M. Skeppstedt, M. Kvist, D. Mowery, B. Chapman, H. Dalianis and W. Chapman. Cue-based assertion classification for Swedish clinical text - developing a lexicon for pyConTextSwe. Special issue: Text Mining and Information Analysis. Artificial Intelligence In Medicine, 2014.
15. Shinohara EY, Aramaki E, Imai T, Miura Y, Tonoike M, Ohkuma T, Masuichi H, Ohe K. An easily implemented method for abbreviation expansion for the medical domain in Japanese text. A preliminary study. Methods Inf Med. 2013;52(1):51-61.
16. Wu Y, Lei J, Wei WQ, Tang B, Denny JC, Rosenbloom ST, Miller RA, Giuse DA, Zheng K, Xu H. Analyzing differences between chinese and english clinical text: a cross-institution comparison of discharge summaries in two languages. Stud Health Technol Inform. 2013;192:662-6.
17. Schulz S, Ingenerf J, Thun S, Daumke P. German-Language Content in Biomedical Vocabularies. Proc CLEF 2013 Evaluation Labs and Workshop - CLEF-ER 2013. 2013.
18. Xu Y, Wang Y, Sun JT, Zhang J, Tsujii J, Chang E. Building large collections of Chinese and English medical terms from semi-structured and encyclopedia websites. PLoS One. 2013 Jul 9;8(7):e67526.

Statement of Participation

The first author affirms that all panel participants have agreed to participate and have contributed to the preparation of this document (as of August 1, 2014)

The PHR Ignite Project: Advancing Consumer-Mediated Exchange

Stephanie Rizk, MS[1]; Deepthi Rajeev, PhD, MS, MSc[2]; Aaron Seib[3]; Caroline Coy, MPH[4]

[1]RTI International, Research Triangle Park, NC; [2]HealthInsight, Salt Lake City, UT; [3]National Association for Trusted Exchange, Washington, DC; [4] Office of the National Coordinator for Health IT, Washington, DC

Abstract

The growth of electronic health information technology provides new opportunities to actively engage patients in their healthcare. Over the last few years, applications such as personal health records (PHRs) have been widely recognized as instruments to promote patient-centered care and influence patient engagement. Despite advances in health information technology, studies have found that PHRs are not widely adopted across the United States of America. Under funding from the Office of the National Coordinator for Health Information Technology (ONC), RTI International convened partners to support the PHR Ignite project which aimed to study the use of PHRs and their potential to support consumer-mediated exchange. The objectives were to: (a) understand attitudes of consumers towards PHRs, (b) identify features and functionalities of PHRs that may influence adoption among consumers, (c) understand the barriers that discourage the use of PHRs, and (d) explore the use of PHRs to demonstrate the bi-directional exchange between patients and healthcare providers. This panel will present findings and lessons learned during this project and discuss opportunities to improve patient engagement and consumer-mediated exchange.

Personal Health Record Use

The increased adoption of electronic health records and other health information systems offers new opportunities to promote patient–centered care. The promise of using personal health records (PHRs) as instruments that evolve patients from passive witnesses of their health care to engaged partners has been recognized in the literature over the last few years[1]. A true patient-centered PHR would represent a complete patient health care record and allow data sharing across health care settings and systems in a timely manner[2]. However, studies have found that PHRs are yet to be embraced by patients and providers in the United States of America[3]. Several initiatives, including the EHR Incentive program focus on meeting the needs of the patients by leveraging health information technology. For patients to realize the potential of the data now available to them, it is imperative to understand the current landscape of PHR applications in the United States, including existing limitations and barriers that influence adoption and use.

Panel Description

In the summer of 2012, the Office of the National Coordinator for Health Information Technology (ONC) funded the PHR Ignite project managed by RTI International under the State Health Policy Consortium. The project was a collaborative effort between multiple entities to advance understanding related to ONC's three pronged Consumer eHealth strategy: to increase patient access to data, support patients in taking action with that data, and to foster a change in attitudes regarding patient use and interaction with their own data. The members of the panel will describe findings from the PHR Ignite project which supported research to support the increased use of PHRs and their potential to support consumer-mediated exchange. Presenters will cover: a) an assessment of awareness within consumer, provider, and payor stakeholders in Utah and New Mexico and investigation into the factors which support active use of PHR applications, performed by HealthInsight; b) an overview of the technical and governance framework developed by the National Association of Trusted Exchange which supported bi-directional exchange between provider EHR and patient PHR systems in three states using Direct secure messaging. A third panelist from the ONC will discuss the impact of the PHR Ignite project, including a review of a second round of consumer-mediated exchange pilots.

Stephanie Rizk, Health IT Manager in the Center for Advancement of Health IT at RTI International, will serve as the panel moderator. Stephanie served as the assistant project director for the State Health Policy Consortium project and managed the PHR Ignite projects. She will provide a brief overview of the PHR Ignite projects, the projects' objectives, and each project's partners. For the past 10 years, Ms. Rizk has supported projects related to health information technology (health IT) and health information exchange (HIE). She is a skilled facilitator, having led a number of multistate workgroups focused on developing concrete, replicable solutions to interstate exchange challenges. Her areas of expertise include policy and governance related to interoperable HIE, implementation and use of Direct secure messaging for interstate exchange

Deepthi Rajeev, a Medical Informaticist at HealthInsight, was part of the team that assessed the current environment and functionalities of PHRs. She will present findings related to the assessment of attitudes and awareness of stakeholder groups representing consumers, providers, and payors in Utah and New Mexico. She will describe features of PHRs that facilitate in their adoption and illustrate barriers that hinder adoption among patients and providers. She will also present lessons learned from a pilot implementation to explore the use of stand-alone PHRs as a mechanism for consumer-mediated and bi-directional exchange of healthcare data. These findings will reveal the limitations of existing data sharing tools from the perspectives of the patients and providers and help inform the requirements needed to support the meaningful bi-directional exchange of data between patients and providers.

Aaron Seib is CEO of the National Association for Trusted Exchange (NATE), an organization which brings the expertise of state programs together to find common solutions that optimize the appropriate exchange of health information for greater gains in adoption and outcomes. The goal of the NATE PHR pilot project was to enable the wider use of PHRs as a vehicle for patients to bi-directionally exchange data with their providers and inform privacy and security policies as well as operational policies to scale the growth of trusted exchange with patients across the nation.

Caroline Coy is a Program Analyst in the Office of the National Coordinator for Health IT, where she works on adoption of health information exchange. In this role, she has served as a Project Officer for 15 states implementing health information exchanges. Currently, she leads consumer engagement efforts for the HIE program, including a pilot project to demonstrate working examples of consumer-mediated exchange. Caroline will discuss the impact of the PHR Ignite project, including a review of a second round of consumer-mediated exchange pilots.

All proposed panel participants have agreed to take part in the panel.

Discussion Questions

1. What can organizations do to overcome the barriers of PHR adoption and how will they benefit from increased PHR adoption?
2. How do tethered PHRs and untethered PHRs serve patients' needs differently in today's environment?
3. Should efforts to support consumer-mediated exchange focus on the general population or would it be valuable to identify specific patient populations that would most benefit from the use of PHRs?

References:

1. Reti SR, Feldman, HJ, Ross SE, Safran D. Improving Personal Health Records for Patient-centered Care. J Am Med Inform Assoc 2010;17:192-195.
2. Reid PP, Compton WD, Grossman JH, et al. Building a Better Delivery System: A New Engineering/Healthcare Partnership. National Academy of Engineering (US) and Institute of Medicine (US) 2005.

3. Tang PC, Ash JS, Bates DW, Overhage JM, Sands DZ. Personal Health Records: Definitions, Benefits, and Strategies for Overcoming Barriers to Adoption. J Am Med Inform Assoc 2006; 13(2):121-126

Tentative title: Squaring the circle: Managing local healthcare terminologies in the age of standardization

Program theme: Terminology and standards

Track: Applications of Informatics

Organizer:

Titus Schleyer, DMD, PhD
Clem McDonald Professor of Biomedical Informatics
Director, Center for Biomedical Informatics, Regenstrief Institute, Inc.
Professor, School of Medicine, Indiana University

Speakers:

Daniel J. Vreeman, PT, DPT, MSc
Assistant Research Professor, Indiana University School of Medicine
Research Scientist, Regenstrief Institute, Inc.

Mark Tuttle, FACMI
Board of Directors
Apelon

James J. Cimino, MD, Chief, Laboratory for Informatics Development
NIH Clinical Center
Lister Hill National Center for Biomedical Communication
National Library of Medicine
Bethesda, Maryland

Abstract

Healthcare has benefited from the increasing maturity, availability and implementation of standardized vocabularies. However, many healthcare organizations continue to maintain local terminogies. In doing so, healthcare institutions, individually as well as collectively, expend significant resources on local terminology creation, maintenance and mapping to standardized vocabularies. The purpose of this panel is to describe current approaches to managing both standardized and local terminologies, elucidate challenges and opportunities, and discuss future-oriented strategies for making the process more efficient and effective. The Regenstrief Dictionary, Logical Observation Identifiers Names and Codes (LOINC), Medical Entities Dictionary (MED) at Columbia University Medical Center and the Research Entities Dictionary (RED) at the NIH will serve as case studies. The panel will discuss how a variety of tools, such as the Distributed Terminology System from Apelon, can support local terminology efforts.

After participating in this session, the learner should be better able to:

- describe the major standardized vocabularies for representing content in electronic health records (EHR), and their strengths and weaknesses;
- discuss reasons and approaches for managing local terminologies; and

- describe developments to will make managing standardized vocabularies and local terminologies in a system context easier and more efficient.

General description of panel

Healthcare has benefited from the increasing maturity, availability and implementation of standardized vocabularies, such as the International Classification of Diseases (ICD), the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT), Logical Observation Identifiers Names and Codes (LOINC), and RxNorm. Resulting benefits include enhanced interoperability, communication among health care providers, measurability of outcomes and reusability of data for various purposes. The implementation and integration of standardized terminologies into electronic health record systems (EHR) is fairly well-understood and common. Many standards developing organizations (SDO) are using well-structured and formalized processes for updating and maintaining their standards.

However, despite these developments, many healthcare organizations continue to maintain local terminologies. Reasons for this circumstance include the need to represent content not available in standardized terminologies, specific preferences and needs of clinicians, researchers and administrators, and the needs of local communities. In doing so, healthcare institutions, individually as well as collectively, expend significant resources on local terminology creation, maintenance and mapping to standardized vocabularies. The purpose of this panel is to describe current approaches to managing both standardized and local terminologies, elucidate challenges and opportunities, and discuss future-oriented strategies for making the process more efficient and effective. The panel will focus on issues such as:

- How useful and appropriate are major standardized vocabularies for representing content in EHRs? Where do they excel/fall short?
- Why would healthcare institutions need/want to maintain a local terminology?
- How are standardized and local terminologies managed together in the context of commercial EHR systems?
- What are good operational processes for implementing standardized vocabularies and, at the same time, creating/maintaining necessary local terminology/ies to achieve a smooth, integrated whole?
- What tools are available to manage local terminologies? How well do they work? Where do they fall short?
- How can the results of local terminology creation/management processes feed back to the "greater good", e.g. others who are engaged in the same kinds of efforts or standards developing efforts?
- What novel directions, such as crowdsourcing, exist for local terminology management?

Dan Vreeman will describe the history, purpose, and ongoing evolution of the Regenstrief Dictionary. The Regenstrief Dictionary is a homegrown terminology created in 1972 as a core component of the Regenstrief Medical Record System. Now, more than 40 years later, it has given birth to the most widely adopted standard for observations, LOINC, and continues a central role in the operation of the Indiana Network for Patient Care, the country's largest health information exchange. The Regenstrief Dictionary contains 46,000 concepts covering areas such as diseases, drugs, lab tests, nursing orders and survey questions/answers. Many of these terms are linked to standard vocabularies. However, we have become acutely aware of the need to evolve our strategy for managing our terminology needs. Current efforts duplicate much work that already has been done in standard vocabularies. In addition, the software applications

Regenstrief is developing would be more generalizable if they were written on top of standard vocabularies. Dr. Vreeman will present Regenstrief's current strategy for adopting standard terminologies while retaining advantages of local dictionary development. In this context, he will discuss challenges and opportunities for terminology management, personnel resources, needed software tools and collaboration with local healthcare institutions. He will also share lessons learned by "eating what we cook" from Regenstrief's unique role as both an SDO and end user of standard vocabularies.

Mark Tuttle will discuss tools and processes for terminology management, drawing on the example of the NLM Unified Medical Language System (UMLS) Metathesaurus. He will describe Apelon's product and service suite focused on support for local terminology efforts, including the Distributed Terminology System, an open source terminology server. Mark will illustrate terminology development, maintenance and deployment with examples from the National Cancer Institute, the Veterans Health Administration and the Department of Defense. He will review lessons learned from more than two decades of terminology mapping and describe emerging best practices in terminology mapping workflow management.

James Cimino will describe the history, current state, purpose of and strategy to evolve the Medical Entities Dictionary (MED) at Columbia University Medical Center and the Research Entities Dictionary (RED) at the NIH Clinical Center. Each of these is a homegrown terminology that contains the controlled terminologies used by clinical and clinical research systems in their respective organizations. Each consists of an ontological structure that includes a multiple hierarchy, a set of high-level organizing concepts, low-level terms used to code actual patient data, and knowledge concepts to support representation of additional knowledge. Development of the MED was begun in 1988 and includes over 160,000 terms. Development of the RED was begun in 2008 and includes over 280,000 terms. Dr. Cimino will describe how these resources adapt to new terminologies (including standards), remain current with existing terminologies, and support a variety of patient care and research functions.

Why the topic of this panel is timely, urgent, needed and attention grabbing

There is an inherent tension between increasing standardization of content in healthcare and the continuing need for representing local content. As more healthcare institutions adopt standardized vocabularies, their limitations, as well as the limitations inherent in maintaining/evolving them, become more apparent. Standard and local terminology management tend to consume a significant, but largely invisible, amount of resources at healthcare organizations. In an age of increasing cost reduction, institutions are looking for answers to solve this problem.

A list of discussion questions to enhance audience participation

- How many of you represent healthcare delivery organizations? How many of those use ONLY standard terminologies in the EHR?
- How much effort/money do you think you spend on local terminology creation, management and mapping?
- What are your main challenges in managing the implementation of standard and/or local terminologies?

All panelists have agreed to participate on this panel at #AMIA2014.

What's In A Name: Precision Medicine and a New Nosology

Presenters:

Mark S. Tuttle, FACMI, Stuart J. Nelson, MD, FACMI, Yves A. Lussier, MD, FACMI, Funda Meric-Bernstam, MD

Abstract:

In 2012 the National Research Council published a white paper[1] on the need for a new taxonomy of disease, noting the failure of our current disease names to address the needs of precision medicine, based on a deeper understanding of nosology and the requirements of personalized medicine. The panel will discuss the implications of this call in several ways. Dr. Lussier will discuss promising translational bioinformatics innovations in precision therapeutics. Dr. Meric-Bernstam will discuss the importance of finding actionable abnormalities in the genome of the tumors she treats with precision. Dr. Nelson will discuss how this new knowledge can be incorporated into the Bloisian model of disease descriptions[2] as well as how the "classical" patient findings can be used to further develop these descriptions. Mr. Tuttle will discuss the growing experience with supervised and unsupervised machine learning as a tool of discovery, and the application of this paradigm to nosology.

The Call for a New Nosology:

The NRC report explains in considerable depth why current taxonomies of disease are impeding progress. Their "good" example is a patient with breast cancer, including a family history of breast cancer.

> Patient 1 is consulting with her medical oncologist following breast cancer surgery. Twenty-five years ago, the patient's mother had breast cancer, when therapeutic options were few: hormonal suppression or broad-spectrum chemotherapy with significant side effects. Today, Patient 1's physician can suggest a precise regimen of therapeutic options tailored to the molecular characteristics of her cancer, drawn from among multiple therapies that together focus on her particular tumor markers. Moreover, the patient's relatives can undergo testing to assess their individual breast cancer predisposition (Siemens Healthcare Diagnostics Inc. 2008). (p. 9)

Their "bad" example is a patient with Type II Diabetes.

---

[1] National Research Council. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*. Washington, DC: The National Academies Press, 2011. See … http://www.ucsf.edu/sites/default/files/legacy_files/documents/new-taxonomy.pdf AMIA colleagues who contributed to this report include Zak Kohne, Christopher Chute, Dan Masys, and Atul Butte.

[2] Blois, Marsden S. *Information and Medicine: The Nature of Medical Descriptions,* University California Press, Berkeley, 1984.

In contrast, Patient 2 has been diagnosed at age 40 with Type II diabetes, an imprecise category that serves primarily to distinguish his disease from diabetes that typically occurs at younger ages (type I) or during pregnancy (gestational). The diagnosis gives little insight into the specific molecular pathophysiology of the disease and its complications; similarly there is little basis for tailoring treatment to a patient's pathophysiology. The patient's internist will likely prescribe metformin, a drug used for over 50 years and still the most common treatment for type II diabetes in the U.S. No concrete molecular information is available to customize Patient 2's therapy to reduce his risk for kidney failure, blindness or other diabetes-related complications. No tests are available to measure risk of diabetes for his siblings and children. Patient 2 and his family are not yet benefitting from today's explosion of information on the pathophysiology of disease (A.D.A.M. Medical Encyclopedia 2011, Gordon 2011, Kellett 2011) (p. 10)

A central point in their argument is the need for a nosology that is data-driven, and continuously so. Current nosologies are not only inadequate, but their update processes – including the means by which updates are deployed by users – are too slow to keep up with important advances such as those that are part of precision medicine.

The report does not lay out a detailed plan but it does describe many opportunities to improve the naming and classification of diseases. Specifically, it proposes a period of bottom-up innovation constrained only by a desire to share and re-use data.

Stuart Nelson, formerly Head of the MeSH (Medical Subject Headings) Section at the National Library of Medicine (NLM), and lead developer of RxNorm, the current standard for medication terminology in the United States, will lead off with a proposed structure for a new nosology. This structure is based on patient descriptions and builds on the work of Marsden Scott Blois, MD, PhD[3], and is influenced by Dr. Nelson's tenure at the NLM. Simply put, precision medicine will need to be based on more than the diagnostic classes available in currently available taxonomies. As the NRC Report acknowledges, while a major driver behind the need for a new taxonomy of disease is our expanding understanding of molecular foundations, said taxonomy needs to include "incidental patient characteristics, or socio-environmental influences." (p. 14)

Precision Medicine:

While cautionary tales continue to appear[4,5], progress on molecular understanding continues. Specifically, while we're all used to reductionistic science making ever – finer distinctions regarding the structure of the world, occasionally reductionism in

---

[3] In addition to his book, see also Blois MS. "Medicine and the nature of vertical reasoning." N Engl J Med. 1988 Mar 31;318(13):847-51.
[4] Pelkoff K. "I Had My DNA Picture Taken, With Varying Results", *NY Times*, December 30, 2013.
[5] Feero W. "Clinical Application of Whole-Genome Sequencing: Proceed With Care" . *JAMA*. 2014;311(10):1017-1019.

one place – at the molecular level, for instance – can lead to the unified understanding of otherwise disparate events. Yves Lussier, Professor of Medicine and Associate Director of Precision Health and Cancer Informatics at the University of Arizona, and longtime contributor to ontologies supporting translation medicine, will describe an example of such unification, namely how pathophysiologic processes can start in different ways but end up sharing protein synthesis characteristics.

Funda Meric-Bernstam, Chair of Investigational Cancer Therapeutics and Professor of Surgery, at the MD Anderson Cancer Center, will illustrate how a focus on actionable, personal genomics affects the treatment of breast cancer patients. Put differently, her challenge is identify, evaluate and separate the actionable results from interesting, potentially important, but not yet actionable results, as they related to individual patients.

Data Science in Support of Nosology:

Mark Tuttle will address the need to evaluate taxonomies empirically. For example, given two taxonomies ostensibly classifying an overlapping population of patients how can we assess whether one is better than the other? Of course, this implies that we develop evaluative criteria. While, clearly a multi-dimensional challenge, prediction of therapeutic outcome is a unifying theme of this panel. Given this, and related, objectives, one can propose application of – for lack of a better name – the Google paradigm. This paradigm, described by Peter Norvig, Director of Research at Google, can be summarized as follows. First, domain experts "classify" data instances in some previously unclassified, or under-classified, domain, something physicians do already. Second, these classified examples are used as a "training set" during supervised learning of a predictive model. Third, the results of supervised learning are used to initialize unsupervised learning – what used to be called "clustering". Fourth, and finally, an oversight process using ongoing data streams maintains the resulting clusters.

Can such an approach prove useful in healthcare? Can it address some of the objectives of the NRC report? Can it support Dr. Nelson's approach to a new nosology? Can it aid Drs. Lussier and Meric-Bernstam in their effort to leverage emerging results from molecular biology and genomics? These, and related questions, will be left to the audience.[6]

---

[6] One of the drivers for this panel was a presentation by Phillip Bourne, new head of data science at NIH, at the *2014 ACMI Symposium* on "What Makes us Special?: The Genotype and Phenotype of Informatics through the Lens of Personalized Medicine". See "NIH Names Dr. Philip E. Bourne First Associate Director for Data Science" http://www.nih.gov/news/health/dec2013/od-09.htm "Phil will lead an NIH-wide priority initiative to take better advantage of the exponential growth of biomedical research datasets …" Francis Collins.

Going Digital: Transforming Medical Checklists for Improved Patient Care

Panelists
Bradford Winters, MD, PhD[1], Randall S. Burd, MD, PhD[2], Jesse Cirimele, PhD[3], and, Leslie Wu, BS[3]

Organizer & Moderator
Aleksandra Sarcevic, PhD[4]

[1]Johns Hopkins Medicine, Baltimore, MD; [2]Children's National Medical Center, Washington, DC; [3]Stanford University, Palo Alto, CA; [4]Drexel University, Philadelphia, PA

Abstract

This panel is aimed at addressing the implications and challenges of designing, developing, implementing and evaluating digital checklists in clinical settings. Panelists will share their experiences with checklist design and development, and discuss how digital formats may further improve the impact of the checklist in a range of settings. This is an important panel to have at AMIA in order to engage the community in discussing critical questions about advancing the checklist mechanisms in increasingly digital medical environments. Learning objectives include: (a) formulate an approach for designing digital checklists in a clinical setting; (b) formulate an approach for implementing digital checklist in a clinical setting; and (c) evaluate the effectiveness and impact of digital checklists using both simulated and clinical settings.

General Description

Checklists have become increasingly widespread in health care. Because their use has been associated with a significant decrease in omission errors, death rates, and inpatient complications, checklists are now used across different medical settings, including operating rooms (OR),[1] intensive care units,[2] and anesthesia.[3] An example of an effective use of a medical checklist is the WHO Surgical Safety Checklist, which improves team communication and consistency of care, while markedly reducing complications and deaths associated with surgery.[4]

Most medical checklists used today are paper-based, requiring care providers to manually record the presence or absence of the checklist items. Although paper checklists offer multiple benefits, questions have arisen about their ease of introduction into workflow and their impact on safety; they may complicate tasks, reduce efficiency, and require additional time and attention.[5] While many hospitals continue to use paper checklists, some have started implementing computer-based checklists and interactive cognitive aids to improve patient care and augment user experience.[6] This trend toward checklist digitization does not come as a surprise as medical work has become increasingly digital over the last decade. Similar to paper checklists, however, introducing digital checklists into clinical settings is not simple, and requires consideration of content, format, timing, trial, and feedback, followed by formal testing and evaluation. In addition, clinical settings pose many challenges to designing digital checklists. Checklist design principles taken from the aviation industry have worked well for static, paper-based checklists,[7] but may not be applicable to designing digital checklists for the dynamic and, often more chaotic, medical work.

This panel gathers four researchers and practitioners who have been involved in the design and development of both paper and digital checklists in different clinical settings. Although the work of all panelists focuses on critical care events—surgery, trauma resuscitation and anesthesia crisis—the approaches to designing, implementing and evaluating the checklists are diverse. The four panelists will share experiences in conducting their research and discuss the implications of their studies. These individual stories will then serve as the material for an open discussion with the audience.

Structure & Discussion Topics

The panel session will begin with a ten-minute presentation by the moderator to describe the purpose and format of the panel, and to introduce the panelists. This will be followed by three, 15-minute presentations by panelists detailing experiences, approaches and outcomes of their exemplar studies. We will then open the floor to discussion

among panelists and with the audience in the remaining 35 minutes. Although we expect that topics will emerge through the discussion, some of the topics that may help structure the conversation are as follows:

Design principles & issues: Most medical checklists have been designed based on principles adopted from the aviation industry.[7] These principles, however, may not be applicable to designing digital checklists for the dynamic and, often more chaotic, medical work. Panelists will discuss approaches they used to gather design requirements, as well as overcome design challenges they faced. We also recognize that some systems may be built in house, while some may be commissioned or purchased from a third party. Questions arising here relate to managing and overseeing the design process, and ensuring that the final product meets the user needs.

Implementation issues: Prior research has identified several barriers to the successful implementation of paper-based checklists, including poor communication among team members, lack of leadership, inappropriate timing for checking an item, and time taken up by checklist completion.[5] Do the same barriers apply to implementing digital checklists? What are the issues we need to consider when implementing digital checklists?

Evaluation issues: Evaluating the impact and effectiveness of checklists is an important aspect in checklist design. Panel members have extensive experience in testing and evaluating both paper and digital checklists in simulated and clinical settings. We will discuss how to design evaluation studies for digital checklists, obtain human subjects protection approvals, collect data, and define evaluation metrics.

Anticipated Audience & Importance of the Topic

This panel session will be of interest to all AMIA members, and particularly to healthcare providers and informatics professionals working in critical care settings. This is a timely topic to discuss given the widespread use of medical checklists and the trend toward digitization of medical care. The panel will educate about the implications and challenges associated with designing, developing, implementing and evaluating digital checklists in a range of clinical settings. Panel members will synthesize their perspectives on these issues and likely future developments in this area, exploring a diverse set of topics and engaging in thoughtful discussion with the audience.

Brief Description of Panelists & Presentations

Bradford Winters, MD, **PhD** is an Intensivist and Anesthesiologist at the Johns Hopkins Hospital and a Core Faculty member of the Armstrong Institute for Patient Safety and Quality at the Johns Hopkins University School of Medicine. His research interests include reducing patient harm through the implementation of Rapid Response Systems, applying broader modalities for patient monitoring and the use of cognitive tools such as checklists and technological solutions to reduce patient harm and adverse events.

Dr. Winters will present a framework for checklist development, and discuss the development and implementation of checklists in the Intensive Care Unit (ICU) environment to improve safety and reduce gaps in the quality of care, including reduction in hospital acquired infections, improved mechanical ventilator care, and appropriate deep vein thrombosis prophylaxis. He will discuss the challenges of implementing and especially digitizing such cognitive tools into the workflow of the ICU.

Randall S. Burd, MD, PhD is the Division Chief of Emergency Trauma and Burn Services and a pediatric surgeon at Children's National Medical Center in Washington, DC. His research interests are in pediatric trauma, with a particular focus on prehospital prediction of severe injury and emergency department resuscitation. Over the past eight years, Dr. Burd has directed a multidisciplinary team that is developing new analytic and technological approaches for rapidly identifying severely injured children, and providing them with safe and efficient care in the emergency department. His currently-funded research has given him expertise in directing the clinical development and assessment of new technologies to support teamwork during resuscitation, designing and implementing simulation for assessment of new technologies, and analyzing the information needs of resuscitation teams.

Dr. Burd will present his recent work on the checklist development for trauma resuscitation. Adverse outcomes in this domain have been associated with omission of key steps in the initial management of injured patients despite continued simulation training and didactic teaching. Dr. Burd and his research group have developed a checklist for trauma resuscitation that improves compliance within this process. Dr. Burd will discuss the approach they used to develop and evaluate this checklist in both simulated and clinical settings. He will also discuss how the checklist impacts communication patterns and potential ways in which a digital format may further improve the checklist's impact in this setting. Finally, Dr. Burd will present how they used digital pen technology to inform the design of a digital checklist, using an approach that allows analysis of the accuracy and timeliness of checkbox completion.

Jesse Cirimele, PhD is a Postdoctoral scholar between the Computer Science Department and the Medical School at Stanford University. He received his PhD in Computer Science from Stanford University with a focus on human-computer interaction. Dr. Cirimele has Bachelors degrees in Cognitive Science and Mathematics from University of California, San Diego. His research has looked into decision support in critical care, adaptive mobile interfaces, mobile health interventions, and improved browser history interfaces.

Leslie Wu, BS is a PhD candidate in Computer Science at Stanford University with a focus on human-computer interaction. Leslie has a BS in Computer Science from the University of Illinois at Urbana-Champaign. Her previous research has investigated web service development and on-demand techniques for data integration and visualization. She has worked on several mobile health startups and is interested in applying crowdsourcing and mobile applications and technology to improving healthcare systems.

Jesse Cirimele and Leslie Wu will co-present their joint work on Dynamic Procedure Aids, an approach for addressing four key problems in the use of medical checklists: ready access to the aids, rapid assimilation of their content, professional acceptance of their use in medical procedures, and limited attention available to their users. To understand the efficacy of Dynamic Procedure Aids for crisis response, Jesse and Leslie created dpAid, a software system for crisis medicine. dpAid's design was based on a year-long observational study of medical teams responding to simulated crises. dpAid was deployed on tablets and large-screen displays, and evaluated in simulated medical crises using eye-tracking hardware.

Aleksandra Sarcevic, PhD is an Assistant Professor of Informatics in the College of Computing and Informatics at Drexel University. Her research interests are in computer supported cooperative work and medical informatics, with a focus on ethnographic studies of work practices and interface design for safety-critical medical settings. Her recent work is in the area of emergency medical resuscitations, where she hopes to reduce human errors and increase teamwork efficiency by introducing a series of technological interventions. She was awarded a 2013 National Science Foundation Early CAREER Grant to continue her work on information technology design for fast-response medical teams.

Dr. Sarcevic will moderate the panel.

Participation statement: All proposed panelists have agreed to participate in the panel.

References

1. Lingard L, Espin S, Rubin B, et al. Getting teams to talk: Development and pilot implementation of a checklist to promote inter-professional communication in the OR. Qual Saf Health Care. 2005 Oct;14(5):340-6.
2. Berenholz SM, Pronovost PJ, Lipsett PA, et al. Eliminating catheter-related bloodstream infections in the intensive care unit. Crit Care Med 2004;32:2014-20.
3. Hart EM, Owen H. Errors and omissions in anesthesia: A pilot study using a pilot's checklist. Anesth Analg. 2005 Jul;101(1):246-50.
4. Haynes AB, Weiser TG, Berry WR, et al. A surgical safety checklist to reduce morbidity and mortality in a global population. N Engl J Med 2009;360:491–9.
5. Fourcade A, Blache JL, Grenier C, Bourgain JL, Minvielle E. Barriers to staff adoption of a surgical safety checklist. BMJ Qual Saf 2012;21:191-7.
6. Robbins, J. Hospital checklists: Transforming evidence-based care and patient safety protocols into routine practice. Crit Care Nurs Q 2011;34(2):142-9.
7. Gawande, A. The Checklist Manifesto: How to Get Things Right. Metropolitan Books, New York, NY, USA, 2009.

Evolving Career Landscapes in Biomedical and Health Informatics

Rui Zhang, PhD[1,2], William Hersh, MD[3,4], Genevieve B. Melton, MD, MA[1,2], Yang Huang, PhD, MSCS[5], Laura Wiley, MS[6], Julie Doberne[3], Nawanan Theera-Ampornpunt, MD, PhD[7]

[1]Institute for Health Informatics, [2]Department of Surgery, University of Minnesota, Minneapolis, MN; [3]Department of Medical Informatics, [4]Clinical Epidemiology, Oregon Health & Science University, Portland, OR; [5]Kaiser Permanente Southern California, San Diego, CA; [6]Department of Biomedical Informatics, Vanderbilt University, Nashville, TN; [7]Medicine Ramathibodi Hospital, Mahidol University, Bangkok, Thailand

Abstract

The career landscape for biomedical and health informatics students continues to expand. With the recent successful administration of the first Clinical Informatics Subspecialty Board Exam and trends in the fields such as healthcare data analytics, the informatics workforce must be well-prepared for diverse career opportunities and a changing healthcare landscape. Following tradition, the AMIA Student Working Group proposes a "career panel" to offer perspectives and advice for students on the success of their career opportunities and professional development. This year, panelists include an academic program director, hospital chief medical informatics officer (CMIO), and industry Research and Development (R&D) director, each with different training backgrounds (e.g., MD, PhD, MA). They will share their experiences and discuss upcoming trends in informatics careers with students. This panel will help current informatics students and early-career professionals to better prepare for and develop their careers.

General Description

Since 2002, the American Medical Informatics Association (AMIA) Student Working Group (ST-WG) has organized a panel during the AMIA annual symposium each fall to provide perspectives and advice on career development for current informatics students, recent graduates and early-career professionals. The career panel has become a valuable component of the symposium for the students and other attendees to actively interact with successful informaticians.

The career opportunities in the field of informatics are growing exponentially due to clinician and researcher needs for biomedical and health information, as well as the continued adoption of health information technology in clinical practice. In 2013, the first group of physician-informaticians passed the first *American Board of Preventive Medicine Clinical Informatics Subspecialty Board Exam*. These board-certified professionals are only the first of many more to come, and AMIA has additional planning underway to develop certification pathways for other informatics professionals besides physicians. In this exciting age, career development advice is more important than ever. Last year's career panel featured panelists from academia and industry with clinical and technical backgrounds. This panel attracted over one hundred attendees and raised many interesting questions.

This year we not only build on this successful tradition but also enhance it in the following ways. This panel will

(a) Feature panelists who are involved in the Board Exam, providing perspectives and helpful advice for professionals;
(b) Give an overview of the entire workforce of informatics professionals (not limited to clinicians);
(c) Discuss current shift in focus from implementation to analytics;
(d) Provide suggestions for different degree levels of informatics, including MD, PhD, and Masters students; and
(e) Share experiences of various career paths: faculty, Chief Medical Information Officer (CMIO), and research directorship.

The planned format is the traditional panel discussion. Each of the panelists is allotted 15 minutes for a short presentation followed by 5 minutes of questions. After 10 minutes for closing discussion, there is a 30-minute session to increase audience participation, with a list of questions.

1) Which experience or factors played a key role in your informatics career?

2) What are the expectations for professional, PhD, and master students, both in academia and industry?
3) What are the differences in career paths for professional, doctoral and master students?
4) How should students prepare and ready themselves for the coming career opportunities and trends?
5) How should early career informatics graduates continue to improve skills and achieve success in their careers?
6) What level of programming knowledge should an "informatics expert" have, clinicians or non-clinicians?

Given the different career experiences and educational backgrounds of the panelists, this panel will provide valuable perspectives for professional, PhD and master's students, who are studying, looking for jobs or just entering the field and beginning to develop their careers in informatics. By the end of the panel, attendees should be able to:

1) Gain familiarity with the clinical informatics board certification;
2) Gain a sense of the multitude of career paths in biomedical informatics;
3) Understand current trends in the biomedical informatics field;
4) Understand the goals and pathways for various degree levels (MD, PhD, Masters, etc.)

Organizers

AMIA Student Working Group Executive Committee and Volunteers
Rui Zhang, PhD, Chair
Laura Wiley, MS, Chair-Elect
Nawanan Theera-Ampornpunt, MD, PhD, Immediate Past Chair
Tiffany Kelley, PhD, MBA, RN, Student Representative to the AMIA Board
Julie W. Doberne, Member-at-large (Resident-Fellow)
Kate Fultz Hollis, MS, Member-at-large (Masters)
Onyinyechi U. Enyia Daniel, MA, Member-at-large (Doctoral-PhD)
Shauna M. Overgaard, Member-at-large (Doctoral-PhD)
Andrew M. Harrison, Member-at-large (Doctoral-Professional)
Anthony Omosule, Member-at-large (International)
Edmond Ramly, PhD Candidate, Editor-in-Chief, ST-WG Newsletters
Scott M. Sittig, MHI, RHIA, ST-WG Public Policy Liaison to the Public Policy Committee
Kourosh Ravvaz, MD, MPH, ST-WG Representative to the Implementation Forum Analysis Group
David Marc, MS, AMIA Academic Forum Member, Health Informatics Graduate Program Director and Assistant Professor at The College of St. Scholastica
Christopher Macintosh, RN, BSN, Student Representative to the Nursing Informatics Working Group

Panel Participations

Informatics Graduate Program Chair
William (Bill) Hersh, MD, FACP, FACMI
Course Director, Clinical Informatics Board Review Course
Diplomate, Clinical Informatics Subspecialty, American Board of Preventive Medicine
Professor and Chair, Department of Medical Informatics and Clinical Epidemiology, School of Medicine, Oregon Health & Science University, Portland, OR

CMIO
Genevieve Melton-Meaux, MD, MA, FACS, FASCRS
Chief Medical Information Officer, University of Minnesota Health
Diplomate, Clinical Informatics Subspecialty, American Board of Preventive Medicine
Associate Professor, Department of Surgery
Core Faculty, Institute for Health Informatics
University of Minnesota, Minneapolis, MN

Research and Development (R&D) Director
Yang Huang, PhD, MSCS

Director of Research and Development Medical Informatics
Systems, Solutions, and Deployment
Kaiser Permanente Southern California, San Diego, CA

Moderator
Rui Zhang, PhD
Chair, AMIA Student Working Group
Assistant Professor, Department of Surgery
Clinical Assistant Professor, Core Faculty, Institute for Health Informatics
University of Minnesota, Minneapolis, MN

Statement

It is confirmed that all panelists listed in this proposal have agreed to participate in this panel. Panelists are aware that there are no travel or registration funds available. Panelists are also aware that the Student Working Group is unable to reimburse their registration costs.

Panel Title:

Technology transfer from biomedical research to clinical practice: measuring innovation performance

Abstract

Background Earlier studies documented 17 years of transfer time from clinical trials to practice of care. Launched in 2002, the NIH translational research initiative needs to develop metrics for impact assessment. A recent White House report highlighted that R&D productivity is declining as a result of increased research spending while the new drugs output is flat.

Goals To support measurement and improvement, this study developed an expanded model of research based innovation and identified performance thresholds of technology transfer from research to practice.

Methods A wide range of models for transfer of research to practice have been collected and reviewed. Subsequently, clusters of models have been created based on common characteristics. Additionally, measures of research transfer were reviewed based on nationally available data. Milestones of progress have been identified based on the 2012 AUTM performance reports.

Results An integrated, Intellectual Property Transfer (IPT) model is described. The central but often disregarded role of research innovation disclosure is highlighted. Numeric milestones of technology transfer are recommended at threshold (top 50%), target (top 25%) and stretch goal (top 10%) performance levels. Measures and corresponding target levels include research spending to disclosure (<$1.88M), disclosure to patents (>0.81), patents to start-up (>0.1), patents to licenses (>2.25) and average per license income (>$48,000). Several limitations of measurement are described.

Conclusions Academic institutions should take strategic steps to bring innovation to the center of scholarly discussions. Technology transfer regulations must be transparent and must support researchers and institutions needs in the innovation to practice cycle (IPC). Research on research, particularly on pathways to disclosures, is needed to improve R&D productivity. Researchers should be informed about the technology transfer performance of their institution.

Panel Members:

Dr. Balas and Dr. Elkin recently published:

Technology transfer from biomedical research to clinical practice: measuring innovation performance. Balas EA, Elkin PL. Eval Health Prof. 2013 Dec;36(4):505-17.

E. Andrew Balas, MD, PhD

Andrew Balas serves as Dean and Professor at Georgia Regents University. His expertise includes development of priorities for the production of innovative scientific knowledge responsive to societal needs and application of advanced digital technologies for transferring research to practice.

He is member of the Board of Directors of the Friends of the National Library of Medicine and also the Allied Health Research Institute. He is an elected member of the American College of Medical Informatics and the European Academy of Sciences and Arts.

Andrew Balas has been effective in taking on the status quo, achieving breakthrough performance improvements and fighting for better public access to scientific discoveries. His studies about delay and waste in the transfer of research results to health care are often cited as reference points in translational research initiatives. As a Congressional Fellow working for the Public Health and Safety Subcommittee of the United States Senate, he drafted the Healthcare Quality Enhancement Act of 1999 that, among others, first achieved government action on reducing errors in health care and was signed into federal law (Dec. 6, 1999).

His leadership emphasizes positive response to community needs, teamwork and measurable improvement. During six years of his previous service as Dean, the College of Health Sciences achieved many successes at Old Dominion University in Norfolk, Virginia (e.g., double digit increases in enrollment, solid accreditations, launching of new programs, tenfold increase in externally funded research; multimillion dollar fundraising, new R&D partnerships with industry). Previously, he served as Dean of the School of Public Health in St. Louis, Director of the Missouri European Union Center and Weil Distinguished Professor of Health Policy at the University of Missouri.

His academic credentials include over 100 publications, externally funded research in excess of 10 million dollar and publications that cumulatively attracted thousands of citations. He obtained degrees in medicine, medical informatics (Ph.D.), and applied mathematics.

Peter L. Elkin, MD, MACP, FACMI, FNYAM

Dr. Elkin serves as Professor and Chair of the UB Department of Biomedical Informatics. He is also a Professor of Medicine at the University at Buffalo. Dr. Peter L. Elkin has served as a tenured Professor of Medicine at the Mount Sinai School of Medicine. In this capacity he was the Center Director of Biomedical Informatics, Vice-Chairman of the Department of Internal Medicine and the Vice-President of Mount Sinai hospital for Biomedical and Translational Informatics. Dr. Elkin has published over 120 peer reviewed publications. He received his Bachelors of Science from Union College and his M.D. from New York Medical College. He did his Internal Medicine residency at the Lahey Clinic and his NIH/NLM sponsored fellowship in Medical Informatics at Harvard Medical School and the Massachusetts General Hospital. Dr. Elkin has been working in Biomedical Informatics since 1981 and has been actively researching health data representation since 1987. He is the primary author of the American National Standards Institute's (ANSI) national standard on Quality Indicators for Controlled Health Vocabularies ASTM E2087, which has also been approved by ISO TC 215 as a Technical Specification (TS17117). He has chaired Health and Human Service's HITSP Technical Committee on Population Health. Dr. Elkin served as the co-chair of the AHIC Transition Planning Group. Dr. Elkin is a Master of the American College of Physicians and a Fellow of the American College of Medical Informatics. Dr. Elkin chairs the International Medical Informatics Associations Working Group on Human Factors Engineering for Health Informatics. Dr. Elkin is the Editor of the Springer Informatics Textbook, Terminology and Terminological Systems. He was awarded the Mayo Department of Medicine's Laureate Award for 2005. Dr. Elkin is the index recipient of the Homer R. Warner award for outstanding contribution to the field of Medical Informatics.

Ross Koppel, PhD

Dr. Koppel is a Professor of Sociology at the University of Pennsylvania who is world renowned for his identification of unintended consequences of Health IT. He has published widely, is a dynamic speaker and is a strong advocate for safer healthcare through the appropriate implementation of tested technologies. He joins the panel to discuss quality in Informatics applications development and implementation. He has published numerous high impact articles and is a fellow of ACMI.

Patient health records (PHRs), patient access to their records/medical information: issues and challenges

Abstract

The AMIA ClS, ELSI, EVAL, and POI Working Groups, with collaboration from the Society for Participatory Medicine (SPM), propose an interactive panel to address issues and challenges regarding PHRs and patient access to their records and medical information. Four panelists will address the topic from their varied experiences and expertise as informaticists, clinicians, clinical staff, and an healthcare attorney. Rather than include a single panelist to represent the patient point of view, we have invited five engaged patients from SPM to participate in the audience, comment, ask questions, and react to and interact with the panelists and other audience members.

Introduction

The AMIA ClS, ELSI, EVAL, and POI Working Groups, with collaboration from the Society for Participatory Medicine (SPM), propose an interactive panel to address issues and challenges regarding PHRs and patient access to their records and medical information, which involve the complex and sometimes competing interests and needs of many stakeholders. Patient access to their medical information and engagement in their care through HIT is a critical topic because it is an objective of Meaningful Use (MU). Four panelists will address the topic from their varied experiences and expertise as informaticists, clinicians, clinical staff, and an healthcare attorney. Rather than include a single panelist to represent the patient point of view, we have invited five engaged patients from SPM to participate in the audience, comment, ask questions, and react to and interact with the panelists and other audience members. The moderator will introduce them as audience-participants and ensure that they and our more traditional participants have ample time for interaction with each other and our panelists. Each presenter will take 8 minutes to discuss questions/points from those posed below, and as a springboard for discussion, take a position on what is most crucial to happen to achieve the following goals: optimum patient engagement, the most productive patient-provider relationships, and improved outcomes and satisfaction. The objective for the intended audience -- all interested members and patients -- is to gain understanding from many perspectives, and despite varying interests, needs, and existing and ongoing challenges, find ways to collaborate to achieve these mutually beneficial goals.

1) How can access to medical records be improved and are PHRs and electronic delivery the best or only (de facto) option for empowered patient engagement (e.g. How far down the PHR push are we? What about socioeconomic disparities that pose barriers to computer access; varying patient literacy/numeracy; patient preferences re level of engagement and electronic format?)
2) How can PHRs/portals be used to successfully engage patients in their own care -- what do and don't patients want, need, what concerns them, and how do we know? (e.g. Will electronic forms of patient engagement improve outcomes, or might they perpetuate or exacerbate health disparities - how will we all know?)
3) What are clinician concerns regarding PHRs and patient access to their medical information? (e.g. Ethical, legal, impact +/- on patient-clinician relationship, workflow issues, clinical team communication, etc.)
4) What are institutional, regulatory, and legal concerns/issues (e.g. Operational, required resources, technology options, HIPAA, Meaningful Use, what information can, should or must be kept out of medical records)?
5) How can the comprehensibility and actionability of medical data be improved?
6) What are key areas related to PHRs and access that need to be evaluated, and what evaluation methods will be needed to assess the effects of patient access to medical records and portals?

Panel members (All listed panel members and engaged-patient participants have agreed to participate.)

1) Catherine K. Craven, MLS, MA (Panel Organizer and Moderator), Chair, EVAL WG, is a doctoral candidate in Health (clinical) Informatics and a past NLM Informatics Fellow at the MU Informatics Institute, University of Missouri, Columbia, Missouri, USA. Her research covers implementation processes in small, rural Critical Access Hospitals; technology impact on ICU clinical team communication; and she is a member of Univ. of Missouri's CMS Health Care Innovations Challenge award team to leverage HIT to improve primary care coordination and patient engagement for better patient and population health, improved care, and cost outcomes.

2) Joseph Kannry, MD, Chair, CIS WG, is Professor of Medicine, a practicing board-certified Internist, and Lead Technical Informaticist for the Mount Sinai Health System, New York, New York, USA. In 2004, he led Mount Sinai's Ambulatory EMR Selection. Since 2005, he has been the Lead Technical Informaticist for the EMR Clinical Transformation Group. He oversees Ambulatory EMR implementation for the Hospital Based Practices and Faculty Practice Associates, PHR implementation, Enterprise CDS, and assisted with Inpatient implementation. Mount Sinai received the 2013 Davies Award for Enterprise EHR.

3) Jessica Ancker, MPH, PhD, Chair-elect, EVAL WG, Assistant Professor at Weill Cornell Medical College, New York, New York, USA, studies evaluation challenges in HIT, numeracy and decision-making issues among patients and providers, and has published on patient portal use among underserved populations. She is the recipient of an AHRQ K award to study patient portals and patient information needs.

4) Paul R. DeMuro, CPA, MBA (Finance), MBI, JD, FHFMA, FACMPE, CHC, is Co-Chair, Healthcare Information & Technology Group, Schwabe, Williamson & Wyatt, and an NLM Post-Doctoral Fellow in the Biomedical Informatics PhD program, Clinical Informatics, Dept. of Medical Informatics and Clinical Epidemiology, School of Medicine, OHSU, Portland, Oregon, USA. With 35 years of health law experience, Paul is a legal architect who develops clinically integrated models with aligned incentives using informatics technologies. Patient access to medical information is a key component of such models.

5) Carolyn Petersen, Chair, ELSI WG, is a Patient Advocate for NCI's Cancer Bioinformatics Grid since 2009, a Consumer Representative to FDA advisory panels since 2002, a Patient Stakeholder Reviewer for PCORI (Advancing Research Methods area) since they started making awards in 2012, and a managing editor of Mayo Clinic's (Rochester, Minnesota, USA) internationally renowned Web site since 2000. She's been a survivor of a pediatric cancer for more than 30 years, a driving reason for her career path and her advocacy work.

Engaged-patient audience participants from the Society for Participatory Medicine

1) Donna Cryer, JD, is an engaged patient herself as an IBD and liver transplant patient. Among many advocacy activities, she serves as a patient representative to the U.S. FDA, a merit reviewer for the Patient Centered Outcomes Research Institute (PCORI), and a member of the Stakeholder Advisory Group to the NIH Learning Health System Research Collaboratory. Most recently, Donna served as Chair, President, and CEO of the American Liver Foundation, the first patient to lead the organization in its 36-year history.

2) Gayle Embt, was caregiver for a parent with Alzheimers and multiple chronic conditions and is a caregiver for a child with complex needs, including pediatric mental health issues. She lives with the impact that the lack of understood expectations, technology enabled communications, and consistency of data sharing, has on the quality and accuracy of the care being delivered. Gayle co-chairs the Patient Experience Council at the Sullivan Institute for healthcare Innovation, served as a caregiver reviewer on the 2013 WEDI roadmap for data exchange.

3) Kym Martin, MBA, CNC, CFT, is an engaged patient as a 30-year, three-time cancer survivor. She is an appointed Co-Chair of the Patient Experience Council affiliated with The Sullivan Institute for Healthcare Innovation and Workgroup for Electronic Data Interchange (WEDI). Her major advocacy effort revolves around the need to shift focus to the overall patient *experience* as a key pathway to sustainable patient engagement.

4) Anna McCollister-Slipp builds platforms for better understanding of and engagement with the needs of patients, which is rooted in her personal experiences living with type 1 diabetes for 28 years. Anna was a member of the ONC HIT Policy Committee's FDASIA Workgroup, charged with advising the government on a regulatory pathway for HIT that would protect patients and promote innovation. As a co-founder of Calileo Analytics, she was named a "Woman to Watch" at Health Datapalooza and was invited to participate in "The Hive" at TEDMED 2013.

5) Mary Anne Sterling has 18 years of experience as a healthcare navigator for aging parents; three of her and her husband's parents suffer from or have died of forms of dementia. Among many engaged caregiver activities, Mary Anne serves on the ONC HIT Policy Committee's Consumer Empowerment Workgroup, and is Alzheimer's Association Ambassador to Sen. Mark Warner's office for implementation of the National Alzheimer's Project Act.

Brief description of panelist presentations

J. Kannry: Will start with what is a PHR, the difference between tethered vs. non-tethered PHRs, and discuss, via key studies, why this matters in the context of Meaningful Use (MU). He will review how MU Stage 2 and beyond has outstripped the evidence in the peer-reviewed literature [1, 2] Because MU is a huge driver of patient engagement through HIT, which is one of 6 listed objectives of the entire MU program, he will address the following questions: 1) Do we have any Informatics evidence that supports meaningfully used PHRs? [3-5]; 2) Based on this evidence, what do we know about patient needs and wants?; 3) What does evidence say about clinician wants, needs, and concerns regarding PHRs? He'll discuss challenges for the commercial market,

institutional, provider and patient challenges, and explore what an ideal PHR might look like.

J. Ancker: Will highlight current literature about the prevalence and use of electronic patient portals and PHRs [6-7] and their efficacy and effectiveness for improving outcomes and behavior [8-9]. She will discuss disparities in usage on the basis of socioeconomic predictors and implications for their impact, and place these technologies in context of health behavior theories to suggest future directions for practice and research.

P. DeMuro: There is much discussion over who owns the medical record and medical information. [10] The US and other countries often have contrasting views, and whether the answer differs for medical records vs. PHRs. This section also will address what a patient can access, view, copy, and amend in his or her medical record [11], and what types of information can be kept out of the medical record, given the HITECH Act, for example, state law considerations and psychotherapy notes. [12] How one can get as much of his or her EHR as possible will be discussed, along with how components of an EHR provided to others might be monitored and/or tracked. [13]

C. Petersen: To be most effective, engagement efforts must meet patients "where they live" via technologies and approaches they will readily use to access care and manage their health. [14, 15] Patient portals are one approach but require patients to possess technical competency, language and health literacy skills, access to technology, and stability in living situation. [16,17] Many patients who would gain most from a stable, engaged relationship with healthcare providers (e.g., Medicare/Medicaid patients, people with cognitive/physical disabilities) are unable to experience the full benefit of such relationships. [18] This segment will discuss benefits and challenges associated with patient portal use and other patient-engagement approaches from the perspective of the patient, as well as technologies and approaches proven to facilitate effective patient engagement, e.g. patient portal, mobile phone interventions, and social media. Expectations as expressed by patients and in the literature will be presented.

References

[1] Kannry J, Beuria P, Wang E, Nissim J. Personal health records: meaningful use, but for whom? Mt Sinai J Med. 2012 Sep;79(5):593-602.
[2] Davis Giardina T, Menon S, Parrish DE, Sittig DF, Singh H. Patient access to medical records and healthcare outcomes: a systematic review. J Am Med Inform Assoc. 2013 October 23, 2013.
[3] Kaelber DC, Jha AK, Johnston D, Middleton B, Bates DW. A Research Agenda for Personal Health Records (PHRs). J Am Med Inform Assoc. 2008 November 1, 2008;15(6):729-36.
[4] Halamka JD, Mandl KD, Tang PC. Early Experiences with Personal Health Records. J Am Med Inform Assoc. 2008 January 1, 2008;15(1):1-7.
[5] Tang PC, Ash JS, Bates DW, Overhage JM, Sands DZ. Personal health records: definitions, benefits, and strategies for overcoming barriers to adoption. J Am Med Inform Assoc. 2006 Mar-Apr;13(2):121-6.
[6] Ancker JS, Barron Y, Rockoff M, Hauser D, Pichardo M, Sczerenczy A, Calman N. Use of an electronic patient portal among disadvantaged populations. JGIM 2011; 26(10): 1117-1123.
[7] Delbanco T, Walker J, Bell SK, Darer JD, Elmore JG, Farag N, Feldman HJ, Mejilla R, Ngo L, Ralston JD, Ross SE, Trivedi N, Vodicka E, Leveille SG. Inviting patients to read their doctors' notes: a quasi-experimental study and a look ahead. Ann Intern Med 2012; 157(7): 461-70.
[8] Ammenwerth E, Schnell-Inderst P, Hoerbst A. The impact of electronic patient portals on patient care: A systematic review of controlled trials. JMIR 2012; 14(6): e162.
[9] David Giardina T, Menon S, Parrish DE, Sittig DF, Singh H. Patient access to medical records and healthcare outcomes: a systematic review. J Am Med Inform Assoc.(epub ahead of print Oct. 23, 2013)
[10] Hall MA, Schulman KA. Ownership of medical information. JAMA. 2009 Mar 25;301(12)1282-84.
[11] 45 CFR 164.524, 526
[12] 45 CFR 164.508(a)(2)
[13] sec. 164.528 of the HIPAA Privacy Rule
[14] Lee EO, Emanuel EJ. Shared decision making to improve care and reduce costs. New Engl J Med 2013 Jan 3;368(1):6-8.
[15] Legare F, Witteman HO. Shared decision making: examining key elements and barriers to adoption into routine clinical practice. Health Aff (Millwood) 2013 Feb;32(2):276-84.
[16] Chakkalatal RJ, Kripalani S, Schlundt DG, Elasy TA, Osborn CY. Disparities in using technology to access health information: race versus health literacy. Diabetes Care 2014 Mar;37(3):353-4.
[17] Berkman ND, Sheridan SL, Donahue KE, Halpern DJ, Crotty K. Low health literacy and health outcomes: an updated systematic review. Ann Intern Med 2011 Jul 19;155(2):97-107.
[18] Henning-Smith C, McAlpine D, Shippee T, Priebe M. Delayed and unmet need for medical care among publicly insured adults with disabilities. Med Care 2013 Nov;51(11):1015-9.

Nursing Data to Support the C-CDA, eMeasures, and Big Data Science - Ready or Not?

Connie W. Delaney, PhD, RN, FAAN, FACMI[1]; Gay Dolin, MS, RN[2]; Susan A. Matney, MSN, RN, PhD-C, FAAN[3]; Judith Warren, PhD, PhD, RN, FAAN, FACMI[4]; Bonnie L. Westra, PhD, RN, FAAN, FACMI[1]

[1]University of Minnesota, School of Nursing; [2]Intelligent Medical Objects; [3]3MHealth Information Systems; [4]Warren Associates, LLC

Abstract

Standardization of data and data structures are essential for sharable and comparable data to support the Consolidated Clinical Document Architecture (C-CDA), development of eMeasures (quality measure from electronic health records (EHRs)) for meaningful use of EHRs, and reuse of the data for big data science. Data standards and eMeasures are prescribed by the Federal government – LOINC for Assessments, and SNOMED-CT for nursing problems and interventions. Nursing has a long history of terminology development and mapping of nursing terminologies. This panel addresses the urgency of assuring that nursing data is ready to support these efforts and to learn strategies form the audience to rapidly propel standardization of data and data structures to support national efforts.

Introduction

Under the Health Information Technology for Economic and Clinical Health (HITECH) Act, the Office of the National Coordinator (ONC) and the Centers for Medicare and Medicaid Services (CMS) provide funds to increase the use of EHRs and require providers and hospitals to meet meaningful use of the EHR criteria. The Health Information Technology Policy Committee developed recommendations on the assignment of code sets to clinical concepts [data elements] for use in quality measures. Data must be extracted from EHRs as defined by the Quality Data Model (QDM) and expressed with the eMeasure format (AKA the Health Quality Measures Format (HQMF)). The results must be reported in the Quality Reporting Document Architecture (QRDA) I or III. The QRDA I leverages the templates in C-CDA and adds further requirements necessary for expressing Clinical Quality Measures (CQM) data. To support continuity of care, EHR technology must be able to electronically receive transition of care/referral summaries in accordance with the Consolidated Clinical Document Architecture (C-CDA). Simultaneously, to achieve the triple aim of better care, health and efficiencies in care delivery, there is an increasing emphasis on interprofessional practice that patient-centered and uses evidence-based practice (EBP) guidelines. Structured and standardized EHR is foundational to evaluate outcomes of care as well as discover new knowledge through comparative effective effectiveness research.

The American Nurses Association has recognized 12 terminologies and data sets; however, the HITECH requirements for the C-CDA and eMeasures requires data to be stored as Logical Observation Identifiers Names and Codes (LOINC) codes for assessment tools and Systemized Nomenclature of Medicine Clinical Terms (SNOMED CT) codes for nursing problems, clinical findings, and procedures (interventions) The major issue is whether nursing is ready to share data via the C-CDA and demonstrate quality with eMeasures using these standards. This panel addresses the urgency of assuring that nursing data is ready to support these efforts and to learn strategies form the audience to rapidly propel standardization of data and data structures to support national efforts.

Objectives

This panel will provide brief presentations and engage the audience in discussion to:

1. Describe the drivers for standardization of nursing and other health professional's data.
2. Explore the relationship of standardized nursing data to the C-CDA and specifically the Care Plan.
3. Relate the requirements for nursing data for eMeasures.
4. Examine initial efforts to create a nursing problem list to compliment the Core Problem list of meaningful use of EHRs using SNOMED-CT to bridge disparate nursing terminologiesl
5. Discuss the current status of nursing assessments included in LOINC
6. Contribute to an action plan to propel use of standardized nursing terminologies to support federal requirements to achieve the triple aim.

Moderator: Connie W. Delaney, PhD, RN, FAAN, FACMI

Dr. Delaney will provide a brief overview of national efforts that demonstrate the need for standardized data and data structures. She will address the synergy of the "Big Data" agenda that includes the Clinical

Translational Science Awards (CTSA) funded by the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health, the Patient Center Outcomes Research Institute (PCORI) goals, the National Center for Interprofessional Practice & Education, and how this synergy relates to standardization of nursing data and structures to achieve the Triple Aim.

Question(s):

As this synergy and integration occurs, what transformation of scientist training must occur?

1. Will knowledge representation of our interprofessional practice be accommodated by disciplinary specific representations?
2. What implications does the integration have for EHR design?
3. As this synergy and integration occurs, what transformation of scientist training must occur?
4. How will the practice models change?

eMeasure Development: Judith J. Warren, PhD, RN, FAAN, FACMI

The evolution of quality indicators, manually extracted to eMeasures electronically extracted, reveals issues concerning the implementation of data standards and the need for data integrity. The challenge is that every hospital has implemented their own documentation system so that data queried from one hospital is not in the same format, or even data type, as another hospital, and thus not comparable. Furthermore, not all implemented EHRs use the ONC/CMS mandated terminologies. eMeasure developers must identify data elements that capture the essence of the quality indicator without a human involvement to validate the comparability of the clinical data. Second, many members of the health care team document on the same data elements. When these do not coincide, the source of truth needs to be identified for the query. Below are a series of questions that are generated as eMeasures are developed and tested by one quality indicator developer.

Question(s):

1. What must be in place to have semantic interoperability? How can data standards facilitate the extraction and comparison of clinical data across clinical sites?
2. If terminology standards and maps are established, where would they be posted for all to use? Do we need agreed upon approaches to turn evidence-based practice protocols into data collection strategies? Who would do this?
3. How is the logic developed to generate the indicator? If anyone, who should endorse the logic set?
4. Where should maps and the logic set for eMeasure be posted to increase consistency?
5. How would eMeasures be collected: queries, clinical decision support rules, or another innovative strategy? How would the eMeasure data be processed and used in benchmarking—do we need quality indicator developers to do this or another type of vendor?
6. Then how do we know that the data extracted by the queries is correct and has integrity? As you look at data pulled from a query, you don't always see what is expected. Data is charted in the wrong place or put in a comment (thus unseen by a query).

Standardized Data Requirements for the C-CDA: Gay Dolin, MSN, RN

Gay Dolin will present a brief overview of the new Care Plan document type in C-CDA R2 focusing on how the HL7 CDA model together with standardized vocabularies, very effectively express the relationship between Care Plan components. A diagram will be shown to demonstrate the components of a Care Plan and the flow between them as expressed by HL7 CDA model. We will discuss how, through the syntax provided by the HL7 CDA model, refined though templates in the C-CDA R2 Care Plan, together with the semantics provided by controlled vocabularies, can fully express not only the requirements of a nursing Care Plan, but also a complete instance of an interdisciplinary patient centered care plan at a point in time.

Question(s):

1. The importance of syntax to fully express a concept using standardized vocabulary. E.g. The observation, diagnosis, intervention was made by who in what role.
2. A Care Plan is a living, changing plan. A CDA instance is a static expression of that plan. What is the value of this?

SNOMED CT Nursing Problem List: Susan Matney, RN-BC, MS, PhD-C, FAAN

The problem list is a key component of patient care and has been acknowledged as critical by the EHR Meaningful Use criteria. Specifically, SNOMED CT has been mandated as the terminology to use when storing

coded problems in the problem observation section of the C-CDA. Nursing diagnoses (problems) on the problem list are foundational for constructing a comprehensive care plan. A multidisciplinary patient problem list will facilitate communication and evaluation of the contribution of nursing care to the patient's clinical care experiences and outcomes. A nursing problem list subset of SNOMED CT has been developed by examining cross mappings between SNOMED CT and other nursing terminologies. The subset contains a complete list of all SNOMED CT concepts relevant for nurses to code a nursing diagnosis, is freely available from the National Library of Medicine website. Making use of this subset will decrease terminology development effort, reduce inconsistency and facilitate sharing of problem list data. A process for ongoing maintenance and continued work with UMLS in mapping nursing diagnoses concepts to SNOMED CT is in place to ensure that all ongoing concepts are proposed to address this limitation in future nursing problem subsets.

Question(s):

1. What is the current state of using the SNOMED CT problem list for exchange of data across multiple nursing terminologies?
2. What actions are needed to continue the development of SNOMED to fully represent the domain of nursing?

Nursing Assessments in LOINC: Bonnie L. Westra, PhD, RN, FAAN, FACMI

LOINC is typically considered a standardized coding system for laboratory data, however, it also include other domains such as document names, nursing management data, and clinical assessments as well as outcome ratings. LOINC includes coded assessment data as single measures such as blood pressure as well as panels and surveys such as vital signs, the Braden Scale, or the Minimum Data Set used in nursing homes. To evaluate the coverage of nursing assessments, a "gold standard" is required to both determine what assessments are in LOINC and the additional assessments that are needed. A brief overview of this project will be presented and participants will be asked about strategies to assure adequate coverage of assessments.

Question(s):

1. What should be the "gold standard" to determine coverage of assessments for inclusion in LOINC?
2. How should assessments be organized to assure consistent coding across settings, vendors, and EHRs?
3. What resources are needed and available to support collaboration in order to expedite the work needed?

Conclusion

Standardization of nursing and other health professional terms are essential as well as the models for reuse of the data for quality and research. Progress has occurred over the past 40 years, but further work is needed. The discussion from the audience will provide additional insights to propel this work forward in the future.

References

Fung, K. W., McDonald, C., & Srinivasan, S. (2010). The Umls-Core Project: A Study of the Problem List Terminologies Used in Large Healthcare Institutions. *Journal of American Medical Informatics Association, 17*(6), 675-680.

Matney, S. A., Warren, J. J., Evans, J. L., Kim, T. Y., Coenen, A., & Auld, V. A. (2012). Development of the Nursing Problem List Subset of Snomed Ct®. *Journal of Biomedical Informatics, 45*(4), 683-688. doi: 10.1016/j.jbi.2011.12.003

Dolin, RH., Dolin, G., Gaunt, S., Gonzaga, Z., Marquard, B., et. al. (Ballotted 2013, published 2014) HL7 Implementation Guide for CDA® Release 2: Consolidated CDA Templates for Clinical Notes (US Realm) Ann Arbor, Mich.: Health Level Seven, Inc.

Dolin, G., Dolin RH., Gaunt, S., et al (2012) HL7 Implementation Guide for CDA® Release 2: Quality Reporting Document Architecture – Category I, DSTU Release 2(US Realm) Ann Arbor, Mich.: Health Level Seven, Inc.

Wood L , Dolin RH., Dolin, G., Gaunt, S. et al (2012) HL7 Implementation Guide for CDA® Release 2: Quality Reporting Document Architecture – Category III, DSTU Release 1(US Realm) Ann Arbor, Mich.: Health Level Seven, Inc.

Dolin, RH., Alschuler, L., Boone, K,. Beebe C., Schadow, G., Dolin (ne Giannone) G., et al. (2010) HL7 Version 3 Standard: Representation of the Health Quality Measures Format (eMeasure), Release 1 Ann Arbor, Mich.: Health Level Seven, Inc.

HIE Enablers: A Crucial Need for Care Coordination Communication Among Long-Term and Post-Acute Care Front Line Nursing and Other Staff

Jennie Harvell (Moderator)[1], Gregory L. Alexander PhD, RN, FAAN,[2] Colene M. Byrne, PhD[3], Michelle Dougherty, MA, RHIA[4]

[1]U.S. Department of Health & Human Services Office of the Assistant Secretary for Planning and Evaluation, Washington, DC [2]University of Missouri, Sinclair School of Nursing, Columbia, MO; [3]Westat, Cambridge, MA; [4]AHIMA Foundation, Chicago, IL

Statement of Participation: All presenters have participated in development of this proposal. All have confirmed their participation in this panel presentation should it be accepted.

Abstract

We propose a panel to explore and discuss the current conditions and enablers to support care coordination including transitions in Long-Term and Post-Acute Care (LTPAC). Our four expert panelists will provide position statements, based on their experience in current research and initiatives, about the state of HIT adoption and use of Health information exchange (HIE) to support LTPAC coordination. We will discuss policy levers that enable HIE, as well as barriers, such as the lagging HIT adoption by LTPAC providers and the status of HIE interoperability standards to support LTPAC. Our positions are grounded in important themes of Technology, Adoption, Communication, and Workflow, specific to LTPAC settings. A learning objective for persons participating in this panel discussion is to present a real-world view of LTPAC HIT and HIE adoption, and interactively explore opportunities to advance HIE with LTPAC providers, as senders and receivers of key patient information.

Intended Audience

The intended audience will cover a range of disciplines, including LTPAC staff, telehealth providers, health care administration, clinicians with knowledge of Health IT, Chief Medical Officers, Chief Medical Information Officers, Chief Nursing Information Officers, researchers (e.g., health IT, telehealth, home health, nursing home), and nurse informaticists. Some audience members may not have deep knowledge of LTPAC and HIE opportunities. Through panel discussions they will gain new knowledge about opportunities to support HIT and HIE adoption in LTPAC.

Introduction of the Topic

LTPAC providers play an important role in the U.S. healthcare system, providing care for 1.5 million frail elders. Over a third of all Medicare patients discharged from hospitals receive subsequent LTPAC services[1]. Annual costs for LTPAC services in the U.S. have reached over $200 billion, with 69% being paid by Medicare and Medicaid[2]. As the U.S. population ages, persons 65 and older are expected to triple to 1.5 billion by mid-century, the demand for LTPAC services is expected to increase[3].

HIE to support care coordination with LTPAC providers is a timely topic as HIE is expected to help achieve national health policy goals (e.g., reduce hospital readmissions) and transform delivery and payment systems with new, improved models of care coordination. Sharing and exchanging health information, particularly during transitions of care, can reduce medication errors and other adverse events associated with preventable hospitalizations[4,5,6]. Front line nursing home staff and their patients, who are composed mostly of nursing personnel, will benefit most from these exchanges.

Funding initiatives and incentives, such as those funded and authorized under the Patient Protection and Affordable Care Act[7] promote and highlight the importance of care coordination around transitions and shared care. These initiatives and incentives include federal, state, and private healthcare payment and integrated care delivery models such as Accountable Care Organizations (ACOs) and Patient-Centered Medical Homes (PCMHs). Some initiatives are designed to reduce LTPAC transfers to hospitals, including readmissions. In addition, the Health Information Technology for Economic and Clinical Health (HITECH) Act enables development of the nationwide health IT infrastructure that will allow for the electronic use and exchange of health information.

Use of EHR and HIE technology by LTPAC lags behind other sectors despite the benefits and drivers of HIE[8]. Barriers to HIT adoption are well-documented and include differences in clinical processes and information needs between LTPAC and other healthcare providers, costs and limited resources for LTPAC to adopt HIT, and not being eligible for Medicare and Medicaid EHR Incentive Programs. Further, LTPAC EHR solutions are generally outdated and do not support the use of health IT standards expected to enable efficient interoperable HIE.

Despite these barriers, LTPAC providers are participating in initiatives to adopt HIT and exchange information[9] and early studies are showing great promise[10].The panelists' position is that with recognized enablers, such as the right stakeholder groups, use cases, and infrastructure, LTPAC providers want and will adopt and use HIE. In the absence of these enablers, LTPAC providers are developing other ways of exchanging, which are less efficient, not interoperable, and may introduce data security issues. Given this real-world understanding of the state of HIE to support LTPAC, we will engage in discussions around how to move forward and advance interoperable HIE.

Contribution of Speakers

Jennie Harvell will open and moderate this panel discussion. Ms. Harvell is a nationally recognized LTPAC policy and Health IT standards expert, particularly in the area of EHR/HIE initiatives. Ms. Harvell will offer her assessment of the state of HIE to support LTPAC, and provide an overview of key initiatives and strategies to advance HIE in LTPAC. She will discuss current data interoperability and information exchange standards to support common LTPAC HIE care coordination use cases.

Dr. Byrne will present findings from an environmental scan, literature review and key informant interviews on the state of HIE (electronic, other means) to support care coordination for persons receiving LTPAC. She will describe and characterize HIE to support LTPAC, using frameworks to characterize the health IT around HIE. The frameworks capture the care coordination workflow that provide opportunities and use cases for HIE. These frameworks also capture the entities involved, the context and drivers for HIE, types of information exchanged, technology used, and HIE outcomes.

Dr. Alexander's position is that strategies for HIT and HIE adoption in LTPAC must build on existing communication networks to support care coordination and transitions. Strategies include identifying required infrastructure to support HIT and HIE, financial incentives to encourage adoption and networking, and building stakeholder partnerships and opportunities for collaboration among healthcare providers. He draws on preliminary findings from an AHRQ-funded 4-year study about the relationship of HIT adoption to nursing home quality measures[11], findings from a CMS Innovations demonstration project on care coordination, communication, and workflow in nursing homes using HIE technology, and working with hospitals to reduce hospital readmissions[8].

Ms. Dougherty's position is that exchange is a common practice in LTPAC using various methods including electronic HIE through an HIE organization, but that interoperable solutions are relatively non-existent and/or don't address clinical processes relevant to LTPAC. She will include case study results to illustrate how technology is being used to support shared care and care transitions between LTPAC and their clinical partners, including which care staff are using HIE and how, the types of information that are exchanged, and the means of exchange. Ms. Dougherty will identify the gaps in technical approaches to support exchange (e.g., to support medication reconciliation), challenges for engaging LTPAC in HIE (lack of needed clinical information available through an HIE organization to support processes), and opportunities to accelerate interoperable HIE by LTPAC providers.

Expected Discussion

Ms. Harvell will synthesize the panelists' positions and then lead an interactive discussion around identified themes. Expected discussion includes:

- Important barriers that need to be overcome so that interoperable HIE can be successfully used by LTPAC providers to support care coordination and safe transitions.
- Whether financial incentives are sufficient to break through longstanding LTPAC/medical care boundaries that have impeded information exchange.
- LTPAC industry's readiness, benefits and risks for participating in electronic/interoperable HIE.
- Feedback on the opportunities identified by the panelists to accelerate interoperable HIE by LTPAC and if there are other opportunities.

- How the AMIA community can accelerate the development of tools and systems to support HIE and build the evidence base on the effectiveness of HIE interventions in LTPAC.
- Whether the identified barriers to 21st century information exchange can be mitigated by effective integrated care models, and as these models continue to evolve, what elements must be put into place to support full participation of LTPAC as key HIE trading partners in integrated delivery systems.
- How the AMIA community can help better define and validate measures that reflect HIE, electronic and other means, to support care coordination across providers, including LTPAC providers.

References

1. Dougherty M, Harvell J. Opportunities for engaging long-term and post- acute care providers in health information exchange activities: exchanging interoperable patient assessment information. Washington, DC: U.S. Department of Health and Human Services Assistant Secretary for Planning and Evaluation/Office of Disability, Aging and Long-Term Care Policy; 2011.
2. U.S. Department of Health and Human Services. National Clearinghouse for Long Term Care Information. 2010. Available at http://longtermcare.gov/costs-how-to-pay/costs-of-care/ Accessed May 22, 2011.
3. Pew Research Center (2014). Attitudes about Aging: A Global Perspective. Available at www.pewresearch.org.
4. Chhabra PT, Rattinger GB, Dutcher SK, et al. Medication reconciliation during the transition to and from long-term care settings: a systemic review. Res Soc Admin Pharm 2012;8(1):60.
5. LaMantia MA, Scheunemann LP, Viera AJ, et al. Interventions to improve transitional care between nursing homes and hospitals: a systemic review. J Am Geriatr Soc 2010;58(4):777-782.
6. Metzger J. Preventing hospital readmissions: the first test case for continuity of care. Falls Church, VA: Computer Sciences Solutions Global Institute for Emerging Healthcare Practices; 2012.
7. Affordable Care Act. National pilot program on payment bundling; Public Law 111-148 and 111-152, 2010:Section 3023.
8. Dougherty M, Williams M, Millenson M, and Harvell, J. EHR payment incentives for providers ineligible for payment incentives and other funding study. Washington, D.C. Prepared for Office of Disability, Aging and Long-Term Care Policy, U.S. Department of Health and Human Services; June, 2013. Available at http://aspe.hhs.gov/daltcp/reports/2013/EHRPI.pdf. Accessed July 14, 2013.
9. Byrne C, Dougherty M. Long-Term and Post-Acute Care Providers Engaged in Health Information Exchange: Final Report. Washington: D.C. Prepared for the Office of Disability, Aging and Long-Term Care Policy, U.S. Department of Health and Human Service; October, 2013. Available at http://aspe.hhs.gov/daltcp/reports/2013/HIEengage.shtml Accessed March 1, 2014.
10. Rantz MJ, Alexander GL, Galambos C, et al. Initiative to test a multidisciplinary model with advanced practice nurses to reduce avoidable hospitalizations among nursing facility residents. J Nurs Care Qual. 2014;29(1):1-8.
11. Alexander GL: A National Report of Nursing Home Quality Measures and Information Technology 1R01HS022497-01: Agency for Healthcare Research and Quality; 2013.

Data Governance Dilemmas for Research and Clinical Care

Bonnie Kaplan, PhD, FACMI, Yale University, New Haven, CT; Paul R DeMuro, CPA, MBA, MBI, JD, Oregon Health & Sciences University, Portland, OR; Frank Pasquale, JD, MPhil, University of Maryland, Baltimore, MD; Jan Talmon, PhD, FACMI, Maastricht University, The Netherlands; Peter Winkelstein, MD, MS, MBA, FAAP, University at Buffalo, Buffalo, NY

Abstract

Legal and ethical considerations distinguish between patient care and research. Data collection, protection, privacy, and permissions are governed differently for routine care and research data. Yet, neither patients nor clinicians can so easily separate the two; the distinction is not as clear-cut as it appears. Today's clinical data becomes tomorrow's research data. New technologies, medical advances, and the move toward learning health care systems further blur the distinction. Panelists will discuss and debate how the clinical/research distinction holds up in practice, and whether it should (or can be) refined or eliminated. They address dilemmas that informaticians, patients, clinicians, and policy makers face because of the distinction between research and clinical care as it plays out in electronic record keeping, data quality, data storage, data sharing, data linking, and secondary use. The panel considers such timely and controversial issues as: Exactly which aspects of care, and the documentation of that care, are research, and which are clinical, when many may be both? Should consent procedures differ according to whether treatment is considered research? Should patient care be contingent on permission to use care data for research? Are quality assurance, biobanks, or learning systems data clinical data or research data? When is a record review quality improvement and when is it research? What are the benefits and costs of maintaining the status quo and of proposals for change? How do different approaches compare internationally? Panelists and audience will exchange ideas about how they might like to see regulations evolve.

Legal and ethical considerations distinguish between routine patient care and research. Data collection, protection, privacy, and permissions are governed differently for routine care and research data. Yet, neither patients nor clinicians can so easily separate routine care from research. The legal and ethical distinction is not as clear-cut as it appears.

As one example, the Institute of Medicine's call for "learning health care systems" has focused attention on the need to use individual patient data collected in the course of clinical practice to improve health care in general. In a learning health care system, patient data is used for continuous quality improvement and to produce knowledge useful throughout the health care environment. These goals for learning health care systems render obsolete the distinction between data use for patient care and data use for research. Some authors controversially advocate requiring patients to allow their data to be used for research, and some institutions routinely make this requirement a condition of care. Further, as observational studies are increasingly replacing randomized controlled trials, patient care data also is research data. What is the purpose of the clinical/research distinction when today's clinical data becomes tomorrow's research data?

Newer developments may make the clinical/research distinction even more problematic and relevant for informaticians. Data shared through health information exchanges (HIEs) and the US National Health Information Network (NHIN) are a natural basis for research studies. Genomic and clinical data are needed for research to develop personalized targeted therapies. Biobanks collect specimens and related data from clinical procedures, data that later may be used for research. Is all this research data or clinical data? Which regulations apply? Data generated from experimental medical devices, mobile health devices, and sensors that broadcast physiological readings and other health indicators for inclusion in patient records, and patients' self-reports, present additional considerations. What might be learned from data governance approaches in different countries and how would international clinical and research collaborations be affected by governance differences?

Panelists will discuss and debate how the clinical/research distinction holds up in practice, and whether it should (or can be) refined or eliminated. They address dilemmas that informaticians, patients, clinicians, and policy makers face because of the distinction between research and clinical care as it plays

out in electronic record keeping, data quality, data storage, data sharing, data linking, and secondary use. Each panelist will outline areas where the clinical/research distinction creates legal, clinical, research, policy, or ethical difficulties and propose possible solutions. Bonnie Kaplan, past chair of the American Medical Informatics Association's Ethical, Legal, and Social Issues Working Group, will introduce the issues and moderate the panel. Health law expert Paul De Muro begins by reviewing laws and regulations that govern each area, pointing to overlapping and divergent areas that are problematic in their application, and to needed changes in light of technological developments. CMIO Peter Winkelstein, Executive Director of the University at Buffalo Institute for Healthcare Informatics, presents difficulties these distinctions create for clinicians, researchers, Institutional Review Board members, and HIT institutional policy committees, based on his experience in all these roles. Jan Talmon, a medical informatician from the Netherlands and past co-chair of the AMIA Evaluation Working Group, will present the approach of the Dutch Academic Centers to collectively establish prospective biobanks for which data are collected as much as possible during the clinical encounters. Health law privacy scholar Frank Pasquale explores what becomes of patient record (clinical or research) data, focusing on consequences of the distinction for social benefit and for patient privacy. He discusses opportunities (and perils) of new forms of health information exchange and addresses the role of regulators in trying to improve the health care innovation environment. The panel thus addresses such timely and controversial issues as: Exactly which aspects of care, and the documentation of that care, are research, and which are clinical, when many may be both? Should consent procedures differ according to whether treatment is considered research? Should patient care be contingent on permission to use care data for research? Are quality assurance, biobanks, or learning systems data clinical data or research data? When is a record review quality improvement and when is it research? What are the benefits and costs of maintaining the status quo and of proposals for change?

After brief panelist presentations, panelists and audience will exchange opinions on data governance and ideas about how they meet regulatory requirements and how they might like to see regulations evolve.

Learning objectives

After participating in this session, participants should be better able to:

1. distinguish between clinical and research health care data based on legal, ethical, and regulatory distinctions;
2. resolve dilemmas created by legal, ethical, and regulatory distinctions between clinical and research health care data; and
3. evaluate implications of new technologies and medical advances in terms of legal, ethical, and regulatory distinctions between clinical and research health care data.

Intended Audience:

The session is relevant to all American Medical Informatics Association Symposium attendees involved in collecting, distributing, and analyzing health care data; regulatory issues; Institutional Review Boards (IRBs); Quality Improvement (QI); data and biobanking registries and repositories; Health Information Exchanges (HIEs); data science and analytics; and patient consenting.

PANELISTS

Paul R. DeMuro, CPA, MBA (Finance), MBI (Biomedical Informatics), JD, FHFMA, FACMPE, CHC National Library of Medicine Post-Doctoral Fellow in the PhD program in Biomedical Informatics, Clinical Informatics Track, at Oregon Health & Science University School of Medicine Department of Medical Informatics and Clinical Epidemiology; Co-Chair, Healthcare Information & Technology Group, Schwabe, Williamson & Wyatt. In his 35th year of practicing health law, Paul is a legal architect who designs clinically integrated health system models using informatics for aligned financial incentives. Paul's main area of research interest is the use of informatics tools to improve quality and cost-effectiveness. Paul is a former Chair of the American Bar Association Health Law Section. He will address such issues as: the current state of the law and regulations governing data collection, privacy, and security; and how they have not kept up with advances in care and research. He also will address special issues involved in the use of mobile medical applications, mHealth, and social media in the context of the issues which will be discussed by the other panelists.

Bonnie Kaplan, PhD, FACMI, of the Yale Center for Medical Informatics, is a Yale Interdisciplinary Bioethics Center Scholar, and a Faculty Fellow of the Yale Law School's Information Society Project. She has long experience in evaluating effects of health information technology on people and organizations. She is a past chair of the American Medical Informatics Association's (AMIA) Ethical, Legal, and Social Issues Working Group; AMIA's People Organizational Issues Working Group; and the International Medical Informatics Association's (IMIA) Organizational and Social Issues Working Group. Her research addresses informatics ethical, legal, and privacy issues; user perspectives and experiences with health information technology; and ethnographic sociotechnical evaluation. She is a Fellow of the American College of Medical Informatics, and currently a Hastings Center Visiting Scholar. She will outline the issues and advocate for simpler patient consent practices that are the same for both research and clinical data use.

Frank Pasquale**,** JD, MPhil, is Professor of Law at the University of Maryland and Affiliate Fellow at Yale Law School's Information Society Project. His article "Grand Bargains for Big Data" (*Maryland Law Review,* 2013) articulates a broad policy framework for health data governance. His book *The Black Box Society: The Hidden Algorithms Behind Money and Information* (Harvard University Press, 2014) offers a broader perspective on data practices. His research agenda focuses on challenges posed to information law by rapidly changing technology, particularly in the health care, Internet, and finance industries. Pasquale has been a Visiting Fellow at Princeton's Center for Information Technology, a Visiting Professor at Yale Law School and Cardozo Law School, and the Schering-Plough Professor in Health Care Regulation and Enforcement at Seton Hall Law School. He has presented before a Department of Health & Human Services/Federal Trade Commission Roundtable and panels of the National Academy of Sciences on new challenges raised by personal health records and pervasive sensor networks. He will address the benefits and pitfalls of regulatory changes and their effect on innovation in health information exchange.

Jan Talmon, PhD, FACMI, Emeritus Associate Professor in Medical Informatics at the Maastricht University, Maastricht, The Netherlands, is a member of the central team of the Parelsnoer Institute (String of Pearls Institute, PSI), a collaboration among the eight academic medical centers in the Netherlands for prospective biobanking. He is past co-chair of the American Medical Informatics Association's (AMIA) Evaluation Working Group, he is co-editor of the *International Journal of Medical Informatics*, and is a Fellow of the American College of Medical Informatics. He will address the governance model of the PSI and the approach taken to address the legal and ethical issues related to the use of clinical data for clinical research. Responsible for PSI's data model, he will describe the different disease-dependent approaches in which some institutions have fully harmonized the data collected in practice and in care, while others have add-ons for research purposes, but all collect data for clinical audit which overlaps with what is collected for research purposes. He will relate this approach to the more general rules of conduct that govern secondary use of health data in the Netherlands.

Peter Winkelstein**,** MD, MS, MBA, FAAP serves as Executive Director of the University at Buffalo Institute for Healthcare Informatics (IHI) and as chief medical informatics officer (CMIO) of both UBMD, the multispecialty faculty practice of UB physicians, and Kaleida Health, the largest healthcare provider in Western New York. He is a physician executive with extensive experience in medical management and medical informatics, including big data and data science. Dr. Winkelstein has served in multiple medical and informatics leadership roles and his background in clinical, administrative and technological medicine enables him to interface effectively with multiple healthcare environment stakeholders. The mission of the IHI is to support big data healthcare research and economic development at UB and its partners by providing a secure computing environment, data science and security consulting, and extensive healthcare datasets. As CMIO, Dr. Winkelstein sees the EHR as not only supporting clinical practice, but also as collecting clinical data that can be used for quality improvement and academic research. Dr. Winkelstein is Professor of Clinical Pediatrics at UB and a Certified Medical Staff Leader. He will address such questions as: How do you deal with the need to use repositories for both clinical/operational questions (HIPAA) and research (IRB)? How do you distinguish QA/QI (HIPAA) from research (IRB) and is such a distinction meaningful or useful (especially in the context of "learning health care systems")? When should explicit consent be required to use patient data for research—always, never, sometimes?

Towards Developing an Undergraduate Interprofessional Biomedical Informatics Course

Saif Khairat, PhD, MS[1], Martha B. Adams, MD, MA, FACP[2], Glynda Doyle, RN, MSN[3], Elaine Ayres, MS, RD[4], Tiffany F. Kelley, PhD, MBA, RN[5]

[1]University of Minnesota, Minneapolis, MN; [2]Duke University, Durham, NC; [3]British Columbia Institute of Technology, Vancouver, Canada; [4]National Institutes of Health, Washington, DC, [5]Nexus Consulting, Boston, MA

Participants

• Organizer, Moderator, and Panelist: Saif Khairat, PhD, MS, Institute for Health Informatics, University of Minnesota

• Panelist: Martha B Adams, MD, MA, FACP, Duke Center for Health Informatics, Duke University

• Panelist: Glynda Doyle, RN, MSN, British Columbia Institute of Technology

• Panelist: Elaine Ayres, MS, RD, NIH Clinical Center, National Institutes of Health (NIH)

• Panelist: Tiffany Kelley, PhD, MBA, RN, Nexus Consulting

Abstract

Biomedical Informatics training has expanded to include almost all clinical specialties. Whether it is a Masters, PhD, or even the new Sub-Specialty in Clinical Informatics, graduate students and/or returning professionals have various informatics training options. However, Informatics education at the undergraduate level is not as well established. Despite initiatives such as TIGER and the Health Information Technology Scholars Program there is still a need to provide interprofessional education. As Informatics continues to grow and healthcare becomes more complex and data driven, there is a growing need to introduce fundamental concepts to undergraduate students to enhance their knowledge base and stimulate interest in the career of biomedical informatics.

The panelists are leading efforts to develop a Biomedical Informatics course that addresses fundamental concepts and core competency skills. This panel will explore the needs and challenges to build an undergraduate course that includes core competencies for various clinical specialties including medicine, nursing, and allied professions such as nutrition. The proposed course will serve as a foundation for health professionals, which is different from the advanced ONC funded Health IT Workforce Curriculum Components. Panelists specifically aim to further understand the expectations of the AMIA community with regards to developing an interprofessional informatics course. This course will be designed to accommodate undergraduate students. This effort intends to provide fundamental knowledge of Biomedical Informatics to pre-med, allied health, and IT undergraduate students. The aim is to present students with a general overview of the role of Informatics in clinical, research, and operation practices and hence, facilitate determining future career directions.

Learning Objectives

• Better understand the needs for interprofessional undergraduate informatics courses

• Identify core undergraduate informatics competencies (general and specialty specific)

• Learn about on-going and future efforts to integrate informatics in undergraduate education

• Collect feedback on the initial structure of the course

• Identify and invite qualified Informatics professionals to content writing

• Provide a report summarizing audience opinions

Audience

Due to its high relevance, this session will be of interest to all AMIA members.

Topic controversy

The controversy of undergraduate informatics lies not in the topic, rather in the presentation, structure, and organization of an undergraduate course. A multidisciplinary field such as Informatics includes multidimensional views and opinions, and the integration of all opinions may be a challenge. Therefore, the debate remains to be identifying the main goals of the course and its core competencies. For that reason, panelists and the audience will discuss what needs to be included in an undergraduate Informatics course, and what is considered advanced knowledge and skills.

Panelist Presentations

Saif Khairat, PhD, MS has been involved with Health Informatics since 2006. He is a Clinical Assistant Professor at the Institute for Health Informatics at the University of Minnesota. Dr. Khairat is the Co-Principal Investigator to a federal grant award from the Health Resources and Services Administration (HRSA) entitled "Telehealth Resource Center Grant Program", at the University of Minnesota. Dr. Khairat earned his PhD in Health Informatics at the Informatics Institute at the University of Missouri with a focus on ICU clinical communication. During his Informatics training, Dr. Khairat worked as a Research Fellow at the Division of Clinical Informatics at Harvard Medical School. He also has track record of computer science training. Prior to pursuing a career in Health Informatics, Dr. Khairat has been involved with the design and development of Health IT systems. Dr. Khairat is lead author to numerous publications and serves as a scientific reviewer to national and international conferences and journals.

Dr. Khairat is the currently the Chair-Elect of the Education Working Group at AMIA, member of the AMIA Working Group Steering Committee, and he has served on Student Working Group committees. Among other goals, Dr. Khairat intends to develop an undergraduate Informatics course within his term period. Dr. Khairat also serves on the Clinical Informatics Board Review Course (CIBERC); he participates to the development of exam questions for the Simulated Exam.

Dr. Khairat is the Co-lead of the EHR Task Force at the University of Minnesota. Dr. Khairat developed a set of core competencies that aim at identifying and acquiring electronic health records and informatics systems tools around which a high quality curriculum can be developed to educate and train health professions students and informatics specialist students at all levels within the context of interprofessional education and collaborative practice.

Dr. Khairat will address the importance of undergraduate informatics education; he will provide an overview of current progress of the proposed course. Dr. Khairat will talk about the course structure, the need for highly trained informaticians to participate in content development, and the need for content reviewers. He will pose questions to the audience related to the need for an undergraduate informatics course.

Martha B. Adams, MD, MA, FACP is a member of the Duke Center for Health Informatics. She is an emeritus professor, a clinician in applied informatics (handheld technology, telegenetics, information security, social media), author of an enterprise-wide research data security plan for Duke and co-founder of an antimicrobial stewardship framework deployed at Duke University Hospital and 30 hospitals in the Netherlands; she is recently advisor to projects of continuous learning quality improvement and another in personalized medicine involving genomics and cardiovascular disease. Her leadership and informatics track record position her well in academia as former vice chair for clinical affairs in the Department of Medicine, member of the Curriculum Committee of the School of Medicine at Duke, and the University's Academic Council and Open Access Advisory Group. She continues membership in the AAMC Group on Information Resources, its Security Working Group.

Dr. Adams will provide insights from her own career and the challenges of implementing informatics education. She will submit provocative questions for the panelists and audience about solutions, all towards stimulating career interest and responding to the goal set by the Institute of Medicine that, "by the year 2020, at least 90% of clinical decisions will be supported by accurate, timely, and up-to-date information that reflects the best available evidence", a goal that requires more than technology, it requires an interdisciplinary approach of the science and how we use data.

Glynda Doyle. RN, MSN teaches at the British Columbia Institute of Technology (BCIT) in Vancouver, British Columbia. She completed her MSN at the University of British Columbia in 2011 where she first discovered her passion for health informatics. Ms. Doyle is focused on the integration of informatics into the BCIT Bachelor of

Science in Nursing (BSN) and Specialty Nursing curricula. She is particularly interested in the role of mobile technologies and their impact on nursing student's clinical judgment and decision making.

Ms. Doyle has many years of national and international experience in critical care in South Africa, the United Kingdom and the United States. She has been teaching at BCIT in the BSN program since 2005, and is dedicated to engaging students with stimulating and relevant educational environments. Although fairly new to the Health Informatics arena, Ms. Doyle is passionate about integrating Health Information Technologies and the Science of Informatics into nursing education to help improve patient safety and quality of care.

Ms. Doyle is a co-investigator in several inter-disciplinary research projects within BCIT and also in collaboration with other Canadian nursing schools studying the impact of mobile devices laden with clinical resources, social networks and e-portfolios on nursing students and their education.

Ms. Doyle intends to ensure that there is a strong nursing component to the development of this course and is aware of the extensive work already accomplished in the nursing field. She would like to share this information and knowledge with other professionals and provide support to educators who are not informaticians but who are tasked with the integration of informatics in their programs and courses. She would also like to see this course available not only in its entirety to educators, but also as subsections to support the integration of informatics into undergraduate courses such as Ethics, Professional Practice, Evidence Based Practice and Communication.

Ms. Doyle hopes to gather more information from the audience as to their experiences and knowledge of currently available resources and to hear their suggestions as to how best we can format this course and ensure its relevance to as many undergraduate programs as possible.

Elaine Ayres, MS, RD is a registered dietitian who through training and practical experience now works in the field of informatics. Ms. Ayres completed her undergraduate training in nutrition at Cornell University, and her master's in nutrition at the University of Maryland, College Park. She completed her dietetic internship at the Massachusetts General Hospital in Boston, MA. Trained as a research dietitian, Ms. Ayres was asked to manage a food and nutrition computer system at the NIH Clinical Center in 1992. This was in addition to her responsibilities as the Director of the NIH Dietetic Internship, a post-graduate didactic and experiential training program required for registered dietitians. The NIH Dietetic Internship has always ensured that students were well versed in informatics principles and practices as a result.

Ms. Ayres then spent a decade in hospital administration (1998-2008) learning the business and funding of large IT systems, including the current NIH Clinical Center electronic health record (CRIS). While involved in the implementation of CRIS, Ms. Ayres identified the need to involved dietitians in the selection and implementation of health care systems. Working closely with what is now the Academy of Nutrition and Dietetics; Ms Ayres chaired the Nutrition Informatics Committee and the Subcommittee on Interoperability and Standards. She was also the principle on the Delphi Study to develop nutrition informatics competencies. Ms. Ayres is now the Project Manager for the Biomedical Translational Research Information System (BTRIS) at the NIH Clinical Center.

Ms. Ayres will address how undergraduate program content in informatics can be used to enhance the curriculum of nutrition students, and how specific topics and competencies from the nutrition domain will serve to enhance the understanding and engagement of students in other disciplines. Audience suggestions on how to maximize the value of interdisciplinary undergraduate education in informatics will be solicited.

Dr. Tiffany Kelley, PhD, MBA, RN is a Registered Nurse for 11 years, currently; she is a Senior Consultant at Nexus Consulting and the founder of Nightingale Apps, a Health information technology company offering mobile applications to nurses in hospital settings. She earned a PhD in Nursing from Duke University with a focus on Health Informatics and an MBA from Northeastern University. She was the past chair of the Student Working Group and the elected Student Representative to AMIA Board of Directors. She has lead numerous AMIA committees looking to introduce various health profession students to Informatics.

Dr. Kelley will address the vision for an undergraduate informatics course that facilitates pursuing more advanced and specialized informatics training. She will shed light on the challenges novice students face in their post-graduate studies, as well as the challenges of today's professionals in industry and ways to improve training new informatics workforce. Dr. Kelley will talk about the demand for informatics professionals who may not possess graduate education to serve in various organizational roles.

All panelists agree to participation if the panel is accepted.

Innovative Approaches to Medication Reconciliation within the Veterans Health Administration: Designing the 'Magic Pill'

Blake J. Lesselroth, MD, MBI[1,2], Kathleen Adams, MPH[1], Steven R. Simon, MD, MPH[3,4], Kenneth S. Boockvar, MD, MS[5,6], Peter J. Kaboli, MD, MS[7,8]

[1]Portland VA Medical Center, Portland, OR; [2]Oregon Health Sciences University, Portland, OR; [3]VA Boston Healthcare System, Boston, MA; [4]Harvard Medical School, Boston, MA; [5]James J. Peters VA Medical Center, Bronx, New York; [6]Icahn School of Medicine at Mount Sinai, New York, New York; [7]Iowa City VA Medical Center, Iowa City, IA; [8]University of Iowa Carver College of Medicine, Iowa City, IA

Abstract

Medication discrepancies at interfaces-in-care are an important source of preventable iatrogenic injury, causing an estimated 1 million hospitalizations and 7,000 deaths in the US annually at an estimated cost of $500 million annually. While medication reconciliation (MR) has been described as an effective process to surface medication discrepancies and avoid adverse drug events, most institutions have struggled to implement durable MR interventions. The Veterans Health Administration (VHA) features a robust electronic health record and manages the entire medication distribution supply chain from provider order entry to home medication delivery. Consequently, VHA is well positioned to develop and study innovative strategies to manage MR. This panel, which includes clinician-informaticians and health services researchers from across the VHA enterprise, will describe an array of quality improvement initiatives designed to address MR throughout the care continuum. Panelists will discuss: 1) multimedia applications intended to improve clinician accuracy, 2) patient portals to improve patient engagement and self-efficacy, 3) the use of regional health-information exchanges to close information gaps, and 4) implementation studies designed to identify best practices. Panelists will also engage the audience in a dialog exploring the emerging data management and sociotechnical challenges that continue to besiege quality improvement initiatives.

Introduction and Problem Statement

Medication discrepancies at interfaces-in-care are an important source of preventable iatrogenic injury, causing an estimated 1 million hospitalizations and 7,000 deaths in the US annually at an estimated cost of $500 million US annually[1]. Medication reconciliation (MR) a standardized process intended to close gaps in medication information and avoid downstream preventable adverse drug events (PADE)[2]. Although implementations vary depending upon context, the archetypal approach describes four crucial steps including 1) collecting a medication history, 2) comparing the history to institutional records or current prescription orders, 3) identifying and reconciling unintended discrepancies, and 4) documenting and communicating actions and care plans to the next team, patient, or caregiver[3].

While the MR process is conceptually simple and has been shown in the literature to reduce discrepancies, most institutions have struggled to implement durable and highly reliable systems[4]. Even with published toolkits speaking to organizational transformation (e.g. Project Red, IHI's Medication Safety Reconciliation Toolkit, AHRQ's MATCH toolkit), clinicians and healthcare executives report difficulty implementing, acculturating, and sustaining successful MR programs[5]. Also, there are formidable sociotechnical and human factors challenges including cognitive overhead, organizational climate, and patient health literacy[6,7].

The Veterans Health Administration (VHA) has long been recognized as an informatics leader, using an electronic health record (EHR) with physician order capabilities for well over a decade[8]. However, its development largely predates the human factors and usability movement in healthcare. Unsurprisingly, the aging clinician interface does not suitably support the organic and opportunistic characteristics of MR. The VHA, and organizations in general, must model the operational and cognitive requirements of MR systems to inform tools that improve clinician performance and engage patients. Until clinicians and informaticians completely understand MR system requirements, recognize the failure modes that beset MR implementations, and consider the patient as a user in the design paradigm, developers will be doomed to commit the same design missteps and organizations will continue to wrestle with costly, poorly designed, and potentially ineffective tools.

Interactive Panel Aims and Description
The VHA has the medication data and technology infrastructure to prototype and pilot MR strategies addressing common implementation barriers and known failure modes. Presently, the VHA is engaged in a needs-assessment phase, testing technology prototypes, studying system performance, and building an evidence-based foundation to support enterprise-wide programs. This work figures prominently at a time where there are tightening medication safety accreditation standards for healthcare organizations and more stringent meaningful use expectations for EHR technologies. While many organizations emphasize compliance over intent, investing misplaced confidence in routine procedures, most front-line clinicians recognize MR is only partially effective and potentially toxic (like any new class of treatment). This symposium will present some ways in which improvements are being implemented and tested to get to "next generation" MR.

This panel assembles clinician-informaticians and health services researchers engaged in MR research and development across the VHA to discuss quality improvement initiatives for inpatient, outpatient, and home health settings. Each discussant will describe a major MR project within their portfolio, using each case study to highlight lessons learned and controversial informatics issues impacting implementation success. Approximately half of the time will be dedicated to an interactive question-and-answer format intended to engage audience members. Topics for discussion will include: 1) the hidden complexity of MR tasks, 2) the overlooked human factors requirements of MR, 3) the resource burden of strong MR systems, 4) the role of the patient as a participant in MR, 5) the weak cultural climate of implementation within many organizations and 6) the generalizability of work to non-federal systems. Both panelists and audience members will have an opportunity to ask questions.

Panelist 1: Blake J. Lesselroth MD, MBI is a teaching hospitalist, applied informatician, and director of a VHA Patient Safety Center of Inquiry (PSCI) – a research and development "clearinghouse" dedicated to the design, implementation, and evaluation of MR technologies. The Center recently completed the design and pilot of a mobile medication review software application. Pharmacists used the software, which pairs prescription data with images, to collect a bedside admission medication history. Human factors methods including user-centered design, simulation testing, and workflow analysis were used to evaluate the target activity system, the potential implementation barriers, and the anticipated effect upon hospital performance. He will describe the technology, the techniques used to measure an implementation, and the sociotechnical lessons learned from a pilot.

Panelist 2: Steven R. Simon, MD MS is the Chief of General Internal Medicine at the VA Boston and a health services researcher at the Center for Healthcare Organization and Implementation Research (CHOIR). His research emphasizes the evaluation of information technology interventions, including consumer informatics portals, to improve the quality and safety of healthcare. His most recent work has focused upon the design of patient-facing MR portals that can be integrated within the VHA's personal health record, MyHealtheVet. He will be discussing the challenges of using technology to improve home-bound patients' healthcare. He will focus upon barriers to patient technology adoption, the role of secure-messaging mediated MR, and future strategies for using patient-centered technologies to engaged patients and care-givers with medication adherence and monitoring.

Panelist 3: Kenneth S. Boockvar, MD MS is an academic geriatrician, primary care provider, and health services researcher at the James J Peters VA Medical Center in the Bronx, New York. He is also Associate Director for a regional VHA Geriatric Research, Education, and Clinical Center (GRECC). He has been principal investigator on two studies that used mixed methods to examine the implementation of inpatient MR, and health information exchange-enhanced discharge MR. He has published extensively on the topic of MR and was one of the first investigators to measure a direct and measurable relationship between medication discrepancies and adverse events. His VA is currently participating in the Bronx Regional Health Information Organization (RHIO) and he directed a VA project adapting inpatient MR processes to include RHIO access. He will describe the use of this VA-non-VA exchange to enhance MR, reflecting upon preliminary measures of effect, barriers to implementation, and observations of providers using the system. He will also forecast how VHA activities may support the transformation of larger regional health networks.

Panelist 4: Peter J. Kaboli, MD, MS is a teaching hospitalist and health services researcher with over a decade of experience directing and evaluating hospital-based quality improvement interventions. He is currently a co-investigator on an ambitious multi-center MR implantation study that also includes the Partners Healthcare System and Northwestern University. The goals of the Multi-Center Medication Reconciliation Quality Improvement Study (MARQUIS) are to operationalize MR best practices for the inpatient setting and evaluate the effect upon unintentional medication discrepancies. He will be discussing human factors lessons drawn from a discharge process re-engineering initiative using health record technologies and clinical pharmacists. He will also share

insights on the MARQUIS study, including study design, measurement decisions, current progress, and important opportunities for further research.

Moderator: Kathleen Adams, MPH is an informatics research manager and usability specialist at the Portland VA Medical Center. She has over 20 years of research experience in the areas clinical outcomes research, applied informatics, technology usability, and implementation science. A certified Clinical Research Professional, she serves as an Associate Director of the VA Clinical Research Alliance which promotes research best practices through education, mentorship, and outreach.

Intended Audience

The topics covered in this interactive panel discussion should appeal to AMIA members since MR is an interdisciplinary effort leveraging implementation science, healthcare quality, decision support systems, consumer informatics, human factors, and health data exchange. The panel is comprised of clinician-educators and health service researchers with expertise in requirements development, process implementation, and outcomes measurement.

A significant part of the session will be dedicated to an interactive dialog with the audience. Attendees will be able to pose questions to the panelists. We also hope to gather feedback and insights from the audience. Important questions to consider include:

- What components of MR are the most challenging to implement or measure?
- Do facilities and clinicians share an accepted set of definitions for MR activities?
- What are the most important steps informatics specialists can take to support broad adoption of MR
- What software or technologies do you think are needed to support MR?
- What information would increase clinician trust in the accuracy of recorded medication information?
- What are some of the unintended consequences of MR programs?
- What measures do you believe would be the most meaningful to support QI efforts?

Learning Objectives

After participating in this interactive panel discussion session, the learner should be better able to:

- Describe at least four informatics strategies the VHA is piloting to improve the quality of MR
- Identify at least two barriers that frequently undermine successful MR program implementations
- Propose how informatics can be used to improve patient engagement or self-efficacy with regard to MR
- Select metrics to evaluate MR system usability, deployment effectiveness, and tool performance or effect
- Identify at least three research gaps or development opportunities in the domain of MR informatics

References

1. Budnitz DS, Lovegrove MC, Shehab N, Richards CL. Emergency hospitalizations for adverse drug events in older americans. N Engl J Med. 2011 November 24, 2011;365(21):2002-12.
2. Institute for Healthcare Improvement. How-to guide: prevent adverse drug events with medication reconciliation. In: IHI, editor. Cambridge: Institute for Healthcare Improvement; 2011.
3. The Joint Commission. Using medication reconciliation to prevent errors. Sentinel Event Alert [Internet]. 2006 March 1, 2012; (35). Available from: http://www.jointcommission.org/assets/1/18/SEA_35.PDF.
4. Bassi J, Lau F, Bardal S. Use of information technology in medication reconciliation: A scoping review. Ann Pharmacother. 2010 May 2010;44(5):885-97. Epub April 6, 2010.
5. Lesselroth BJ, Holahan PJ, Adams K, Sullivan ZZ, Church VL, Woods S, et al. Primary care provider perceptions and use of a novel medication reconciliation technology. Inform Prim Care. 2011 2011;19(2):105-18.
6. Lesselroth BJ, Adams K, Tallett S, Wood SD, Keeling A, Cheng K, et al. Design of admission medication reconciliation technology: a human factors approach to requirements and prototyping. HERD. 2013 Spring 2013;6(3):30-48.
7. Boockvar K, Santos S, Kushniruk A, Johnson C, Nebeker J. Medication reconciliation: Barriers and facilitators from the perspectives of resident physicians and pharmacists. J Hosp Med. 2011;6(6):329-37
8. Brown SH, Lincoln M, Groen P, Kolodner R. VistA--U.S. Department of Veterans Affairs national-scale HIS. Int J Med Inform. 2003 March 2003;69(2-3):135-56.

Predictive Analytics in Healthcare (HPA): Considerations and Challenges
Moderator: Suchi Saria, PhD
Assistant Professor of Computer Science, Health Policy and Informatics, Johns Hopkins University, Baltimore, MD

The US healthcare system struggles with high cost, poor quality and uneven performance. Electronic predictive algorithms are becoming widely recognized by caregivers, policy makers and practitioners as a means for achieving the Triple Aims. Such algorithms have shown tremendous promise in other industries, and have fueled the recent "big data" revolution. Successful adoption in healthcare, however, will require careful consideration of issues surrounding HPA. In September 2013, the Moore foundation in collaboration with other industry stakeholders and Health Affairs assembled experts from informatics, health services, computer science, bioethics, and law to identify key challenges for HPA, and plan a roadmap for progress. The panelists will begin by presenting takeaways on four key questions regarding HPA: What have we learnt from past deployments? Where the most promising opportunities are in the near term and what are the current barriers for adoption? Should HPA be regulated? How should informatics education adapt to improve HPA adoption? An open discussion will follow with the goal of identifying opportunities and priorities for HPA.

Since clinical decision support, population health, and risk stratification are important applications of HPA, and of critical interest to the AMIA community, we hope this panel will generate insightful and productive discussions.

1. **Dr. B. Alex Dummett, MD,** Kaiser Permanente Northern California Region, CA
2. **Dr. Suchi Saria, PhD,** Assistant Professor of Computer Science, Health Policy and Informatics, Johns Hopkins University, Baltimore, MD
3. **Dr. Paul Tang, MD, MS,** Vice President, Chief Innovation and Technology Officer at the Palo Alto Medical Foundation (PAMF),
4. **Dr. Lucila Ohno-Machado, PhD,** Assistant Professor of Computer Science, Health Policy and Informatics, Johns Hopkins University, Baltimore, MD

Enhancing Patient Engagement in the Inpatient Care Setting

David K. Vawdrey, PhD[1], Patricia Dykes, RN, PhD, MA[2], S. Ryan Greysen, MD, MHS, MA[3], Ann O'Brien RN, MSN[4], Jaap Suermondt, PhD[5]

[1] Department of Biomedical Informatics, Columbia University, New York, NY
[2]Brigham and Women's Hospital and Harvard University, Boston, MA
[3]University of California San Francisco School of Medicine, San Francisco, CA
[4]Kaiser Permanente, Oakland, CA
[5]Hewlett-Packard Laboratories, Palo Alto, CA

Abstract

Patient engagement has been compared to a "blockbuster drug" that has potential to improve the quality and reduce the costs of healthcare. This panel will explore how health information technology—specifically tablet computers—can facilitate patient engagement in the inpatient care setting. The panelists will summarize their individual experiences leading projects at their respective institutions where tablet computers have been employed to deliver tailored information to patients and to enhance patient-provider communication. These projects represent a variety of hardware/software platforms, EHR systems, patient populations, and organizational models. The panelists will discuss the current and future state of inpatient engagement technology, debating topics such as: 1) what information should be shared with patients, and in what form should it be displayed; 2) how to best engage patients' family members and care partners while protecting patients' privacy interests, 3) how can technology support patients with low health literacy and health numeracy, and 4) how have institutions' patient engagement efforts affected frontline physicians, nurses, and other healthcare providers. Ample time will be provided for audience questions and group discussion.

General Description

"If patient engagement were a drug," wrote health information technology consultant Leonard Kish in 2012, "it would be the blockbuster drug of the century and malpractice not to use it." A key ingredient of patient engagement is access to timely, clear, and understandable information for patients and their caregivers. As anyone who has been a hospital patient or witnessed the hospitalization of a loved one knows, the lack of information in the inpatient care setting contributes to anxiety and feelings of helplessness (1). Cumbler and colleagues reported that only 28% of hospital patients were provided with the opportunity to review their inpatient medication list (2), and O'Leary and colleagues found that just 32% of the hospital patients that they surveyed could correctly name even one of their hospital physicians (3).

Health information technology can support patient engagement. Foundational work in this area was performed by Safran and colleagues (4), Mandl and colleagues (5), and others (6), but until recently, it has not been feasible to disseminate the applications they developed into broad clinical practice. With the rapid adoption of electronic health records, personal health records, and mobile devices such as tablet computers, informaticians are creating novel and exciting systems to support patient engagement.

This panel will educate healthcare providers and informatics professionals about the benefits and challenges associated with the use of tablet computers to enhance patient engagement in inpatient care settings. The members of the panel have extensive experience implementing and evaluating patient engagement technologies across a variety of care settings using commercial and locally-developed EHR and PHR systems. Panel members will synthesize their perspectives on the biomedical literature and likely future developments in this area, exploring the following diverse set of topics and engaging in thoughtful discussion based on questions from the audience.

Discussion Topics

Determining what information should be shared with patients
The federal Meaningful Use financial incentive program requires eligible hospitals to provide patients the ability to view online, download and transmit their health information within 36 hours of discharge. In their experience with inpatient information sharing, panelists have discovered that patients increasingly want greater access to their information while they are in the hospital. At the same time, there is concern that patients may misunderstand or misinterpret information, resulting in increased anxiety and more work for nurses and physicians. Many clinicians anticipate that if patients have greater access to the inpatient records, the content of progress and consult notes may change, possibly hindering clinician-to-clinician communication.

Engaging patients' family members and care partners while protecting patients' privacy interests
In a recent JAMA editorial, Bates and colleagues explained that "health care systems today do not optimally identify or engage [lay caregivers] and frequently even push them away by creating barriers to obtaining patient information that may help in the care of their family member, often in the name of privacy and security…." There is general agreement that with appropriate privacy and security safeguards, a patient's family members can play a crucial role in managing information and helping to plan for discharge and follow-up care needs and events. However, how to best establish digital credentials and govern access is an area of considerable debate.

Using information technology to support patients with low health literacy and health numeracy
One of the challenges of using information technology to enhance patient engagement is to prevent an increase in the Digital Divide, where educated, affluent, tech-savvy patients benefit while disadvantaged populations are left behind. The panelists have considerable experience working with patients from minority backgrounds, who don't speak English, and who come from economically disadvantaged communities. The panelists will discuss tablet computers as a bridge to overcome health literacy and numeracy barriers.

Understanding the impact of patient engagement technology on physicians, nurses, and other healthcare providers
Health information technology systems have frequently been perceived as causing additional work for care providers. The panel will present evidence about the effects of patient engagement interventions on staff workload, satisfaction, and patient-provider communication practices.

Assessing the influence of patient engagement technology on patient satisfaction
Hospitals across the U.S. have a growing number of reasons to focus on patient satisfaction. For example, many healthcare delivery organizations are concerned about the impact that the Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) survey will have on their Medicare and Medicaid financial reimbursement rates as well as their public image. The HCAHPS survey is administered to patients after they have been discharged from the hospital, and includes satisfaction measures relating to communication with nurses, pain management, and patient education among others. Survey results will be publicly reported and will be used in part to determine CMS reimbursement rates. Panelists will discuss the impact of patient engagement technology on patient satisfaction at their institutions, presenting a case for the financial benefits of engaging patients in the healthcare.

Panel Members

Patricia Dykes, RN, PhD, MA is Senior Nurse Scientist and Program Director for Research in the Center for Patient Safety Research and Practice and the Center for Nursing Excellence at Brigham and Women's Hospital (BWH) and Assistant Professor at Harvard Medical School. Her research interests are in quality and safety of care and adverse event prevention, especially as it relates to hospitalized inpatients. While funded by the Robert Wood Johnson Foundation, Dykes and team developed a fall prevention toolkit that significantly reduced falls in hospitals. She has expanded this research to explore the use of technology to provide the core set of information needed by patients and care team members to engage in safe patient care. Dr. Dykes is a fellow of the American Academy of Nursing and the American College of Medical Informatics. She is the author of two books and over 50 peer-reviewed publications, including the 2013 article in Journal of Gerontological Nursing, *Building and testing a patient-centric electronic bedside communication center*, which describes the use of a participatory design process to develop and test an electronic bedside communication center prototype to improve access to health information for hospitalized adults and their family caregivers. The prototype was recently enhanced to include additional tools to support patient engagement in their plan of care and is being implemented and accessed by patients and family caregivers on iPads on the medical intensive care and oncology units at BWH.

S. Ryan Greysen, MD, MHS, MA is Assistant Clinical Professor in the UCSF School of Medicine. His research focus is transitions of care for hospitalized older adults and interventions to improve post-discharge continuity of care including novel uses of patient-centered technology such as mobile devices and social media. Dr. Greysen's mixed-methods evaluation studies have leveraged advanced quantitative and qualitative techniques to inform health policy at the intersection of mobile health, hospital medicine, and geriatrics. His 2014 publication in the Journal of Hospital Medicine shows that hospitalized patients can be more engaged in their care by using tablets to access their patient portal at the bedside and that older or less tech-savvy patients benefit just as much as others if given focused coaching. His ongoing work expands this bedside coaching with portals and integrates mobility sensors to engage older adults/caregivers in goal-setting to avoid complications such as functional decline during hospitalization.

Jaap Suermondt, PhD is Vice President and Director of the Analytics Lab at HP Labs, responsible for the research investments in analytics systems, technologies, and infrastructure as well as applications and visualization, including in clinical informatics. He was previously Director of Healthcare Research at HP Labs, creating a portfolio of projects with clinical partners to develop innovations to improve patient and staff experiences, patient safety, and operational efficiency. Dr. Suermondt holds a PhD in Medical Information Sciences from Stanford University School of Medicine. He is an inventor on more than 40 granted U.S. patents, author of dozens of peer-reviewed publications, a Fellow of the American College of Medical Informatics, and a past program chair for the AMIA Annual Symposium. His most recent publication (*Pediatrics,* February 2014, doi:10.1542/peds.2013-2249) shows an electronic patient dashboard helping to reduce CLABSI's by 70% at a major teaching hospital. He created an ongoing research project called Visual Call Button, currently in trial deployment at a skilled nursing facility, that is aimed to enhance patient engagement by providing language-independent visual communication between patients and staff, improve the ability to track, fulfill, and perform analytics on patient requests and the resulting actions, and provide patients with real-time interactive updates and a feedback mechanism on their care and care plans.

Ann O'Brien, RN, MSN, CPHIMS is the National Senior Director of Clinical Informatics at 36-hospital Kaiser Permanente based in Oakland, California. Her leadership role leverages clinical data and innovative technology to transform nursing care processes and improve outcomes. She is currently a Robert Wood Johnson Executive Nurse Fellow and was recognized by *Modern Healthcare* as one of the Top 25 Clinical Informaticists of 2012. As a nurse informaticist, Ms. O'Brien also is leading Kaiser's initiative to transform care delivery through enabling technology such as the development of clinical information dashboards and providing tailored education to hospital patients using tablet computers.

David K. Vawdrey, PhD will moderate the panel. Dr. Vawdrey is Assistant Professor of Clinical Biomedical Informatics in the Department of Biomedical Informatics at Columbia University and Assistant Director for Medical Informatics Services at NewYork-Presbyterian Hospital. In 2011, he conducted one of first studies of patient engagement using tablet computers in a hospital setting, where iPads linked to the electronic health record were provided to cardiology patients, allowing them to review portions of their charts and interact with their care team members. He is the principal investigator on a project funded by the Agency for Healthcare Research and Quality (AHRQ; R01 HS21816) to investigate hospital patient information needs and how they can be met using information technology.

Participation statement: All members have agreed to participate in the panel.

References

1. Skeels M, Tan DS. Identifying opportunities for inpatient-centric technology. Proceedings of the 1st ACM International Health Informatics Symposium; Arlington, Virginia, USA: ACM; 2010.
2. Cumbler E, Wald H, Kutner J. Lack of patient knowledge regarding hospital medications. J Hosp Med. 2010 Feb;5(2):83-6.
3. O'Leary KJ, Kulkarni N, Landler MP, Jeon J, Hahn KJ, Englert KM, et al. Hospitalized patients' understanding of their plan of care. Mayo Clin Proc. 2010 Jan;85(1):47-52.
4. Safran C. The collaborative edge: patient empowerment for vulnerable populations. Int J Med Info. 2003 Mar;69(2-3):185-90.
5. Mandl KD, Kohane IS, Brandt AM. Electronic patient-physician communication: problems and promise. Ann Intern Med. 1998 Sep 15;129(6):495-500.
6. Cimino JJ, Patel VL, Kushniruk AW. What do patients do with access to their medical records? Stud Health Technol Inform. 2001;84(Pt 2):1440-4.

Patient-Generated Health Data in Practice – Learning from the Early Experiences of Innovators

Jonathan S. Wald, MD, MPH[1,7], S. Trent Rosenbloom, MD, MPH[2], Susan Woods, MD, MPH[3], Carolyn Kerrigan, MD, MHCDS[4], Alistair Eskine, MD[5], Neil Wagle, MD, MBA[6,7]

[1]RTI International, Research Triangle Park, NC; [2]Vanderbilt University Medical Center, Nashville, TN; [3]Veterans Health Administration, Portland, OR; [4]Dartmouth-Hitchcock Medical Center, Lebanon, NH; [5]Geisinger Health System, Danville, PA; [6]Partners HealthCare System, Boston, MA; [7]Harvard Medical School, Boston, MA

Abstract

Patient-generated health data (PGHD) is important for monitoring chronic conditions, treatment adherence, treatment response, symptom severity, satisfaction, and shared decision-making. PGHD arises in a variety of contexts – it may be patient- or caregiver-initiated, passive data from a device, responses to a provider or care team request, or corrections when medical chart information is accessible to the patient. Through these and other uses, PGHD can play a significant role in care quality, patient safety, practice efficiency, and the patient care experience. Innovative organizations are gaining experience in collecting, reviewing, and documenting PGHD. A technical expert panel convened in 2013 on behalf of the Office of the National Coordinator for Health IT examined key PGHD opportunities and issues, supporting PGHD as a federal Stage 3 meaningful use objective. While information submitted by patients can benefit patient care, shared decision making, quality reporting, comparative effectiveness research, quality improvement efforts, and marketing activities, digital PGHD also elicits clinician concern about liability, information veracity, impacts on provider workflow, mismatched provider-patient expectations, and limited resources for timely analysis of PGHD. This interactive panel brings together leading innovators for a discussion of digital PGHD use in a variety of settings and contexts. Topics will include: good practices in selecting apps and data for gathering and managing PGHD; types and value of PGHD; how PGHD is received, reviewed, and documented; provider and patient intentions in using PGHD; policies supporting the use of PGHD; and practical approaches to making PGHD actionable.

Intended Audience

The intended audience includes providers, EHR designers, implementers, management, policymakers, senior leadership, patients, and others concerned with increasing the routine use of patient-generated data in care activites.

Introduction to the Topic

Providers, patients, caregivers, and researchers rely routinely on information obtained directly from the patient. Medical history-taking, specialist reviews of treatments with other providers, nutritional habits, self-management for a chronic condition, medication use, decision-making preferences, and service satisfaction are just a few examples in which information from the patient or caregiver, sometimes referred to as patient-generated data (PGHD), plays an important role. PGHD has been described as data "created, recorded, gathered, or inferred by or from patients or their designees to help address a health concern" and is operationally distinct from data captured in clinical settings by providers in two important ways: (1) patients, not providers, are primarily responsible for capturing or recording these data, and (2) patients direct the sharing or distribution of these data to health care providers and other stakeholders[(1)].

Interest in formalizing the use of PGHD is growing for a number of reasons. Models of care such as the patient-centered medical home (PCMH) are increasing. Developed around the Chronic Care Model[(2)], PCMH highlights the use of between-visit information for any condition that is monitored over time, such as blood pressure and glucose data in diabetes[(3)]. Expanding efforts to focus on patient safety in ambulatory care requires routine review and reconciliation of medications[(4)], which can be facilitated by patient-reported medication information collected at or before a visit to help streamline this time-consuming but important task. The aging population and growing prevalence of chronic disease[(5)] is expanding the number of complex patients – those having multiple conditions, medications, specialist providers, and higher risks of having siloed data and care coordination gaps[(6)]. Greater focus

on cost savings by reducing unnecessary treatments and leveraging shared decision-making as part of the care process requires robust and documented patient preferences in order to support personalized decisions.

A technical expert panel convened in 2013 on behalf of the Office of the National Coordinator for Health IT examined key PGHD opportunities and issues, supporting PGHD as a federal Stage 3 meaningful use objective. In addition to the potential benefits of PGHD to patients, it also elicits clinician concern about liability, information veracity, impacts on provider workflow, potentially mismatched provider-patient expectations, and limited resources needed for timely analysis of PGHD[7].

Aim of the Discussion, Expected Discussion, and Timeliness

The aim is to use panelist experiences designing and implementing PGHD processes to prompt the audience to consider the sociotechnical aspects such as policies, preconceptions, analytic methods, and technical challenges of introducing and working with PGHD. Discussion is expected to encompass opportunities and actual or potential solutions, challenges and fears, and areas requiring further exploration relating to PGHD.

This is an important and timely discussion because leading innovators are already introducing PGHD, many organizations view PGHD as an important component of value-based (instead of volume-based) care, it is rather complex to design systems for the collection and review of many different kinds of data leveraging many different technologies in many different home and practice settings, and PGHD is proposed as a menu item for Stage 3 meaningful use.

Specific Contribution of Each Speaker

This interactive panel brings together leading innovators to discuss their growing experience using PGHD in a variety of provider settings and contexts, given its emerging importance and practical challenges. Topics will include: selecting apps for gathering and managing PGHD; data types; how PGHD is received, reviewed, and documented; data sampling, veracity, reliability, and provenance; provider and patient intentions in using PGHD; policies supporting the use of PGHD; practical approaches to making PGHD actionable; and maximizing the value and business alignment of PGHD.

Moderator: Jonathan S. Wald, MD, MPH, Director of Patient-Centered Technologies, RTI International, Research Triangle Park, NC; Dr. Wald served as an author of the 2012 White Paper on PGHD(1), provided testimony to the HIT Policy Committee on PGHD in June, 2012, and co-chaired the 2013 PGHD Technical Expert Panel. He will provide a brief introduction to the opportunities and challenges of PGHD.

Panelist: S. Trent Rosenbloom, MD, MPH, Director of Patient Engagement, Vanderbilt University Medical Center, Nashville, TN; Dr. Rosenbloom will discuss PGHD in the context of Vanderbilt's My Health Team (MHT) chronic care management program including the use of patient journaling, prompted responses from patients between office visits, and emerging alerts and decision support for patients.

Panelist: Susan Woods, MD, MPH, Director of Patient Experience, Connected Health Office, Veterans Health Administration, Washington, DC; Dr. Woods will discuss plans at the VHA to expand PGHD integration into clinical tools, patient-facing health technologies and web and mobile apps, a framework for an enterprise federated architecture, and policies and clinical process requirements for PGHD.

Panelist: Carolyn Kerrigan, MD, MHCDS, Professor of Surgery, Dartmouth Medical School, Hanover, NH; The successful integration of PGHD into care requires coordination of at least 4 essential elements: 1) leaders to sponsor and help set priorities of the initiative; 2) IT to design and build tools; 3) operations to integrate data capture and end-user review into everyday workflows; and most importantly 4) patients to co-design, test, and use. Dr. Kerrigan will share how this strategy is used to successfully implement PGHD in at least 18 clinical areas at Dartmouth-Hitchcock Medical Center.

Panelist: Alistair Eskine, MD, Chief Clinical Informatics Officer, Geisinger Health System, Danville, PA; Geisinger uses a commercially available EMR (Epic 17-years) with a robust patient portal (233,000 registered patients) and gathers PGHD from rheumatology and heart failure patients. Geisinger uses automated phone message, text-messaging reminders and online tools to engage patients, with plans underway to explore more interactive, real-time social networking strategies and additional telehealth assistive technologies to further engage patients and bring their PGHD into the EMR.

Panelist: Neil Wagle, MD, MBA, Medical Director for Quality, Safety, and Value – PROMs, Partners HealthCare System, Boston, MA; Partners has been working to implement a patient-reported outcome measures (PROMs)

platform for over two years, learning a great deal about decision makers in different clinical settings, workflow, operational issues such as enabling the medical and administrative assistant to administer a PGHD platform, data considerations, risk adjustment, issues of privacy and consent, budgeting, technology policy, and arriving at consensus around instruments both internally and with external parties. In summary, Partners has learned a great deal of theoretical and practical knowledge about implementing a PROMs platform.

[Note: Dr. Mattison, below, will attend the interactive session. If any panelists change their plans and are unable to attend, Dr. Mattison has agreed to join the panel.]

Optional Panelist: John Mattison, MD, Chief Medical Information Officer, Kaiser Permanente, Southern California; Dr. Mattison will discuss his work on a wide range of innovation pilots using social, mobile, motivational apps and avatars to both motivate healthy decisions by patients and to better engage them in their own care. He will expose his taxonomies for both a) how wearable sensors will evolve in three classes, and b) how the torrents of PGHD can be managed from both an analytic perspective as well as for safety net sniffing for escalation from machine to human interventions.

Statement of Participation

All participants have agreed to take part in the panel and have sent CVs to be included in this proposal.

References

1. Shapiro M, Johnston D, Wald J, Mon D. Patient-Generated Health Data 2012. Available from: http://www.rti.org/pubs/patientgeneratedhealthdata.pdf.
2. Wagner EH. Chronic disease management: what will it take to improve care for chronic illness? Eff Clin Pract. 1998 Aug-Sep;1(1):2-4. PubMed PMID: 10345255. Epub 1999/05/27. eng.
3. Gruman J, Jeffress, D., Edgman-Levitan, S., et al. The Opportunity for Patient-Centered Medical Homes to Support Patients' Engagement in Their Health and Health Care. Washington, D.C.: Center for Advancing Health, 2011.
4. Haynes RB, Ackloo E, Sahota N, McDonald HP, Yao X. Interventions for enhancing medication adherence. The Cochrane database of systematic reviews. 2008 (2):CD000011. PubMed PMID: 18425859.
5. Dall TM, Gallo PD, Chakrabarti R, West T, Semilla AP, Storm MV. An Aging Population And Growing Disease Burden Will Require ALarge And Specialized Health Care Workforce By 2025. Health Aff (Millwood). 2013;32(11).
6. Samal L, Hasan O, Venkatesh AK, Volk LA, Bates DW. Health Information Technology to Support Care Coordination and Care Transitions: Data Needs, Capabilities, Technical and Organizational Barriers, and Approaches to Improvement (Commissioned Paper)2012 December 13, 2013. Available from: http://www.qualityforum.org/Publications/2012/02/Health_Information_Technology_to_Support_Care_Coordination_and_Care_Transitions.aspx.
7. Deering MJ. Issue Brief: Patient-Generated Health Data and Health IT2013 1/4/2014. Available from: http://www.healthit.gov/sites/default/files/pghd_brief_final122013.pdf.

Interactive Panel:
How safe are users of Consumer Health Informatics?

Thomas Wetter, Heidelberg University (Germany) and University of Washington, Seattle (WA); **Mary Czerwinski**, Microsoft Research, Visualization and Interaction (VIBE) Research Group, Redmond (WA); **George Demiris**, University of Washington, Seattle (WA); **Robert C Hsiung**, Dr. Bob LLC, Chicago (IL) **Holly Jimison**, Northeastern University, Boston (MA)

Abstract

Consumer Health Informatics (ConsHI) has proven beneficial for many medical and mental conditions. Since it assigns the patient an active role reliable patient contributions are a necessary condition for safe operation, but cannot be taken for granted when patient physical, mental or emotional states deteriorate. The panel presents different approaches from within and outside ConsHI (Home monitoring of cognitive and physical functions, emotion tracking, behavior in online communities), that allow early detection and curbing of the risks. Routine application of such safeguards creates large amounts of data that must be protected against privacy breaches and that must be made sense of, and it causes additional cost. Patients and policy makers therefore face choices whether to mandate such additional safety when ConsHI services go routine.

Introduction

Consumer Health Informatics (ConsHI) is the discipline of using information and communication technology for services that enable the citizen to safely play an active role in his health prevention, medical care, rehabilitation, and as ambient assistive technology. ConsHI services often have a Personal Health Record as passive data backbone. In this panel we concentrate on so called active ConsHI: services for medical risk assessment, symptom tracking and interpretation, safeguarded discharge after inpatient treatment, psychotherapeutic modules for depressed or phobic patients, dose adaptation to asthma or diabetes patients based on flowmeter or glucose readings and many more. An increasing number of services fully by-pass the classical provision of medical care. Many such approaches have shown promising results that seem to suggest that ConsHI services can step in when human resources in health care run low – which regarding the demographic development they inevitably will.

Under the umbrella of trials IRB supervision and respective tight monitoring of unanticipated risks is in place all time. However, when such services leave the research arena, they may be considered the illegal practice of medicine without in-person contact with patients (Wetter et al; AMIA Annual Symposium, 2013, Panel 73). There are good ethical arguments to change legislation as long as services are safe enough and potential benefits outweigh risks. To inform the ethical discussion, it is timely to explore how safe ConsHI can be made.

Challenges to fail-safe Consumer Health Informatics

Generally, active ConsHI heavily relies on patient contributions to be faithful. Severe conditions such as depression or diabetes may lead to falsely reported signs, concealed information, or non-adherence to advice, with increased risk of critical life threatening events lurking. So patient willingness, capability and emotional stability to cooperate contribute to “safe enough” ConsHI that a provider should check ***before*** offering a contract. We subsequently concentrate on more subtle and harder to catch threats to patient cooperation that develop over time and turn a responsible patient into someone who is a risk for himself. Degrading capability to cooperate may hit a patient in various dimensions. Cognition may fade due to undiagnosed dementia or neural conditions and patients can no longer faithfully report or reliably follow up upon advice. Circumstances of life, side effects of therapies, etc., may affect moods and then behaviors such as discontinuation of a therapy. Bodily functions deteriorate very gradually and patients perform functions at more and more effort. Apparently normal behavioral patterns may already show abnormalities or

signs of exhaustion that can be used to intervene before catastrophic failure. Or a depressive person has his/her ups and downs over time but eventually a down goes so deep that suicide becomes a tangible risk. If such deteriorations happen in the classical in-person care situation the health care professional has a fair chance to notice and intervene in time. If ConsHI strives for an on par role of service provision it faces the challenge of finding technical substitutes for the good senses enhanced by clinical acumen that humans can apply. Health and humanities professionals play an important role in identifying incremental decays in client responsiveness and evaluating conclusions drawn from big data analysis such as the one presented subsequently.

Approaches to reduce risks of failure

Holly Jimison: Safety and Privacy Issues in Home Monitoring and Health Interventions for Older Adults in the Home

With the growth of the elderly population outpacing other age groups, and with an associated escalation of health care resources spent on chronic disease and conditions of aging, new models of care need to focus on scalable and coordinated interventions to the home. Needs assessments have shown that the highest priorities for older adults are to remain independent and age in place. New technology developments ranging from sensors to mobile computing and communications offer promise for effective health interventions for quality of life and independence. However, the monitoring of activities and behaviors, as well as the need to share key data bring up important privacy and data sharing challenges. In interviews and debriefings, we have found that older adults find the benefits of coaching and feedback to be worth the privacy trade-off. The concerns seem to pale in comparison to losing independence and not being able to "age in place." However, it is particularly important to understand and explicitly model data sharing preferences. The challenge in this domain is how best to assess, infer, and implement tailored data sharing protocols. Another safety concern with home monitoring interventions addressed in this panel has to do with automated alerting, often based on noisy and context based data from sensors in the home. We will present a decision theoretic analysis of over-alerting versus missing important safety events, ranging from medication errors to falls. Finally, through monitoring of cognitive performance, as we do through adaptive cognitive computer games and motor speed monitoring, we can improve the safety of ConsHI services by detecting cognitive impairment and reduced client reliability and by adapting the system interfaces and protocols to account for potential cognitive impairment.

Mary Czerwinski: Emotion Tracking for Awareness, Health and Well-being

In emotional tracking for health and well-being, we flip the typical affective computing world on its ear and give the emotional signal back to the user for interpretation. We aim to empower the users with better awareness and tracking capability about their mental state over time. In addition, we attempt to intervene with psychologically relevant activities that are infused with social media, which make the exercises more enjoyable. In this way, we get users to interact with our intervention platform (a mobile phone application with psychologically relevant suggestions to participate in social or individual popular activities and intervention suggestions) for longer periods of time, and we've been able to show that, if they do so, they develop more positive coping strategies. In addition, we've found that the awareness of their affective state alone is useful for behavioral change. Clients of a ConsHI service might be upset through intimidating medical data (labs, tumor markers) or by signs they observe about themselves and interpret as a beginning of a relapse, etc., or an adverse effect of a medication. They may unconsciously try to ignore these signals. This is true even though expedited treatment would be beneficial if not life-saving (e.g., a graft vs. host reaction). Emotion detectors could also react to affective changes caused via bad medical news. This could be general, and all-encompassing across medical conditions. In some future these signals may be able to distinguish negative affect caused by work or family or health

care. But in the near future, these signals could be used to alert ConsHI service providers of the affect changes of their clients that might require referrals to their clients' GP or equivalent before continuing the service. And finally, we do believe that at some point these clients will be ok with sending this longitudinal information to their doctors or caregivers, which can then help them make better decisions about long-term care and what strategies/medications/exercises are working best for that patient. Even if the patient isn't seen in person for long periods of time, doctors and other caregivers can be fully aware of the patient's affective state and interact with that patient "just in time" with directed interventions (possibly through technology in the home).

George Demiris: Creating value from embedded technologies to the safety of older citizens in their homes

Smart homes, namely residential settings with embedded technology that facilitates passive monitoring of residents, generate very large data sets. With the application of sophisticated data mining and pattern recognition approaches it is possible to make inferences about patterns of activities of daily living and ultimately study trajectories of activities over time in order to automatically detect abnormalities and potentially prevent adverse events without having to rely on human observers or self-report. Such technologies are increasingly used in community based housing where individuals with one or more chronic conditions are provided with a safety net for early detection of behavioral abnormalities that point to severe underlying health problems. This approach could potentially empower consumers who can more actively monitor their overall well-being which aligns well with the emerging quantified self movement. However, challenges pertain to a) ensuring algorithms are accurate and the false-positive and false-negative signals do not become burdensome or disruptive to everyday life nor jeopardize residents' safety; b) visualizing the generated information so easy to understand that consumers themselves can process and annotate the generated knowledge; c) integrating clinicians so that the sensor data enhance clinical decision making but do not burden clinicians with new demands on new data sets; and d) addressing ways to maximize the benefit of smart home data for health consumers in terms of access, awareness, and safety. Smart home sensor data can facilitate not only more effective monitoring to maximize resident safety but often support decision making around transitions of care when it may no longer be safe for an individual to live alone or where specialized care and assistance are due.

Robert C Hsiung: How safe are members of online communities (and users of social media) with depression?

Depression can dramatically lower the quality of, and even end, life. The safety of members of online communities (including users of social media) with depression could potentially be increased by a) improving detection of worsening depression and b) facilitating professional intervention. Shame and stigma can prevent people with depression from receiving support from peers and treatment from professionals. Online community members who are anonymous tend, however, to feel less exposed and more free to share how depressed or suicidal they feel. Worsening depression may be detected by automated Natural Language Processing methods at nearly the precision and recall as expert opinion. Detection may also be considered crowdsourced by the forum software to other community members, who may also have more familiarity with depression and be better able to "see through" denial than "real-life" friends and family members. Intervention may be facilitated by alerting affected community members directly, other community members, community moderators, or even the primary care physicians of the community members. To implement the last while safeguarding the privacy of the community members, all that would be required would be for the community members to obtain the agreement of their primary care physicians (false positives would be a risk) and to trust the community management enough to provide contact information for their primary care physicians. Then if their depression worsened, software or community members could initiate an automated and confidential contact process that could include the content of the posts that were considered to indicate worsening depression.

The Impact of HIT on Cost and Quality in Patient-Centered Medical Home Practices

Julia Adler-Milstein, PhD[1,2], Genna R. Cohen, BS[2] Amanda Markovitz, MPH[3], Michael Paustian, PhD[3]

[1]School of Information, University of Michigan, Ann Arbor, MI; [2]School of Public Health, University of Michigan, Ann Arbor, MI; [3]Blue Cross Blue Shield of Michigan, Detroit, MI

Abstract

While health IT is thought to play a critical role in supporting new models of care delivery, we know little about the extent to which HIT improves cost and quality outcomes. We studied a large patient-centered medical home (PCMH) program to assess which types of HIT led to improvements in composite performance outcomes: PMPM cost, chronic disease management, medication management, and preventive care. At baseline, registries were associated with lower PMPM spending (-$19.37; p<0.05). Over time, practices that newly adopted EHRs had smaller gains in chronic disease management adherence relative to non-adopters (diff-in-diff: -1.55%; p<0.05). We failed to find a relationship between other types of HIT – ePrescribing and PHRs/Portals – and our composite outcomes. The lack of consistent relationship between HIT adoption and improved performance suggest that these tools may not yet support the clinical activities and approaches to patient engagement that enable PCMHs to deliver higher-quality, lower-cost care.

Introduction

The HITECH Act seeks to promote widespread adoption of health information technology (HIT) in order to address persistent quality and efficiency challenges facing our healthcare delivery system, many of which derive from reliance on paper records.(1) HITECH also seeks to put in place the infrastructure for broader efforts to reform healthcare delivery by supporting new models of care.(2) While the role of HIT is thought to be critical to achieving these reforms, we know little about the extent to which HIT is impacting the intended quality and cost outcomes. We therefore studied a large program promoting implementation of the patient-centered medical home (PCMH), a new model of care delivery that has received substantial policy attention, in order to assess whether adoption of various types of HIT was associated with improved performance. Assessing the impact of IT on cost and quality performance in PCMH practices was of particular interest in light of recent evidence documenting the shortcomings of HIT solutions to support key PCMH practice activities.(3, 4)

Background

Patient-centered medical homes were developed in response to widely recognized shortcomings in primary care delivery. While the concept has a long history(5), in their modern form they are based on the principles of providing care that is accessible, continuous, comprehensive, family-centered, coordinated, compassionate, and based on trusting relationships.(6) These principles may be achieved by promoting integrated care teams who are supported by information technology. Medical practices and other types of delivery settings across the nation are implementing and assessing the medical home model through federal, state and regional programs. Many have relied on a set of standards for PCMH achievement developed by the National Committee for Quality Assurance (NCQA).

The use of IT in PCMHs is thought to be essential to successfully achieving the key aims of team-based, coordinated care and meaningfully engaging patients.(7) For example, patient registries enable population-level views of clinical data that can help health care providers identify systematic care gaps and opportunities for targeted outreach. Electronic health records can promote teamwork through clinical messaging, task tracking, and creation of shared lists (i.e., a problem list). These functions should also help ensure that care gaps are filled when the patient is in the

office, and that follow-up on critical test results is performed. ePrescribing can further complement these activities by enabling more careful tracking of medications. Another key promise of the PCMH model is the ability to involve patients in their care. Patient portals and personal health records (PHRs) hold the potential to engage patients through making their clinical data accessible to them. These HIT applications can also serve as teaching tools that allow patients to manage their own conditions as well as facilitate effective patient communication with their care team.

There has been no large-scale empirical assessment of the extent to which various types of HIT are enabling PCMH practices to deliver higher-quality, lower-cost care. We therefore leveraged a PCMH program across the state of Michigan to examine whether HIT adoption was associated with better performance at baseline, and then to assess whether those that newly adopt HIT over time realized greater performance gains relative to those that did not adopt.

Methods

Setting

We examine the impact of HIT adoption on cost and quality outcomes within ambulatory practices that participated in the Blue Cross Blue Shield of Michigan's (BCBSM) Physician Group Incentive Program (PGIP). PGIP is a pay-for-performance program that was launched in 2005 to align financial incentives for structural and quality improvement initiatives. Individual physician practices participate in PGIP through a physician organization (PO) intermediary, such as an Independent Practice Association or a Physician-Hospital Association.(8) The 40 POs participating in PGIP represent nearly 15,500 physicians providing care to nearly two million BCBSM members. (9)

One of the flagship PGIP initiatives focuses on promoting transformation into a PCMH. Physician practices are awarded financial incentives for implementing capabilities aligned with the PCMH model - for example, providing 24-hour patient access to a clinical decision-maker by phone and after-hours urgent care access as well as offer education materials to patients during visits. These capabilities are similar to those developed by the NCQA for their PCMH program.(10)

Population Studied

We studied physician practices participating in PGIP over the course of three years between July 2009 and June 2012. This period lags one year behind the launch of the PCMH program in January 2008, providing practices time to establish sufficient capabilities before assessing their impact. We focused our analysis on adult primary care practices that participated in PGIP continuously over our analytic period and had demonstrated that they were pursuing PCMH capabilities (defined as a PCMH score of greater than 0.2 in June 2009 – see below for explanation and interpretation of score). After applying these inclusion criteria, the final analytic sample included 573 practices.

Data sources

Our study relied on data from five sources. The first -- the BCBSM Self-Reported Database (SRD) -- included IT adoption status, physician demographic information, PO membership, practice affiliation (i.e. which physicians make up each practice), a primary care physician (PCP) indicator and practice PCMH capabilities. It is collected semi-annually from PO representatives and the self-reported capabilities are verified by BCBSM at random. The second and third sources of data -- BCBSM member enrollment and claims data – were used to obtain demographic data on members who received care at PCMH practices and then construct cost and quality measures. The fourth and fifth sources of data -- the 2010 United States Census and 2011 American Community Survey (ACS) – were used to create variables to adjust for potential confounding from market area, socio-economic or demographic factors not available through the SRD or BCBSM administrative data.

Measures

HIT Adoption. We used survey questions from the SRD to measure the adoption (yes/no) of four different types of HIT in each practice: (1) electronic health records (EHRs); (2) patient registries; (3) ePrescribing (eRX); (4) patient portals or personal health records (PHRs) at two time points: as of June 2010 and as of June 2011.

Cost and Quality Outcome Measures. We constructed a practice level measure of cost and three composite measures of quality of care.(11, 12) Total combined medical and surgical allowed cost per member per month (PMPM) was calculated to assess cost of care. To assess quality, we calculated composite measures of chronic disease management (e.g., HbA1c testing for diabetic patients), medication management (e.g., annual monitoring for patients on ACE/ARBs), and preventive care (e.g., breast cancer screening). We relied on composite measures because of concerns about sufficient numbers of patients for any individual measure (13) and heterogeneity in performance across individual measures (14). Outcome measures were calculated for July 2009-June 2010 (i.e., the year prior to our first measure of HIT adoption) and July 2011-June 2012 (i.e., the year following our second measure of HIT adoption).

Practice Controls. We included seven practice characteristics calculated from the SRD and administrative claims data as covariates: practice size, primary care focus, BCBSM patient volume, physician turnover, number of years in the PGIP, and practice movement between physician organizations. We calculated practice size through a list of providers' and their practice affiliations using the SRD. We categorized size as follows: solo (1 physician), small (2-3 physicians) medium (4-5 physicians), and large (6+ physicians). Data were not available to capture other measures of practice size, such as number of clinical or administrative support staff. We used the PCP indicator in the SRD to classify practices as 'primary care-only' if only PCPs were present and 'multispecialty' if both primary care and non-primary care specialty physicians were present. Total BCBSM paid services delivered per PCP were calculated annually for each practice as a proxy for BCBSM volume within the practice. We measured physician turnover as the proportion of practice physicians who left the practice in each time period. Average number of years that the practice's PCPs participated in PGIP was used as the measure of the practice's longevity in PGIP. A binary (0/1) indicator identified practices that changed physician organizations between practice-years.

Patient Cohort Controls. We used the primary care attributed member cohort for each practice and their member enrollment information to estimate the two practice level patient characteristics: proportion of members who were female and mean prospective risk score (OptumInsight® Symmetry version 8) for adult patients in the practice. The prospective risk score employs a large national database of aggregated claims and membership information to derive a numerical, diagnosis-based episode assessment used to predict future medical costs.

Market Controls. We examined six zip code-level market characteristics and one PO characteristic to address additional sources of variation that might influence cost and quality outcomes. We calculated the percent of residents who were non-White or Hispanic and the percent of residents who lived in a rural area using zip code data from the 2010 US Census. We used 2011 ACS data to capture median household income and percent unemployment. We further used the BCBSM Provider Enrollment and Credentialing System, which captures 94% of active physicians in Michigan, and 2010 US Census population estimates to measure total PCPs per 1,000 population estimated at the zip code level. Finally, we calculated BCBSM market share at the zip code level based on member subscriber addresses from BCBSM member enrollment information and the total estimated zip code population from 2010 US Census. All zip code measures were weighted for each practice to account for the proportion of their care provided to members residing in each zip code. We measured PO size based on the total number of affiliated practices with at least one primary care physician.

PCMH Score and Cutoff. In order to ensure that practices in our analytic sample represented those pursuing PCMH status, we calculated a PCMH score at the beginning of our analytic timeframe using information on individual PCMH capability achievement reported in the SRD (115 capabilities within 13 domains). We used the methodology of prior studies that relied on the same data(15) by calculating a PCMH score as follows: we assigned capabilities reported as 'fully in place' a value of 1 and assigned capabilities reported as 'not in place' a value of 0. When capabilities had multiple gradients, the capability score was calculated as a proportion of the maximum gradient. For example, the Extended Access domain asked respondents to identify the percentage of appointments reserved for same day scheduling from the following options: 30 or 50 percent. A response of 30 percent implementation on same-day scheduling was assigned a value of 0.6 (0.3/0.5). We calculated domain-specific scores by summing all capability scores within the domain and dividing by the maximum number of distinct capabilities within that domain. This method gives equal weight to each PCMH domain in order to avoid giving greater weight to domains with more capabilities. Finally, we calculated the overall PCMH score as the mean of the 13 domain-specific scores and only included practices that had a score of at least 0.2 (i.e., 20% of PCMH capabilities in place) in order to ensure that practices have demonstrated sufficient ability to achieve PCMH principles.(15)

Analytic Approach

Our research design relies on change over time in HIT adoption among sample practices, comparing practices that newly adopt each type of HIT in the sample between June 2010 and June 2011 to those that always or never had adopted each type of HIT over the same period. This difference-in-differences approach compares change in outcome between the pre-adoption year (July 2009-June 2010) and the post-adoption year (July 2011-June 2012) for new adopters compared to always- or never-adopters. For each of the four outcomes, we used multivariable cross-classified linear mixed models, incorporating a random effect for practice and a cumulative random effect for the physician organization. The cross-classified model accounts for the longitudinal design, clustering of practices within physician organizations, and movement of practices between physician organizations.(16)

Table 1. Practice Characteristics

Practice Characteristics	n=573	n=573
Practice HIT Adoption (count)	As of June 2010	As of June 2011
EHR	302 (53%)	355 (62%)
PHR or Portal	206 (36%)	238 (42%)
eRX	492 (86%)	527 (92%)
Registry	451 (79%)	466 (81%)
Practice Outcomes (median)	July 2009-June 2010	July 2011-June 2012
Adult PMPM	$364.16	$394.68
Chronic Disease Management	79.0%	77.7%
Medication Management	80.0%	80.0%
Preventive Care	78.2%	76.1%
Continuous practice and patient characteristics (median)	July 2009-June 2010	July 2011-June 2012
Mean prospective risk score for adults [4]	1.57	1.71
Percent of attributed members who are female [4]	52.1%	52.0%
Professional services per PCP in practice [4]	1,648	1,502
Average number of years in PGIP for PCPs in practice [1]	2.5	4.5
Turnover of physicians in practice during study year[2]	0.00	0.00
Categorical practice and patient characteristics (count)		
Practice Size [1]		
Solo physician practice	249	243
2-3 physicians	145	151
4-5 physicians	79	81
6 or more physicians	100	98
Practice Specialty [1]		
Primary Care- Only	541	548
Multispecialty	32	25
Whether practice changed POs during study year [2]		
No	566	506
Yes	7	67
PO and market characteristics (median) [3]		
Total practices in PO with a PCP	99	103
Percent BCBSM market share	32.9%	30.8%

Percent non-white residents	18.4%	18.0%
Percent rural	19.3%	19.4%
Number of PCPs per 1,000 residents	0.77	0.79

(1) As of June 2009; June 2011

(2) Change during time periods between June 2009-2010; June 2011-2012

(3) 2010 Census or 2011 ACS

(4) Attribution using claims from July 2008-June 2010; July 2010-June 2012, prioritizing most recent year

Results

In our analytic sample, the majority of practices were solo or small practices with fewer than 4 physicians. (Table 1) The vast majority of practices were designated as "primary care only" practices. There was minimal physician turnover within practices during each study year and few practices changed POs during each study year.

HIT adoption rates were fairly high at baseline: 53% EHR adoption, 36% PHR/Portal adoption, 86% ePrescribing adoption, and 79% registry adoption. (Table 1) There were, however, increases in adoption over time for all types of HIT. Specifically, 53 practices newly adopted an EHR (9% of all practices); 32 practices newly adopted a PHR or patient portal (6% of practices); 35 practices newly adopted ePrescribing (6% of practices); and 15 practices newly adopted a registry (3% of total practices).

Over time, adult PMPM spending increased from an average of $364.16 to $394.68. (Table 1). Performance on all three composite quality measures was essentially flat over time. Chronic disease management decreased slightly from 79.0% to 77.7%. Medication management remained flat at 80.0%. Preventive care decreased slightly from 78.2% to 76.1%.

In multivariable models for our first outcome of interest, adult PMPM cost, we found that, in our baseline year, practices that had adopted a registry had significantly lower cost: coefficient = -$19.37 (95% CI: -$36.17 to -$2.57, Column 1, Table 2). There were no other significant relationships between IT adoption at baseline and PMPM cost. Over time, there was a slight, but non-significant, increase in PMPM cost. Similarly, for all types of HIT, there were no significant difference-in-differences between new adopters and practices who never- or always-adopted.

For our second outcome of interest, chronic disease management, we found no significant relationships between HIT adoption at baseline and performance. (Column 2, Table 2) Over time, there was a slight, but non-significant, decrease in performance for this composite measure. For PHRs/portals, eRX, and registries, there were no significant difference-in-differences between new adopters and practices who never- or always-adopted. For EHRs, practices that newly adopted performed worse over time compared to those that had always- or never-adopted EHRs (difference-in-differences of -1.55%; 95% CI: -2.99% to -0.12%; Column 2, Table 2).

Table 2. Relationship between HIT Adoption and Cost/Quality Outcomes, July 2009-June 2010 compared with July 2011-June 2012

Independent Variable	Adult PMPM	Chronic Disease Management
Practice HIT Implementation		
EHR	$11.47 (-$2.44, $25.38)	0.77% (-0.47%, 2.01%)
PHR or Portal	$7.92 (-$9.34, $25.17)	-0.39% (-1.92%, 1.15%)
eRX	-$8.85 (-$29.17, $11.47)	0.40% (-1.42%, 2.23%)
Registry	-$19.37 (-$36.17, -$2.57)*	0.92% (-0.60%, 2.43%)
Interaction with Study Period		
Study Period	$6.12 (-$28.34, $40.57)	-1.67% (-4.61%, 1.26%)
Study Period*EHR	$1.93 (-$16.62, $20.49)	-1.55% (-2.99%, -0.12%)*
Study Period*PHR/Portal	-$15.33 (-$36.75, $6.10)	0.40% (-1.24%, 2.04%)
Study Period*eRX	-$7.16 (-$38.01, $23.68)	1.61% (-0.99%, 4.21%)
Study Period*Registry	$12.11 (-$10.87, $35.10)	-0.34% (-2.16%, 1.47%)

Practice and Patient Characteristics		
Mean prospective risk score for adults	$225.07 ($207.73, $242.40)*	2.03% (0.27%, 3.79%)*
Percent female	-$3.86 (-$8.85, $1.13)	-0.20% (-0.73%, 0.33%)
Professional services per PCP in practice	-$0.60 (-$3.82, $2.62)	-0.04% (-0.37%, 0.30%)
PCPs' average number of years in PGIP	-$3.16 (-$10.79, $4.47)	-0.03% (-0.77%, 0.71%)
Turnover of physicians in practice	-$0.03 (-$3.40, $3.34)	-0.12% (-0.44%, 0.19%)
Practice Size		
Solo physician practice	Reference	Reference
2-3 physicians	$8.00 (-$4.02, $20.03)	-0.21% (-1.40%, 0.99%)
4-5 physicians	$2.52 (-$12.79, $17.83)	-0.04% (-1.54%, 1.46%)
6 or more physicians	$10.52 (-$6.01, $27.05)	0.06% (-1.58%, 1.70%)
Practice Specialty (ref: primary care)	$19.60 (-$2.53, $41.73)	-0.69% (-2.64%, 1.25%)
Whether practice changed POs (ref: no)	-$25.62 (-$54.83, $3.58)	0.64% (-1.78%, 3.06%)
PO and Market Characteristics		
Total practices in PO with a PCP	$3.79 (-$5.01, $12.60)	-0.72% (-1.37%, -0.08%)*
Percent BCBSM market share	-$3.86 (-$15.51, $7.79)	0.66% (-0.46%, 1.79%)
Percent non-white residents	$2.76 (-$2.05, $7.56)	-0.22% (-0.70%, 0.26%)
Percent rural	$2.64 (-$0.44, $5.73)	0.11% (-0.21%, 0.43%)
Number of PCPs per 1,000 residents	$3.46 (-$17.99, $24.91)	0.46% (-1.80%, 2.73%)

For our third and fourth outcomes of interest, medication management and preventive care, we found no significant relationships with IT adoption at baseline or new adoption over time (Columns 1 and 2, Table 3).

Table 3. Relationship between HIT Adoption and Cost/Quality Outcomes, July 2009-June 2010 compared with July 2011-June 2012

Independent Variable	Medication Management	Preventive Care
Practice HIT Implementation		
EHR	0.01% (-1.62%, 1.64%)	-0.43% (-1.53%, 0.66%)
PHR or Portal	0.02% (-1.97%, 2.00%)	0.00% (-1.37%, 1.36%)
eRX	0.74% (-1.66%, 3.15%)	0.15% (-1.33%, 1.63%)
Registry	0.88% (-1.12%, 2.87%)	0.58% (-0.71%, 1.88%)
Interaction with Study Period		
Study Period	1.55% (-2.30%, 5.41%)	-2.98% (-5.08%, -0.87%)*
Study Period*EHR	-0.34% (-2.19%, 1.51%)	-0.08% (-1.09%, 0.94%)
Study Period*PHR/Portal	-0.75% (-2.89%, 1.40%)	0.13% (-0.90%, 1.15%)
Study Period*eRX	-1.34% (-4.68%, 2.00%)	-0.06% (-1.91%, 1.79%)
Study Period*Registry	-0.86% (-3.19%, 1.47%)	0.13% (-1.17%, 1.43%)
Practice and Patient Characteristics		
Mean prospective risk score for adults	7.25% (4.83%, 9.67%)*	-1.12% (-2.81%, 0.57%)
Percent female	0.36% (-0.36%, 1.08%)	2.08% (1.55%, 2.60%)*
Professional services per PCP in practice	-0.13% (-0.58%, 0.31%)	-0.03% (-0.36%, 0.29%)
PCPs' average number of years in PGIP	-0.15% (-1.11%, 0.80%)	0.73% (0.10%, 1.36%)*
Turnover of physicians in practice	-0.14% (-0.55%, 0.26%)	0.03% (-0.22%, 0.29%)

Practice Size		
Solo physician practice	Reference	Reference
2-3 physicians	-0.04% (-1.64%, 1.56%)	0.21% (-0.88%, 1.31%)
4-5 physicians	-0.81% (-2.81%, 1.19%)	0.49% (-0.92%, 1.91%)
6 or more physicians	-0.59% (-2.79%, 1.60%)	1.20% (-0.41%, 2.80%)
Practice Specialty (ref: primary care)	-1.13% (-3.63%, 1.37%)	-0.89% (-2.48%, 0.71%)
Whether practice changed POs (ref: no)	0.90% (-2.46%, 4.25%)	0.56% (-0.77%, 1.89%)
PO and Market Characteristics		
Total practices in PO with a PCP	-0.34% (-1.03%, 0.36%)	-0.59% (-1.34%, 0.15%)
Percent BCBSM market share	0.85% (-0.65%, 2.35%)	2.14% (1.05%, 3.23%)*
Percent non-white residents	-0.20% (-0.84%, 0.44%)	-0.39% (-0.89%, 0.12%)
Percent rural	0.15% (-0.29%, 0.60%)	-0.02% (-0.37%, 0.32%)
Number of PCPs per 1,000 residents	-0.45% (-3.53%, 2.63%)	2.97% (0.68%, 5.26%)*

Discussion

This study is among the first to assess the impact of various types of HIT on key PCMH performance measures. Our results suggest that, at least soon after adoption, there is limited direct impact of HIT. Registries were associated with lower PMPM cost, but when we examined new adoption using a difference-in-differences approach, the only significant relationship was for EHRs on chronic disease management performance and in the opposite direction than predicted (i.e., worse relative performance among new EHR adopters). These results suggest that current HIT tools may not immediately support the types of clinical activities and approaches to patient engagement that enable PCMHs to deliver higher-quality, lower-cost care.

While not all types of HIT were associated with better outcomes at baseline, the fact that registries – a core type of HIT that is emphasized in the PCMH model and often adopted early – were associated with lower PMPM cost coupled with the fact that new EHR adoption led to worse relative performance for chronic disease management may suggest that the impact of HIT varies with time. That is, newly adopted HIT may hurt performance in the near term by disrupting workflows, introducing complexity, and reducing the time that providers can devote to clinical practice transformation. However, once practices become comfortable with the technology and are able to invest in optimizing its use, the benefits may appear. Some of this optimization may also come in the form of more advanced HIT functionalities. Prior evidence suggests that use of clinical decision support is consistently linked with gains in adherence to evidence-based medicine. It is likely that practices that had newly adopted EHRs had not yet advanced to consistent use of alerts and reminders for chronic care, medication management, and preventive care. Over the next few years, data from a longer time horizon will become available and enable these hypotheses to be tested.

Our results may also be explained by the fact that the BCBSM PCMH program does not emphasize HIT to the same degree as other PCMH programs. A 2011 Urban Institute report compared the BCBSM PCMH program capabilities with those of NCQA and found that the BCBSM program focuses less on HIT. As a result, PCMH practices in our sample may not rely as heavily on HIT to pursue the targeted domains of performance improvement. While it will be important to assess the impact of HIT in the context of other PCMH programs, our results suggest that BCBSM may want to consider a greater focus on these tools in the PCMH capabilities they promote.

Our analytic approach includes some limitations that should be considered when interpreting the results. First, we rely on observational data and were not able to adjust for all potential sources of confounding. Of particular concern are time-varying factors, such as workflow changes, that may have been implemented in parallel with HIT adoption. Second, SRD data was provided by PO practice consultants who may not have had in depth knowledge of the status of HIT adoption in each practice, and the data only captured whether each type of HIT was adopted, not the extent to which it was used. It is therefore possible that practices had functionalities available that were not widely used. In addition, for some types of HIT, only a small number of practices newly adopted in the one-year time window, limiting our power to detect an impact on our target outcomes. Further, because adoption rates for some types of HIT were already high (i.e., eRx and registries), our findings likely reflect the effect in late adopters that may have been less interested in adopting and using technology. Third, the composite quality measures had fairly high rates of adherence in our sample (~80%), limiting the opportunity to observe improvement. Finally, we were only able to

look at the impact of HIT adoption in the year immediately following adoption. It may take longer for PCMH practices to realize the benefits of these new tools.

Conclusion

We examined the impact of HIT on cost and quality outcomes in 573 practices that had significantly progressed in patient-centered medical home implementation in a large program. We found that registries were associated with lower PMPM cost and that EHRs appeared to negatively impact quality performance for chronic disease management in the short-run. This suggests that HIT tools may require a longer time horizon to be used effectively or they simply may not yet meet the demands of new models of care delivery that are being promoted under health reform.

References

1. Blumenthal D. Launching HITECH. N Engl J Med. 2010;362(5):382-5.

2. Buntin MB, Jain SH, Blumenthal D. Health Information Technology: Laying The Infrastructure For National Health Reform. Health Aff. 2010;29(6):1214-9.

3. O'Malley AS, Grossman JM, Cohen GR, Kemper NM, Pham HH. Are Electronic Medical Records Helpful for Care Coordination? Experiences of Physician Practices. JGIM: Journal of General Internal Medicine. 2010;25(3):177-85.

4. Fernandopulle R, Patel N. How The Electronic Health Record Did Not Measure Up To The Demands Of Our Medical Home Practice. Health Aff. 2010;29(4):622-8.

5. Kilo CM, Wasson JH. Practice Redesign And The Patient-Centered Medical Home: History, Promises, And Challenges. Health Aff. 2010;29(5):773-8.

6. Dickens MD, Green JL, Kohrt AE, Pearson HA. The Medical Home. Pediatrics. 1992;90(5).

7. Bitton A, Flier LA, Jha AK. Health information technology in the era of care delivery reform: To what end? JAMA: The Journal of the American Medical Association. 2012;307(24):2593-4.

8. Wise CG, Alexander JA, Green LA, Cohen GR. Physician Organization-Practice Team Integration for the Advancement of Patient-Centered Care. Journal of Ambulatory Care Management. 2012;35(4):312-23.

9. Blue Cross Blue Shield of Michigan. Physician Group Incentive Program: About Detroit, MI2013 [cited 2013 March 12]. Available from: http://www.bcbsm.com/providers/value-partnerships/physician-group-incentive-program.html.

10. NCQA. Patient Centered Medical Home Program 2011 [March 13, 2013]. Available from: http://www.ncqa.org/Programs/Recognition/PatientCenteredMedicalHomePCMH.aspx.

11. Jaen CR, Crabtree BF, Palmer RF, Ferrer RL, Nutting PA, Miller WL, et al. Methods for Evaluating Practice Change Toward a Patient-Centered Medical Home. Ann Fam Med. 2010;8(Suppl_1):S9-20.

12. Higgins A, Stewart K, Dawson K, Bocchino C. Early lessons from accountable care models in the private sector: partnerships between health plans and providers. Health Aff. 2011;30(9):1718-27.

13. Scholle SH, Saunders RC, Tirodkar MA, Torda P, Pawlson LG. Patient-Centered Medical Homes in the United States. The Journal of Ambulatory Care Management. 2011;34(1):20-32 10.1097/JAC.0b013e3181ff7080.

14. Parkerton PH, Smith DG, Belin TR, Feldbau GA. Physician Performance Assessment: Nonequivalence of Primary Care Measures. Medical Care. 2003;41(9):1034-47 10.97/01.MLR.0000083745.83803.D6.

15. Paustian ML, Alexander JA, El Reda DK, Wise CG, Green LA, Fetters MD. Partial and Incremental PCMH Practice Transformation: Implications for Quality and Costs. Health Services Research. 2013:n/a-n/a.

16. Raudenbush SW, Bryk AS. Models for Cross-Classified Random Effects. Hierarchical Linear Models: Applications and Data Analysis Methods. 2nd ed. Thousand Oaks, CA: Sage Publications, Inc.; 2001. p. 373-98.

Capture of Osteoporosis and Fracture Information in an Electronic Medical Record Database from Primary Care

Sonya Allin, PhD[1], Sarah Munce, PhD[1], Susan Jaglal, PhD[1,2], Debra Butt, MD, MPH[1,2], Jacqueline Young, MS[2], Karen Tu, MD[1,2]
[1]University of Toronto, Toronto, ON, Canada; [2]Institute for Clinical Evaluative Sciences, Toronto, ON, Canada

Abstract

In a large database of EMR records, we explore: 1) completeness in capture of bone mineral density (BMD) T-scores required for diagnosis of osteoporosis; 2) concordance of BMD exam information with other osteoporosis information; and 3) evidence of osteoporosis screening among fracture patients. To explore completeness of exam capture, BMD exams in the EMR were related to a provincial billing database. To explore concordance of information and screening rates, 7500 EMR records were reviewed for osteoporosis and fracture details. Results show that 98% of exams billed to the province for EMR patients were found in the EMR. However, documented osteoporosis was substantiated with BMD results only 55.8% of the time. Of 151 charts for fragility fracture patients, 1 in 4 contained no evidence of osteoporosis investigation. In summary, while EMR information about osteoporosis is of variable quality, EMR records shed light on osteoporosis management indicators and completely capture BMD results.

Introduction

In 2005, the Ontario Ministry of Health and Long Term Care provided funding for the Ontario Osteoporosis Strategy with the goal of promoting appropriate osteoporosis care and preventing fractures in the province[1]. As a part of the Strategy, the Ontario Bone Mineral Density (BMD) Working group has been exploring utilization of BMD testing in the province, specifically. BMD tests provide information required to definitively diagnose osteoporosis[2]; this occurs when a patient's BMD T-score is found to be 2.5 standard deviations or more below the young adult mean. Clinical guidelines currently recommend that all patients over the age of 65 undergo BMD testing in order to screen for osteoporosis, as well as all individuals with a history of fragility fracture[2].

In Canada, some information about BMD testing is available in administrative databases, like the Ontario Health Insurance Plan (OHIP) records. However, while OHIP contains billing information for the province's BMD tests, it does not contain the results required to definitively diagnose osteoporosis (i.e. the BMD T-scores). Incidence of osteoporosis in the Ontario population, then, must instead be inferred from other administrative sources, like prescription information or hospital billings; these sources are known to provide limited accuracy[3]. Administrative data, moreover, does not provide adequate insight regarding about appropriate osteoporosis screening. Specifically, while administrative data indicate rates of post-fracture BMD testing to be low[4], these data cannot identify pathways of care among patients. This means that administrative data alone cannot determine where efforts to improve screening rates should best be targeted[4].

In part due to the limitations of administrative data, researchers have been increasingly turning to Electronic Medical Records (EMRs) in the hopes that these may provide better, more timely and comprehensive information about both osteoporosis incidence and management. EMR data have several advantages over administrative data in that the data include test results and are specific to management decisions that take place within a particular care pathway. The quality of data in EMRs, however, has tended to be variable in part because clinicians use EMRs in a wide variety of ways. For example, many physicians continue to use *both* paper records and EMR records[5]. A 2010 review of EMR usage highlighted the impact of such variability; completeness of blood pressure information captured in EMRs, for example, varied from 0.1% to 51% across studies[6]. EMR data also often has been found to contain misspelled words, missing diagnostic codes, or data that has been entered in incorrect or inconsistent database fields[7,8]. A lack of mature electronic interfaces to EMR has also hindered the completeness of data in Ontario's EMRs; according to the National Physician Survey, in 2010 approximately 40% of family physicians in Ontario

had an electronic interface to diagnostic imaging services[5]. Lack of diagnostic imaging results impacts the completeness of BMD results in EMRs, specifically; a 2005 of EMR records in the U.K. revealed limited capture of these results[9].

Study Objectives

The primary purpose of this study is to provide an updated assessment of data quality related to Bone Mineral Density (BMD) test results in EMRs and to explore the relationship of BMD data to other osteoporosis documentation by primary care physicians. To facilitate this assessment, we are using a large database of Ontario EMR records called the Electronic Medical Record Administrative data Linked Database (EMRALD) [10-12]. At the time of this study, EMRALD contained data from 296 family physicians at 31 practices who use Practice Solutions® EMR and is located at the Institute for Clinical and Evaluative Sciences in Toronto, Ontario. Records in EMRALD include comprehensive electronic patient records, consult notes, as well as diagnostic imaging and lab test results for more than 300,000 individual patients. The distribution of patients represented in EMRALD reflects the age and sex distributions of the province and the average duration of EMR usage by participating clinics was 4 years. Analysis based on EMRALD data therefore can be expected to represent relatively recent usage trends.

The specific research objectives of the present study are:

First, to assess the completeness of BMD exam capture in EMRs and the potential for exam results in the EMR to yield gold-standard diagnoses of osteoporosis based on collected BMD T-scores. To assess the completeness of exam capture, we relate BMD results in EMRALD for all patients over 40 to BMD exam billings in a provincial insurance database (OHIP). We also explore the number of tests that are stored in the EMR as text, and where machine-assisted identification of BMD T-scores can be expected to be relatively straightforward.

Second, to assess the relationship between the information about osteoporosis found in BMD results and osteoporosis information found elsewhere in patient charts. For this level of analysis, we manually reviewed 7500 randomly selected patient records in EMRALD to determine where and how osteoporosis information was coded across the sample.

Third, to explore the information that EMRALD may hold about screening for osteoporosis within primary practice and after fragility fracture. Analysis here focuses on the charts of 151 individuals who were identified in the subset of 7500 as fragility fracture patients. We explore the incidence of BMD screening in this group and characterize other information indicative of osteoporosis in the population.

Methods

Data Collection

To construct EMRALD, clinically relevant data fields from the EMRs of participating family physicians are extracted through a custom software 'plug-in' and securely transferred to ICES. Data are collected from each clinic every 6 months. At ICES, health card numbers for provincial insurance are partitioned from data and the remaining data undergoes de-identification as per standard ICES policies and procedures for maintaining privacy and confidentiality[13].

The research protocol was approved by the Research Ethics Board of the Sunnybrook Health Sciences Centre.

Completeness of BMD Exam Capture

To assess the completeness of EMRALD's capture of BMD exams, BMD results in EMRALD for all patients over 40 were related to billing records for BMD exams in OHIP.

To begin, records for all patients over age 40 as of December 31, 2011 were selected from EMRALD. The diagnostic imaging results for these patients were searched and exams that were both labeled 'BMD' and dated between December of 2011 and January of 2006 were recorded.

For the same group of patients, OHIP billing records were searched for BMD exam billings dated before the end of 2011 and after either January 1, 2006 or the date the patient first appeared in EMRALD (whichever was more recent). Of note is the fact that new fee codes for baseline BMD exams were introduced by OHIP in 2008. This means that exams in EMRALD dated after 2008 were matched with a slightly different set of OHIP billings than exams dated prior to 2008 (refer to Table 3 for OHIP billing codes).

To relate OHIP records to EMR records of BMD examination, bi-directional matches were performed. First, an attempt was made to pair each OHIP billing with a BMD exam in EMRALD that shared the *exact same date*. Because exams may appear in EMRALD several days post-billing to OHIP, a second attempt was also made to match OHIP billings with corresponding exams in EMRALD that were dated *up to 30 days after billings took place.* In the opposite direction, BMD exams found in EMRALD were paired with OHIP billings sharing the *same date* and also with billings that took place *up to 30 days prior* to the exam date recorded in EMRALD.

We report the number of tests located in both EMRALD and OHIP for the patients and the percentage of tests that could be matched between corresponding data sources. We also report the number of BMD tests that were stored in EMRALD in text format rather than as images, in an effort to gauge the potential to access numeric T-scores using computerized searches. Entries for BMD exams in EMRALD that were more than 200 characters in length or which contained instances of the word "T-score" were flagged as accessible.

Relationship Between BMD Results and Other Osteoporosis Information

To assess the relationship between BMD results and osteoporosis information found elsewhere in patient records, 7500 charts for individuals over age 20 were randomly selected from the database and manually reviewed by a trained nurse. Details regarding osteoporosis were abstracted from the machine-readable, structured portions of electronic records; these include patient profiles, notes from individual patient encounters, text-based (but not image-based) diagnostic imaging results, consult notes and lab tests. During abstraction, both content and location of pertinent information in the electronic charts were noted. Information that was abstracted included:

1. Basic Demographic Information, including the age and sex of patients;
2. Evidence of osteoporosis as documented in:
    a. BMD exam results;
    b. Patient summaries (e.g., in problem lists);
    c. Notes attached to individual patient exams (such as annual exams);
    d. Consult notes (e.g., hospital discharge notes, ER consult notes, notes from specialists);
    e. Other exam results (e.g., X-ray investigations).

Documentation that was considered indicative of osteoporosis included presence of BMD T-scores of the femoral neck or lumbar spine below -2.5, vertebral compression fracture as documented by x-ray investigation, documentation of osteoporosis in the patient profile, or a diagnosis attached to a patient exam, given in a consult note or on an imaging result. We report the number of patients identified as osteoporotic in the sample, and the distribution of osteoporosis documentation across the sample.

Documentation of Osteoporosis among Post-Fracture Patients

Current guidelines recommend that all patients who have undergone a recent fragility fracture should be screened for osteoporosis; the gold standard for diagnosis of osteoporosis is based on the result of a BMD examination[2].

To explore consideration of osteoporosis among fragility fracture patients, the subset of 7500 records described above was additionally reviewed for fragility fracture evidence. Evidence that was considered indicative of a fracture history included radiological evidence of fracture (X-rays, CT scans, etcetera),

documentation of a fracture in patient profiles, hospital discharge notes, or mentions of fractures in consult notes from specialists. If fractures were specifically indicated as “fragility fractures” in charts, this was noted. Fractures of the head, foot, toe, hand and finger were excluded from analysis.

We report the number of patients identified as having a history of fracture in the sample, the number with specific mention of “fragility fractures”, and the percentage of fragility fractures patients for whom there is evidence to suggest consideration of, or screening for, osteoporosis.

Results

Completeness of BMD Exam Capture

Records for 79,740 patients over age 40 were located in EMRALD.

OHIP billing records indicated that 14,536 of these individuals had one or more BMD tests after their appearance in EMRALD, resulting in a total of 20,174 OHIP billings for BMD tests, or 1.39 billings per patient (on average) between 2006 and 2011. A summary of the BMD tests as recorded by the OHIP billing database is presented in Table 1.

Table 1. Distribution of BMD billings in OHIP for EMRALD patients over age 40.

OHIP Fee Codes	**Fee Code Description**	**# (%)**
X146*, X145*	Baseline BMD test	1996 (9.9%)
X152, X153	Second BMD test, low risk patient	5250 (26.0%)
X142, X148	Subsequent BMD test, low risk patient	75 (0.4%)
X149, X145	Subsequent BMD test, high risk patient	12,853 (63.7%)
Total billings for tests		**20,174 (100%)**

**indicates a fee code that was added in 2008.*

Of the 20,174 BMD tests that were identified in OHIP, 89.8% could be matched with a corresponding test in EMRALD that shared the exact same date. When matching constraints were relaxed to allow for correspondences between billings and exams in EMRALD dated up to 30 days post-billing, the match rate increased to 97.9%.

In EMRALD, a total of 21,553 BMD exams were located for 15,365 patients, resulting in an average of 1.4 BMD tests per patient between 2006 and 2011. Of these tests, 84.1% could be matched with an OHIP billing that shared the exact same date. Using relaxed matching constraints (i.e., when allowing matches with billings prior to EMR dates), the match rate increased to 91.7%. 829 patients with BMD tests in EMRALD (5.4% of all patients with BMD tests in EMRALD) were found to have no corresponding billings for BMD tests in OHIP.

A total of 72.6% of the BMD tests in EMRALD were stored as text and 66.5% contained the word ‘T-score’ specifically. However, the number of BMD tests stored as text varied widely from practice to practice. The mean percentage of tests stored as text at each clinic was 65% with a standard deviation over 35% and a range that spanned 0% to 100%.

Relationship Between BMD Results and Other Osteoporosis Information

The randomly selected group of 7500 electronic records represented a group of patients with a mean age of 49.6 years, 57.3% of whom were female.

A total of 441 charts (5.9% of the total) were found to contain evidence of osteoporosis. The average age of the patients in the group with osteoporosis was 71.0 years, and 87.3% were female. By far the most common way of documenting osteoporosis was in the context of individual patient exam entries (such as in

an entry for a recent annual exams). A total of 398 cases of osteoporosis, or 90.2% of the total, were documented this way. Only 246 charts, or 55.8% of the osteoporotic group, were found to have evidence of osteoporosis that was substantiated with results from BMD exams. A total of 277 (or 62.8% of the group) contained a diagnosis of osteoporosis in an electronic patient profile.

Results are summarized in Table 2.

Table 2. Documentation of patients with osteoporosis (n = 441)

Documentation found in:	**# (%)**
Notes for Patient Exam(s)	398 (90.2%)
Patient Profiles	277 (62.8%)
BMD Exam Result	246 (55.8%)
Consultation Note (Specialist Consult Note, Emergency Consult Note, etc.)	38 (8.6%)

Documentation of Osteoporosis among Post-Fracture Patients

A total of 1473 fractures were documented in the charts; these were found in the charts of 1048 individual patients (15.9% of the sample). Charts for 151 of these individuals *specifically* mentioned a history of low-trauma fracture; these patients were therefore labeled "fragility fracture" patients. Of these patients, no record of BMD investigation could be located for 55, or 36.4% of the group. For several of these patients, however, alternative evidence was located to indicate consideration of osteoporosis (i.e. notes in patient profiles, consult notes, etcetera). There were a total of 36 fracture patients (23.8% of all fragility fracture patients) for whom no evidence of osteoporosis consideration could be located (i.e. no BMD results or other charted notes).

Results are summarized in Table 3.

Table 3. Osteoporosis documentation among patients with fragility fractures

	# (%)
Patients with fragility fracture	151 (100%)
Fragility fracture patients without evidence of BMD investigation	55 (36.4%)
Fragility fracture patients without evidence of OP investigation	36 (23.8%)

Discussion

Completeness of BMD Exam Capture

The present study indicates that EMR records capture BMD results completely relative to administrative data sources, and that the information required for 'gold-standard' diagnosis of osteoporosis is readily accessible to machine-assisted interpretation in a significant proportion of these results.

In a 2005 study of osteoporosis information in EMRs, a relative shortage of BMD tests was found across the EMR records for 78 practices[9]. In that prior work, only two practices could be located with more than 200 BMD test records on file[9]. Our study of data from 31 clinics, by contrast, identified several comparable practices averaging more than 200 BMD tests over similar stretches of time. More importantly, our results indicate that EMRALD captures BMD exams found in the OHIP database, which are the majority of those that take place in the province. In addition, EMRALD was found to capture records of some BMD tests that do *not* appear in OHIP billing records. Errors in test categorization and a lack of perfect synchronization between the EMR and OHIP may account for some of these extra tests, but EMRALD likely captures additional valid test results, such as those for tests performed outside of Ontario or billed to alternate insurers.

This evidence suggests that key details regarding BMD tests are reliably being provided to and/or labeled in Ontario's EMRs. The fact that 97% of OHIP billings for BMD tests could be matched with exam data in EMRALD contrasts with recent analyses of EMRALD's capture of laboratory tests and prescription information; roughly 67% of EMRALD's data could be matched with administrative records for these services[4]. BMD exams are, however, somewhat unusual in that they are typically ordered directly by family physicians while other imaging and laboratory services (like x-rays) may be ordered by specialists. Increasing comfort with EMRs on the part of clinicians and improvements in interfaces between EMRs and diagnostic imaging labs are likely contributors to the BMD exam coverage in EMRALD. According to the National Physician Survey, in 2010, 40% of family physicians in Ontario reported an electronic interface to diagnostic imaging services and this percentage is climbing[5].

In the 2005 study, moreover, BMD exams could not be associated with the numeric T-scores required for diagnosis of osteoporosis[9]. Results from our study, however, indicate T-scores to be somewhat more accessible, as the word 'T-score' could be automatically drawn from more than 65% of EMRALD's entries for BMD exams. Promising related work has taken additional steps toward automated BMD exam result interpretation: in a study of EMR records for post-menopausal female veterans from the U.S., T-scores were not only pulled from 63% of the exams stored in an EMR but were correctly associated with regions of interest and diagnoses more than 80% of the time[14]. Like the 2005 study, however, the present study documents variation of T-score accessibility by practice. The percentage of BMD reports stored as images rather than text at each practice ranged the entire spectrum (from 0 to 100%).

In summary, EMRALD was found to record incidence of BMD testing comprehensively relative to administrative data; accessibility of the T-scores in test results means that EMRALD may soon be able to contribute to the estimation of key health status indicators in the Canadian population, including prevalence of both BMD test outcomes and osteoporosis[15]. Related research to validate the Canadian Association of Radiologists and Osteoporosis Canada guidelines in the Canadian population required BMD data from roughly 4,000 Manitoban patients per year (or 16,000 patients across 4 years)[16]. The size of population accessible to diagnosis of osteoporosis via BMD results in EMRALD is slightly smaller but comparable.

Relationship Between BMD Results and Other Osteoporosis Information

While BMD results were captured completely by EMRALD, they were not found in all charts of patients with a diagnosis of osteoporosis. By far the most common way in which osteoporosis was documented was in the notes for individual patient exams, like annual exams. About 63% of the time osteoporosis information was found in patient profiles and only about 56% of the time the diagnosis was substantiated with a BMD result.

This is not a surprising result, as some patient charts in EMRALD span a relatively short period of time. Many individuals were likely investigated for osteoporosis prior to their appearance in EMRALD; results substantiating their diagnosis may not be available to the EMR in electronic form. The BMD records in EMRALD, then, may potentially identify osteoporotic individuals with sensitivity, but not with specificity. As EMR usage continues, however, it will most likely capture more of the BMD information required to substantiate those diagnoses of osteoporosis that are on file.

The fact that entries for individual exams proved to be the most common form of osteoporosis documentation is somewhat troubling, perhaps, as notes from individual patient exams tend to become buried over time[17]. Patient summaries, by contrast, were specifically designed to hold persistent clinical information in an accessible fashion and have recognized potential to facilitate coordination of care related to chronic conditions[17]. However, osteoporosis diagnoses were found in this location only 63% of the time.

Documentation of Osteoporosis among Post-Fracture Patients

In the 2005 study of osteoporosis and fracture information in EMR records, few if any fragility fractures were identified[9]. In this study, by contrast, a significant number of fractures were identified but less than 15% could reliably be identified as fragility fractures.

In the present study, 36 of those individuals identified as fracture patients were found to have no documented consideration of osteoporosis in their charts; documentation was defined to include BMD test results and other clinical notes. This is somewhat concerning as currently guidelines state that all fracture patients should be screened for osteoporosis; those without a history of BMD testing should receive a BMD test[2]. However, the result is consistent with evidence that suggests fragility fracture patients are inadequately screened in general. In a 2005 review of osteoporosis management after fracture, for example, less than 50% of patients with fragility fractures followed up with a doctor and investigations for osteoporosis were performed in less than 15% of the group[18]. In a more recent analysis of Ontario's BMD testing rates, less than 20% of fragility fracture patients received a BMD test within 6 months of their fracture[4].

These prior results, however, were not specific to patients in the care of a family physician and of known fracture status. The fact that the current study focuses on a specific care pathway may partly account for the high proportion of fracture patients in EMRALD with records of BMD testing (63.5%) relative to the proportion reported for the Ontario population[4]. Moreover, evidence tells us that testing rates for this population and within this care pathway can be altered using electronic reminders. A 2006 study of fracture patients identified using HMO billings found 40% were tested at six months post-fracture when family physicians received electronic reminders; without reminders testing rates were under 2%[19]. The population focus in the current study, however, focuses on fractures known to family physicians, known to be low-trauma, and identified based on annotations within EMR records.

Limitations

There are several limitations to the current study. For one, while we know that BMD capture in EMRALD records is complete, we do not know the accuracy of any other information about osteoporosis or fracture. Documentation of fractures in the EMR records is particularly problematic, as fragility fractures can be very difficult to characterize. Vertebral fractures, for example, may be difficult to determine given both the lack of fracture recognition by clinicians and the ambiguous terminology in radiology reports[20]. Doctors and patients may also note fractures less forcefully and frequently, as they are transient conditions[21]. Many fragility fractures found in administrative data may therefore be missing from EMRALD records. It seems relatively probable, however, that those fractures specifically noted in EMRALD as "fragility fractures" are indeed the result of low-impact events.

EMRALD's documentation of fragility fractures is also limited in that it often excludes specific fracture dates. History of fracture is sometimes noted in patients' problem lists, for example; dates provided here may be vague or in the relatively distant past. This limits our ability to determine if BMD exams were ordered after a given fracture as guidelines recommend. Direct validation of the fractures documented in EMRALD against administrative data sources, like emergency room records and physician billings, is subject for future research; this will include validation of fracture dates.

In addition, some patients may have been screened for osteoporosis at clinics not represented in EMRALD or at times prior to the commencement of their EMR record such that these BMD results may not be present in EMRALD as a result. This is likely why many diagnoses could not be substantiated by a corresponding BMD result, despite the fact that EMRALD's coverage of BMD results after 2006 is complete.

Finally, the sample of EMR records in EMRALD only reflects practices that make use of a single EMR system (Practice Solutions); results from EMRALD may therefore not generalize to data from other EMR databases. However, EMRALD has been shown to accurately capture the presence of other medical conditions in a way that reflects province-wide rates[22].

Despite these limitations, EMRALD remains a promising source of information about osteoporosis and fracture in the Ontario population, and it holds the potential to contribute to the estimation of important health status indicators. Coverage of BMD results, in particular, is significantly more complete than it appeared to be in 2005 and T-scores required for diagnosis are accessible to machine assisted interpretation

in a significant proportion of these. Moreover, data in EMR charts may shed light on key osteoporosis management indicators in the context of primary care. Results presented here, for example, demonstrate the EMR records to contain information about screening for osteoporosis among fracture patients and on the part of family physicians.

References

1. Jaglal SB, Hawker G, Cameron C, Canavan J, Beaton D, Bogoch E, Jain R, Papaioannou A, for the Osteoporosis Research, Monitoring and Evaluation Working Group. The Ontario Osteoporosis Strategy: implementation of a population-based osteoporosis action plan in Canada. Osteoporos Int. 2010;21(6):903-8.
2. Papaioannou A, Morin S, Cheung AM, Atkinson S, Brown JP, Feldman S, Hanley DA, Hodsman A, Jamal SA, Kaiser SM, Kvern B, Siminoski K, Leslie WD, for the Scientific Advisory Council of Osteoporosis Canada. 2010 clinical practice guidelines for the diagnosis and management of osteoporosis in Canada: summary. CMAJ 2010; 182(17):1864-73.
3. Leslie WD, Lix LM, Yogendran MS. Validation of a case definition for osteoporosis disease surveillance. J Clin Epidemiol. 2008; 61(12):1250-60.
4. Jaglal S, Hawker G, Croxford R, Cameron C, Schott A-M, Munce S, and Allin S. Impact of a change in physician reimbursement on bone mineral density testing in Ontario, Canada: a population-based study CMAJO 2:E45-E50
5. National Physician Survey, College of Family Physicians of Canada, Canadian Medical Association, Royal College of Physicians and Surgeons of Canada. http://www.nationalphysiciansurvey.ca. Accessed Sept 16, 2013.
6. Chan KS, Fowles JB, Weiner JP. Review: electronic health records and the reliability and validity of quality measures: a review of the literature. Med Care Res Rev. 2010;67(5):503-27.
7. Brouwer H, Bindels P, Weert H. Data quality improvement in general practice. Fam Pract 2006, 23(5):529–536.
8. Whitelaw FG, Nevin SL, Milne RM, Taylor RJ, Taylor MW, Watt AH. Completeness and accuracy of morbidity and repeat prescribing records held on general practice computers in Scotland. Br J Gen Pract 1996, 46 (404):181–186.
9. de Lusignana S, Chanb T, Wooda O, Haguea N, Valentinc T, Van Vlymen J. Quality and variability of osteoporosis data in general practice computer records: implications for disease registers. Public Health 2005; 119: 771–780.
10. Tu K, Mitiku T, Lee DS, Guo H, Tu J. Validation of physician billing and hospitalization data to identify patients with ischemic heart disease using data from the electronic medical record administrative linked database (EMRALD). Can J Cardiol 2010; 26(7): e225-e228.
11. Ivers N, Pylypenko, B, Tu, K. Identifying Patients With Ischemic Heart Disease in an Electronic Medical Record. J Prim Care Community Health, 2011; 2(1):49-53.
12. Tu K, Mitiku TF, Ivers NM, Guo H, Lu H, Jaakkimainen L, Kavanagh DG, Lee DS, Tu JV. Evaluation of Electronic Medical Record Administrative data Linked Database (EMRALD). Am J Manag Care. 2014 Jan;20(1):e15-21.
13. Privacy Procedures at the Institute for Clinical and Evaluative Sciences (ICES). http://www.ices.on.ca/webpage.cfm?org_id=119. Accessed September 14, 2013.
14. LaFleur J, Ginter T, Curtis J, Adler R, Agodoa I, Stolshek B, Nelson R, DuVall S A novel method for obtaining bone mineral densities from a dataset of radiology reports and clinic notes: Natural language processing in a national cohort of postmenopausal veterans (abstract). American Society for Bone Mineral Research Annual Conference, Baltimore, MD, USA. November, 2013.
15. LeMessurier J, O'Donnell S, Walsh P, McRae L, Bancej C, Osteoporosis Surveillance Expert Working Group. The development of national indicators for the surveillance of osteoporosis in Canada. Chronic diseases and injuries in Canada 2012; 32(2): 101-7
16. Leslie WD, Fang J, Lix L. Simplified System for Absolute Fracture Risk Assessment: Clinical Validation in Canadian Women. J Bone Miner Res. 2009;24(2):353-60.
17. Lewis J. Cumulative Patient Profile, Can Fam Physician. 1989; 35: 1259–1261.
18. Giangregorio L, Papaioannou A, Cranney A, Zytaruk N, Adachi J. Fragility Fractures and the Osteoporosis Care Gap: An International Phenomenon. Semin Arthritis Rheum. 2006;35(5):293-305.

19. Feldstein A, Elmer PJ, Smith DH, Herson M, Orwoll E, Chen C, Aickin M, Swain MC. Electronic medical record reminder improves osteoporosis management after a fracture: a randomized, controlled trial. J Am Geriatr Soc. 2006 Mar;54(3):450-7.
20. Delmas PD, van de Langerijt L, Watts NB, Eastell R, Genant H, Grauer A, Cahall DL, for the IMPACT Study Group. Underdiagnosis of vertebral fractures is a worldwide problem. J Bone Miner Res 2005 Apr;20(4):557-63.
21. Sale J, Beaton D, Sujic R, Bogoch ER. 'If it was osteoporosis, I would have really hurt myself.' Ambiguity about osteoporosis and osteoporosis care despite a screening programme to educate fragility fracture patients. J Eval Clin Prac 2010: 16(3): 590-96.
22. Tu K, Manuel D, Lam K, et al. Diabetics can be identified in an electronic medical record using laboratory tests and prescriptions. J Clin Epidemiol. 2011;64(4):431-5.

Development of an Alert System to Detect Drug Interactions with Herbal Supplements using Medical Record Data

Melissa Archer, PharmD[1,2], Joshua Proulx, BS[1,3], Laura Shane-McWhorter, PharmD[2], Bruce E. Bray, MD[1,3], Qing Zeng-Treitler, PhD[1,3]
[1]VA Salt Lake City Health Care System, Salt Lake City, UT; [2]University of Utah College of Pharmacy, Salt Lake City, UT; [3]University of Utah Department of Biomedical Informatics, Salt Lake City, UT;

Abstract

While potential medication-to-medication interaction alerting engines exist in many clinical applications, few systems exist to automatically alert on potential medication to herbal supplement interactions. We have developed a preliminary knowledge base and rules alerting engine that detects 259 potential interactions between 9 supplements, 62 cardiac medications, and 19 drug classes. The rules engine takes into consideration 12 patient risk factors and 30 interaction warning signs to help determine which of three different alert levels to categorize each potential interaction. A formative evaluation was conducted with two clinicians to set initial thresholds for each alert level. Additional work is planned add more supplement interactions, risk factors, and warning signs as well as to continue to set and adjust the inputs and thresholds for each potential interaction.

Introduction

Complementary and alternative medicine (CAM) refers to a group of therapies not usually practiced in conventional medicine. The National Institutes of Health (NIH) outlines categories of CAM including natural products (dietary supplements or probiotics), mind and body medicine (meditation, yoga, or acupuncture), manipulative and body-based practices (spinal manipulation, massage therapy), and other practices (traditional healers, traditional Chinese medicine, and other modalities).[1] CAM practices are not typically taught in medical schools and are not readily available in hospitals. Despite this, CAM usage is prevalent in the United States. Up to 40% of Americans use health care therapies considered outside of conventional medicine.[2] In some cases, clinical practice guidelines recommend CAM therapies in addition to conventional medicine. The American College of Cardiology Clinical Expert Consensus Guidelines on Integrating Complementary Medicine Into Cardiovascular Medicine (2005) recommend meditation, biofeedback, Omega-3 fatty acids, stanol/sterol ester margarines, and soy proteins as possible integrative CAM therapies in patients with cardiovascular diseases.[3]

While many CAM practices are considered safe (e.g. meditation, biofeedback), some dietary supplements may interact with conventional pharmacologic therapies, or with the disease state itself, and negatively impact patient outcomes. For example, clinical trials evaluating high dose vitamins (vitamin C > 2,000 mg/day; vitamin E > 1,000 mg/day), oleander, ephedra, and other herbs/botanicals with stimulant properties demonstrate adverse effects in patients with cardiovascular diseases.[3] Ginseng and St. John's Wort are herbal agents known to interact with cardiovascular drug therapies such as digoxin or warfarin.[4-6] Based on the clinical evidence available, CAM therapies should be used with caution in patients with cardiovascular diseases.

Despite the safety concerns, most patients do not discuss CAM use with their providers. Our prior study reported a large discrepancy between self-reported and provider-perceived usage-rate.[7] Lack of communication and documentation of CAM therapies make it difficult for clinicians to assess and mitigate the risk of herb-drug-disease interactions.[8] In addition, literature suggests both providers and consumers lack knowledge of the safety and efficacy of herbal medicine.

Our general goal is to alert physicians and patients of potential herb-drug interactions. Many automated alert systems tracking prescription drug-drug and drug-disease interactions are widely available and highly utilized in pharmacies and clinics. Indeed, automated alerts are one of the best-known and most successful examples of computer-based decision support.[9, 10] However, there is a need for an automated system which educates and alerts physicians and patients of the safety concerns regarding CAM therapies and herb-drug-disease interactions.

This paper describes a prototype herb-drug interaction alert system that we have developed. A formative evaluation was performed using data from the Veterans Health Administration's Electronic Health Record (EHR).

System Overview

The alert system consists of several key components: a module that looks for potential medication and supplement interactions based on medication or medication class, modules that calculate weights of various risk factors and warning signs, and a module that determines what level of alert should be given for each potential interaction. Data on herb/drug use, risk factor, warning signs may be obtained from an EHR or patients directly. Alerts may be provided to the users directly or delivered electronically (e.g. through email).

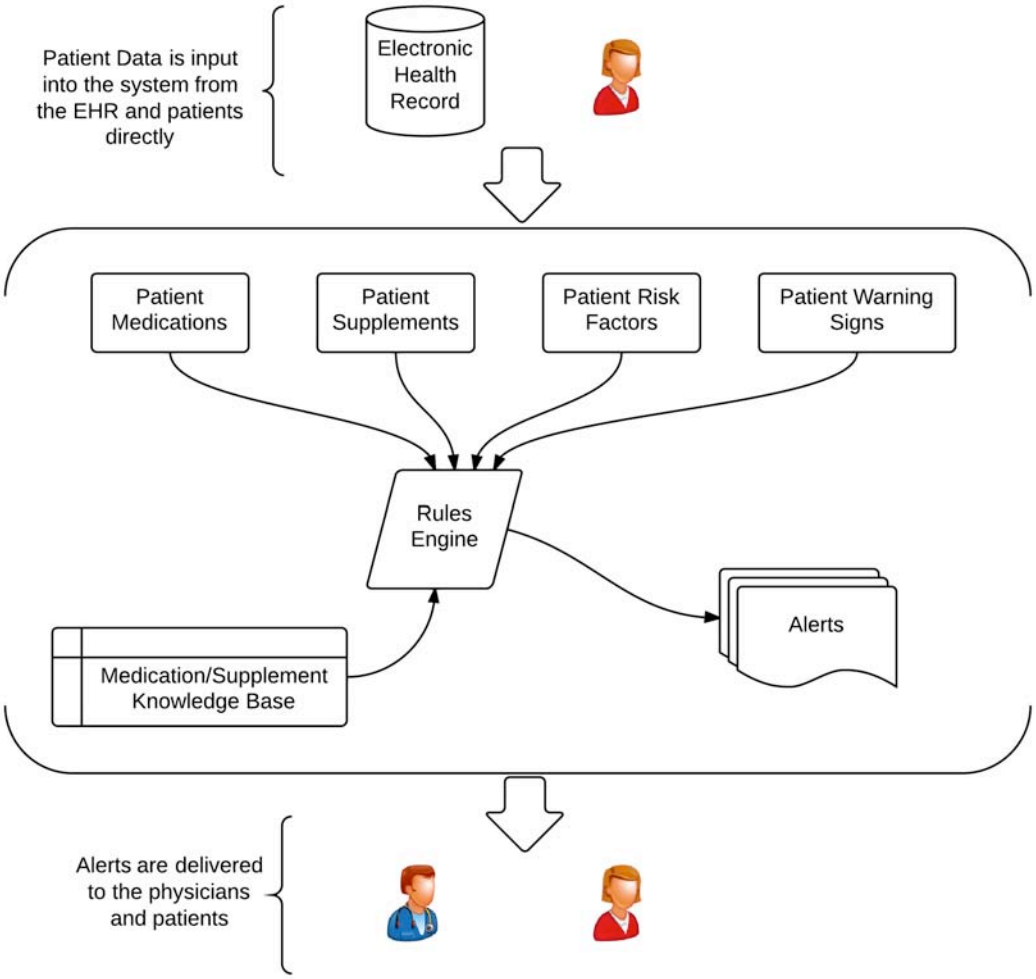


Figure 1. Knowledge base architecture.

Knowledge Base

A clinical pharmacist with expertise in drug information and drug utilization review and an informaticist software engineer developed the drug herb knowledge rules. To focus on common and important herb-drug interactions, nine of the most popular US supplements were identified. We assessed their potential interactions with common medications taken by cardiac patients at the University of Utah Health Sciences Center (Table 1). The nine herbal (non-vitamin, non-mineral) agents were identified, based on the National Health Statistics Report of Complementary and Alternative Medicine Use Among Adults and Children (US, 2007)[2] and the Herbal Gram[13]. Using the Natural Medicines Comprehensive Database (NMCD), current therapeutic guidelines, and available clinical evidence, a list

of potential herb-drug interactions was compiled and weighted based on both the likelihood and the severity of the interaction. Likelihood of interaction was determined by the strength of evidence available for the interaction (anecdotal vs. case reports vs. clinical trials, etc.) and severity was determined by the seriousness of the potential interaction as defined by the NMCD. The Natural Medicines Comprehensive Database is a compendium of information on over 1000 herbs and supplements and is based on clinical literature and details not only background information but potential herb-medication interactions. Other clinical evidence considered was based on primary literature.

Table 1. Herbal Agents and Medications Included in the Herb-Drug Interaction Database

Supplements	Medications		
Echinacea	amiodarone	Flecainide	nifedipine
Garlic	amlodipine	fluoxetine	omeprazole
Ginkgo	aspirin	fondaparinux	pantoprazole
Ginseng	atorvastatin	gemfibrozil	paroxetine
Grape seed	carvedilol	glipizide	prasugrel
Green tea	cilastazol	glyburide	pravastatin
Milk thistle	citalopram	hydrochlorothiazide	prazosin
Saw palmetto	clopidogrel	insulin aspart	propafenone
Soy	dabigitran	insulin glargine	propanolol
	dipyridamole	insulin lispro	ranitidine
	ditiazem	insulin regular	rivaroxaban
	dofetilide	irbesartan	rosuvastatin
	doxazosin	lansoprazole	sertraline
	dronedarone	losartan	simvastatin
	enoxaparin	lovastatin	tamsulosin
	ephedrine	metformin	ticagrelor
	eplerenone	metoprolol	torsemide
	escitalopram	mexiletine	valsartan
	esomeprazole	nebivolol	verapamil
	famotidine	nicardipine	warfarin

We took both pharmacokinetic and pharmacodynamic parameters into consideration when developing the knowledge base. For example, pharmacokinetic interactions occur if the absorption, distribution, metabolism, or elimination of a drug is affected. Most often, isoenzymes of the cytochrome P450 system are involved in pharmacokinetic interactions and may result in increased or decreased serum drug concentrations (SDCs) of the affected drug. This may then result in supratherapeutic or subtherapeutic drug effects. Pharmacodynamic interactions involve additive or oppositional pharmacologic effects. For instance bleeding may occur when two antiplatelet drugs are combined. However, decreased efficacy of an antihypertensive agent may occur when combined with a supplement that may increase blood pressure. In addition, using clinical evidence, patient-specific diagnoses and comorbidities which put the patient at higher risk for developing an herb-drug interaction were identified and used to generate personalized alerts.

To help determine the level of alerts for specific patients, risk factors and warnings of the potential interactions were identified. The risk factors and warnings are patient features that potentially increase the likelihood or severity of an interaction or increase the urgency of intervention. Table 2 lists risk factors and warnings in the system. Risk factors and warning signs can be assigned to all potential interactions, or specific interactions by drug or drug class.

Thresholds were set for each interaction based on the total score possible (including risk factors and warning signs) and likelihood/severity of the interaction.

Data was extracted from the Electronic Health Records (EHR) of patients who receive care at the Veterans Administration (VA) Medical Centers. Patients receiving one or more of the nine identified herbal agents along with one or more of the identified prescription medications were considered for inclusion. Herbal agents and prescription medications found via a text search in either the active VA drug list or in a clinic note were used. Patient-specific diagnoses and comorbidities identified as risk factors and warning signs for a drug-herb-disease interaction were defined by International Classification of Diseases (ICD-9) codes reported in the EMR. Risk factors and warning signs were selected based on clinical information gathered from the NMCD, clinical practice guidelines, and primary literature.

Table 2. Risk Factors and Warning Signs Used for Calculation of Personalized Alerts

Risk Factors	Warning Signs	
>3 comorbidities	Arrhythmia	Hypotension
>5 meds/day	Bleeding	Increased LFTs
Age >65	Bradycardia	Lactic acidosis
Arrhythmia	Bruising	Lethargy
Heart failure	Diarrhea	Myopathy
History of GI bleed	Dizziness	Orthostatic hypotension
Impaired hepatic function	Edema	QT prolongation
Impaired renal function	Electrolyte imbalance	Renal failure
Multiple QT prolongation medicines	Epistaxis	Rhabdomyolysis
Multiple serotonergic medicines	GI hemorrhage	Serotonin syndrome
Patient history of MI	Hearing loss	Signs of drug inefficacy
Recent stroke	Hyperkalemia	Somnolence
	Hypertension	Syncope
	Hypoglycemia	Tachycardia
	Hypokalemia	Thrombosis

Alert Generation

Nine herbal agents and 62 medications were selected for inclusion in the herb-drug database. Overall, 259 drug-herb-disease rules were built into the system.

Drug-herb-disease interactions identified by the new system were assigned one of three severity alert ratings: 3) Most Serious, Action Recommended; 2) Less or Moderately Serious (or “Monitor” and Action Should be Considered); and 1) Least Serious (No specific action is recommended but education is provided so that the clinician and patient can be aware of a potential interaction).

Formative Study

To test the alert system, verify thresholds, and inform further development, we manually extracted patient characteristics from the VA VINCI Database (http://www.hsrd.research.va.gov/for_researchers/vinci/). To identify patients with potential interaction, a computer query was first performed on all VA clinical notes for the nine herbal supplements described above. The text search was limited to the Active VA and Non VA Medication List section that is embedded in various VA notes such as Pharmacy Encounter Notes, Anticoagulant Clinic Notes, Clinic Notes, etc. We then searched the structured medication orders within the 30 day window of the documentation of herbal usage for medications with potential interaction. Twenty patients with potential interactions were randomly selected

for the initial formative evaluation. Potentially relevant patient characteristics such as age, gender and clinical diagnoses (based on ICD codes) were extracted.

A practicing cardiologist and a clinical pharmacist with expertise in herbal medicine were asked to evaluate the 20 patients cases for potential interactions. Without knowledge of the automated rules, they identified and evaluated all potentially harmful herb-drug interactions, risk factors, and warning signs for the alert as well as the level of severity for the alert. These expert-generated alerts were then compared to the automated herb-drug alerts.

Results

A total of 622 subjects taking at least one herbal agent were identified by the automated system. Twenty subjects with a total of 26 herb-drug interactions were randomly selected for evaluation (Table 3). The mean age of subjects was 68 years (range 51-85). All subjects were male. The mean interaction scores calculated by the herb-drug interaction database were 5.3 resulting in 11 Level one alerts, 10 Level two alerts and 5 Level three alerts. A Pearson's coefficient of 0.78 was calculated between the automated alert levels and the expert alert levels.

All 26 automated alerts generated by the system were deemed appropriate by the clinicians, i.e. excellent positive predictive value. During the manual review, the clinical team also identified 3 additional herb-drug interactions not originally identified by the automated system (Table 3). The missing alerts were due to lack of related rules. The team also identified 4 additional risk factors deemed important for identification and evaluation of the severity of the potential herb-drug interaction including: poly-herb use, smoking history, alcohol use and a dementia diagnosis. The expert-generated alerts produced 10 Level one alerts (34.5%), 15 Level two alerts (51.7%) and 4 Level three alerts (13.8%). Per clinician review, herb-drug combinations that were associated with a possible risk of hypoglycemia generated higher level alerts.

Overall, 15 unique herb-drug interactions were included in the analysis (Table 3). In this study of 20 patients, pharmacodynamic herb-drug-disease interactions were identified more frequently than pharmacokinetic herb-drug interactions (93.1% vs. 6.9%). The most frequently identified interaction was concurrent administration of the herb saw palmetto and an antiplatelet agent with the potential for a pharmacodynamic interaction resulting in an increased risk of bleeding (34.5%; n = 10). The second most frequently identified interaction pair was garlic with an anticoagulant or antiplatelet agent and a potential pharmacodynamic interaction associated with increased risk for bleeding (27.6%; n = 8). The most frequently identified risk factors based on medication lists and ICD-9 codes reported in the active problem list of the clinic note for the patent were > 5 medications (95%; n = 19) and age greater than 65 (60%; n = 12).

Table 3. Summary of Study Patients

Patient Number	Age	Number of Risk Factors	Herb/Supplement	Drug	Pharmaco-dynamic Interaction	CYP450 Interaction	Risk Score	Automated Alert Level	Expert-Based Alert Level
1	70-74	2	Saw Palmetto	Warfarin	X		6	2	2
2	50-54	1	Ginkgo Biloba	Sulfonylurea	X		5	3	3
2	50-54	1	Ginkgo Biloba	Metformin	X		5	2	2
3	70-74	2	Saw Palmetto	Aspirin	X		6	1	1
4	70-74	2	Garlic	Clopidogrel	X		6	3	3
5	65-69	2	Garlic	Aspirin	X		6	3	2
5	65-69	2	Garlic	Dipyridamole	X		6	3	3
6	65-69	4	Ginkgo Biloba	Warfarin	X		6	2	2
7	75-79	3	Saw Palmetto	Warfarin	X		5	2	2
7*	75-79	3	Ginseng	Warfarin	X		-	-	2
8	70-74	1	Saw Palmetto	Aspirin	X		5	1	1
8	70-74	1	Saw Palmetto	Dipyridamole	X		5	1	1
9	85-89	1	Garlic	Aspirin	X		5	1	1
9*	85-89	1	Garlic	Simvastatin		3A4	-	-	1
10	60-64	1	Garlic	Aspirin	X		5	1	1
10*	60-64	1	Garlic	Amlodipine		3A4	-	-	1
11	60-64	0	Garlic	Warfarin	X		4	1	1
11	60-64	0	Garlic	Aspirin	X		4	1	1
12	55-59	1	Green Tea	Aspirin	X		6	2	1

13	75-59	2	Saw Palmetto	Warfarin	X		5	2	2
14	80-84	3	Saw Palmetto	Warfarin	X		5	2	2
15	60-64	1	Ginseng	Metformin	X		6	2	2
16	50-54	1	Ginkgo Biloba	Sulfonylurea	X		5	3	3
16	50-54	1	Ginkgo Biloba	Metformin	X		5	2	2
17	60-64	1	Saw Palmetto	Aspirin	X		5	1	2
18	60-64	1	Garlic	Clopidogrel	X		4	1	2
18	60-64	1	Saw Palmetto	Clopidogrel	X		4	1	2
19	70-74	2	Ginseng	Metformin	X		7	2	2
20	85-89	2	Saw Palmetto	Clopidogrel	X		6	1	2

Key: *=additional alerts identified during expert-generated validation process

Discussion

This study demonstrates the feasibility of extracting data from the VA VINCI Database to clinically assess herb-drug-disease interactions. This was the first iteration of the development of an herb-drug-disease interaction knowledge base which is useful for personalized, automated alerts. The initial qualitative assessment demonstrates the complexity of identifying and evaluating potential herb-drug interactions in an EHR. In 20 selected patients, a total of 29 expert-generated herb-drug alerts were generated. Both pharmacodynamic and pharmacokinetic interactions were identified. The range of the severity of the alerts varied based on risk factors and warning signs present in the patients' EHR. The most frequently reported alert severity was a Level 2 or Moderately Serious alert ("Monitor," and Action Should be Considered). The least frequently reported alert severity was a Level 3 or Most Serious alert (Action Recommended). This is appropriate to mitigate alert-fatigue and is in line with current clinical evidence for the likelihood and severity of potential herb-drug-disease interactions.

When the clinical team met to identify and rate the severity of the herb-drug interactions, several issues emerged that warrant further evaluation and possible inclusion as additional risk factors. One risk factor that was identified was whether the patient was on five or more daily medications. However in examining patient profiles it was apparent that some patients were often using more than one herb or supplement. Thus poly-herb or poly-supplement combinations should be considered. Also, some patients were identified as smokers and since this may increase the risk for thromboembolic events this may be another important consideration. Furthermore, alcohol use may increase risk for hemorrhagic events, such as gastrointestinal bleeds, as well as hypoglycemic events, since alcohol use inhibits gluconeogenesis. Hence that may warrant inclusion as additional risk factors. Mention of memory impairment or early dementia should also be considered for risk factor inclusion since this may affect medication use behavior. Finally, codes for strokes should be clarified to determine if the event was ischemic or hemorrhagic since those are two different outcomes that could be part of the herb-drug interaction profile.

In future iterations of the system design, we will refine the rules engine formulas to improve data acquisition to gain a more comprehensive list of herb-drug interactions and include additional clinical data for identification of potential risk factors and warning signs. We will also customize the alert content based on intended recipient (physician, nurse or patient). We will continue to update the knowledge base based on new information on interactions from the literature. For example, clinical information regarding the hypoglycemic effects of many herbal agents (milk thistle, garlic, etc.) is becoming available and should be considered for inclusion in our decision support framework. This study demonstrates feasibility of using EHR data to assess potential herb-drug-disease interactions using a rule-based alerting system. Additional refinement of patient characteristics and interaction rules is needed.

References

1. Medicine NCfCaA. What is complementary and alternative medicine? *National Institutes of Health.* 2013. <http://nccam.nih.gov/health/whatiscam>. Accessed 6 March 2014.

2. Barnes PM, Bloom B, Nahin RL. Complementary and alternative medicine use among adults and children: United States, 2007. *Natl Health Stat Report.* Dec 10 2008(12):1-23.

3. Vogel JH, Bolling SF, Costello RB, et al. Integrating complementary medicine into cardiovascular medicine. A report of the American College of Cardiology Foundation Task Force on Clinical Expert

Consensus Documents (Writing Committee to Develop an Expert Consensus Document on Complementary and Integrative Medicine). *J Am Coll Cardiol.* Jul 5 2005;46(1):184-221.

4. Fasinu PS, Bouic PJ, Rosenkranz B. An overview of the evidence and mechanisms of herb-drug interactions. *Front Pharmacol.*3:69.

5. Izzo AA. Interactions between herbs and conventional drugs: overview of the clinical data. *Med Princ Pract.*21(5):404-428.

6. Izzo AA. Herb-drug interactions: an overview of the clinical evidence. *Fundam Clin Pharmacol.* Feb 2005;19(1):1-16.

7. Scarton LA, Zeng Q, Bray BE, Shane-McWhorter L. Feasibility and potential benefit of collecting Complementary and Alternative Medicine data through a computerized patient interview. *AMIA Annu Symp Proc.*2011:1217-1223.

8. Krasuski RA, Michaelis K, Eckart RE. The cardiovascular patient's perceptions of complementary and alternative medicine. *Clin Cardiol.* Apr 2006;29(4):161-164.

9. Hoonakker P, Khunlertkit A, Tattersal M, Keevil J. Computer decision support tools in primary care. *Work.*41 Suppl 1:4474-4478.

10. Tamblyn R, Huang A, Perreault R, et al. The medical office of the 21st century (MOXXI): effectiveness of computerized decision-making support in reducing inappropriate prescribing in primary care. *CMAJ.* Sep 16 2003;169(6):549-556.

11. Phansalkar S, Desai A, Choksi A, et al. Criteria for assessing high-priority drug-drug interactions for clinical decision support in electronic health records. *BMC Med Inform Decis Mak.*13(1):65.

12. Smithburger PL, Buckley MS, Bejian S, Burenheide K, Kane-Gill SL. A critical evaluation of clinical decision support for the detection of drug-drug interactions. *Expert Opin Drug Saf.* Nov;10(6):871-882.

13. Blumenthal M, Lindstorm A, Ooyen C, Lynch M. HerbalGram. *The Journal of the American Botanical Council.* 2011. <http://cms.herbalgram.org/herbalgram/issue95/hg95-mktrpt.html>. Accessed 14 March 2014.

Factors Contributing to CPOE Opiate Allergy Alert Overrides

Deborah Ariosto, PhD, RN
Vanderbilt University Medical Center, Nashville, TN

Context Increasing regulatory incentives to computerize provider order entry (CPOE) and connect stores of unvalidated allergy information with the electronic health record (EHR) has created a perfect storm to overwhelm clinicians with high volumes of low or no value drug allergy alerts. Data sources include the patient and family, non-clinical staff, nurses, physicians and medical record sources. There has been little written on how to collect hypersensitivity information suited for drug allergy alerting. Opiates in particular are a frequently ordered class of drugs that have one of the highest rates of allergy alert override and are often a component of pre-populated Computerized Provider Order Entry (CPOE) order sets. Targeted research is needed to reduce alert volume, increase clinician acceptance, and improve patient safety and comfort.

Design, Setting, and Patients An FY10 retrospective, quantitative analysis of 30321 unique adults with opiate allergies triggering CPOE alerts at a large academic medical center.

Measurements The prevalence of opiates ordered with opiate allergy alerts triggered and overridden is described. The effect of age, race, gender, visit type (medical, procedural), provider type (physician, advance practice nurse), and reaction/severity (e.g. nausea/mild) on the likelihood of provider override of the patient's first opiate alert was analyzed using Generalized Estimating Equations (GEE).

Results Analysis of a patient's first opiate allergy alert (n=2767) showed that only prescriber role had a significant effect on alert override compared with all other variables in the model. Advanced practice nurses (APNs) were generally less likely to override the patient's first opiate alert as compared to physicians (GEE, β=-.793, p=.001). However, override rates remained high, with 80% for APN's and 90% for physicians.

Over half of all discharges had opiates ordered during their stay. Of those, 9.1% of the patients had recorded opiate allergies triggering 25461 CPOE opiate allergy alerts. The largest sub-group of alerts was triggered by gastrointestinal (GI) "allergies" such as nausea and constipation. Removing these types of non-allergic, low severity GI reactions from the alert pool reduced the first alert volume by 15% and the overall alert volume by 22%. Of note is that a history of codeine allergy triggered a significant volume of opiate alerts, yet was rarely ordered.

Conclusion With an increasingly complex, information dependent healthcare culture, clinicians do not have unlimited time and cognitive capacity to interpret and effectively act on high volumes of low value alerts. Drug allergy alerting was one of the earliest and supposedly simplest forms of CPOE clinical decision support (CDS), yet still has unacceptably high override rates. Targeted strategies to exclude GI non-allergic type hypersensitivities, mild overdose, or adverse effects could yield large reductions in overall drug overrides rates. Explicit allergy and severity definitions, staff training, and improved clinical decision support at the point of allergy data input are needed to inform how we process new and re-process historical allergy data.

Introduction

In 2009 President Obama signed into law the Health Information Technology Economic and Clinical Healthcare Act (HITECH), which significantly ramped up the investment in this nation's health care infrastructure by providing significant reimbursement incentives for "meaningful use" of electronic health record (EHR) technology. Computerized provider order entry (CPOE) adoption and drug-allergy checking, were two of the meaningful use objectives described for phase I. With increasing incentives to implement these objectives comes increasing need to mitigate adverse, unintended consequences of this evolving technology.

While there is evidence to support this safety measure, numerous alert fatigue studies cite the excessive alert volume and low clinical value as significant factors in intentional and unintentional alert overrides (van der Sijs, 2009). Two classes of drugs, antibiotics and opiates, represent the majority of CPOE alerts and alert overrides (Hsieh et al., 2004; Hunteman et al., 2009). This study focused on the latter. Opiates are not usually associated with the life-threatening allergic reactions (i.e. anaphylaxis, angioedema) as are antibiotics. Previous studies have identified that from 31-80% of the patients with opioid/narcotic allergies are labeled inappropriately (Gilbar 2004, Pilzer 1998). Commonly reported are known side effects or mild overdose due to potentiation with other drugs.

The European Academy of Allergy and Clinical Immunology proposed a definition, "Hypersensitivity causes objectively reproducible symptoms or signs, initiated by exposure to a defined stimulus at a dose tolerated by normal subjects" (Johansson I, 2005). It is doubtful that this definition is understood and consistently applied in the medication allergy history by those outside of the allergy field of medicine.

The increase in opiate allergies is likely the by-product of a longitudinal EMR that has moved unconnected, and often unsubstantiated, allergy records into a centralized data repository. It is facilitated by lack of staff education and reinforced by data entry screen design. Many EMRs have an associated data input screen named "Allergy" that was designed to store allergy data, but has expanded to collect any possible adverse drug reaction. The more appropriate term purposed for CPOE is "drug hypersensitivity" which is the umbrella term that includes both allergic (immune system mediated) and non-allergic types of reproducible adverse drug reactions.

Study objectives

The purpose of this study was to identify factors that contribute to high volume, low value alerts that are consistently overridden and which pose minimal patient safety concerns. These low value alerts represent known drug side effects of low/mild severity as well as low/mild hypersensitivities. Opiate alerts represent the most common over-alerting problem in the literature and comfort management is a significant component of clinical care in most settings. Comfort management refers not only to timely and effective pain relief, but management of undesirable opiate side effects such as nausea and constipation. Patient and prescribing provider attributes were also evaluated to determine if factors other than the allergy reaction influenced opiate allergy alerting.

Meaningful Use Stage I was about widespread adoption of the technology. However, the success of subsequent decision support depends heavily upon its ability to deliver trusted, clinically significant, actionable information at the point of care. The following exemplar highlights the need for this study:

> *A patient was urgently admitted and underwent open heart surgery. In recovery, an order was placed for needed pain medication for this intubated, agitated patient. A narcotic was ordered and the CPOE system alerted the provider that the patient was allergic to the ordered class of*

narcotics. Review of the documented allergy history showed an adverse reaction to codeine, but the reaction and severity were not documented. Family was unavailable. Nurse, physician and pharmacist conferred, during which time patient experienced increasing distress. The decision to give morphine was made, the patient was monitored closely, and no adverse response was observed. When questioned prior to discharge, the spouse had told the admission nurse that he "felt funny" after taking cough syrup with codeine a couple of years ago. The patient was discharged, and his electronic allergy history was not updated.

Allergy alerts differ from other types of alerts in that they often depend upon patient recall rather than those that depend on discrete lab values, diagnoses, or known interactions with other drugs. Allergy data is collected by clerical and/or clinical staff across multiple venues (clinic, inpatient, doctor's office) and during care encounters of varied intensity (routine, urgent) from patients or their surrogates (spouse, aide, etc.). This suggests that allergy alerts, while simpler to program, may be more complex to successfully to design and implement than other types of drug alerts.

Setting

The study was conducted, with IRB permission, at a large urban, southeastern academic medical center with approximately 40,000 FY10 adult inpatient discharges. The center has a large biomedical informatics department, and extensive clinical systems development capacity. It has a long history of electronic medical record (StarPanel[tm]) and inpatient CPOE applications (WizOrder[tm] , McKesson/Horizon Expert Orders[tm]). Two internally developed data input applications supplied allergy data which were used in both inpatient and outpatient settings. All allergy data, regardless of originating system is collected by a patient summary service (PSS) application that makes this data available to CPOE and other decision support decisions. The CPOE drug alert logic is commercially supplied by First DataBank, Inc. [tm]

Methods

This was a retrospective, quantitative analysis of 25,461 CPOE opiate allergy alerts across 3,473 discharges (2,767 unique patients). Prevalence of opiate allergy alerts across all FY10 opiate orders was calculated. Descriptives of related attributes included: (1) patient (age, race, sex), (2) prescriber alert override (APN, physician), (3) medical or procedural visit, (4) reaction descriptions (ie. nausea) and (5) severity (mild, moderate, severe). Reaction and associated severity were combined to create a new allergy variable called NALS (Non-Allergic/Low Severity) with three mutually exclusive groups: NALS, Not NALS, and Unknown.

The next step was to extract all opiate orders (n=153,026) from the CPOE orders database in FY10 based on the American Hospital Formulary System (AHFS) opiate class code =28080800. These orders were matched to the MRN to calculate the percentage of discharges with opiates ordered. Of these, physician (82%) and Advance Practice Nurse (15.4%) prescriber roles were retained, excluding orders attributed to other staff (2.6%).

In the third step, the opiate allergy alert response log by prescriber ID was extracted from the CPOE allergy alert dataset. This log recorded whether the prescriber cancelled the order or overrode the alert. The two CPOE alert responses were (1) Override (place order) or (2) Cancel order (accept alert). The reason for override field was available, but was excluded since it was 99% missing.

To reduce the dataset, and minimize the influence of excessive repeating alerts for some patients with long lengths of stay, only the patient's first opiate allergy alert was used in the regression analysis for each patient.

The fourth step was to assign allergy status: allergic vs. non-allergic. Each reaction within each group was evaluated (Table 1). Each patient was assigned to one of the study categories (NALS, Not_NALS, Unknown) using the coded severity and opiate reaction data in the patient allergy file. The following algorithm was used, followed by visual confirmation and assignment to one of the three following:

- Unknown: Both reaction and severity are blank or stated unknown
- NALS (GI): Reaction is nausea, vomiting, constipation, diarrhea, or GI upset and severity is low or unknown
- Not_NALS: Assign everything else to Not_NALS

Table 1 Characteristics of Opiate Reactions by severity and by group

Patients		Reaction	Severity			
#	%		Mild	Moderate	Severe	Unknown
1323	28%	Other	1%	2%	7%	91%
1142	24%	Skin	4%	7%	15%	73%
1046	22%	Gastrointestinal	2%	9%	15%	74%
504	11%	Nervous/Mood	2%	9%	22%	68%
781	16%	Unknown/Blank	0%	0%	0%	99%
4796	100%	All	2.5%	6.5%	16.8%	74.3%

Note. Some patients had more than one reaction

The following is an explanation of how reactions were assigned to or excluded from the non-allergic/low severity (NALS) category.

Skin (24%) - The predominant reaction within the skin category was itching and rash. While there were mild reactions recorded, this group as a class was excluded from the NALS category, since it is difficult to tell if these skin reactions were histamine responses of a pseudo-allergy (NALS), or a true, immune mediated allergic reaction (Not_NALS).

Nervous/Mood (11%) - Those in this category ranged from nervousness, insomnia, confusion, hallucinations, to suicidal ideation. In this group, there were very few coded as mild severity. There were only 2 coded mild reaction types that could have been included in the NALS group (sleepy, headache) – but these were excluded due to low volume compared to the class.

Other Reactions (28%) – The information in this group was difficult to classify by body systems. It was highly variable, and had low volumes in any one category. It also included non-allergy type data such as:

Patient received Narcan in the past (respiratory rescue)
Patient has history of opiate abuse
Patient has liver failure, dose appropriately
Patient allergic to IV morphine, but can tolerate oral

Patient has stomach ulcers, may bleed from NSAIDS

Gastrointestinal (22%) - The predominant reaction in the GI category was nausea and vomiting. Those coded or described within the free text as moderate or severe reactions were assigned to Not_NALS category. Those GI Reactions with mild, unknown, or blank severity were assigned to the NALS group if the descriptions were nausea, vomiting, upset stomach, constipation, diarrhea, or GI upset. The assumption being that if it were more than a mild reaction, it would likely been have recorded. More severe reactions such as bleeding or ulcer (exacerbation) were excluded from NALS.

Results

Four allergies triggered 82% of the opiate allergy alerts: Allergies to Codeine (32%), Morphine (28%), Hydrocodone (11%), and Oxycodone (11%). Table 2 displays the number of discharges with a specific opiate allergy and the number of alerts triggered.

Table 2 Patient Allergies Triggering Alerts

# Allergic	%	Allergy	# Alerts	%
1600	36%	Codeine	8,163	32%
1079	24%	Morphine	7,152	28%
483	11%	Hydrocodone	2,783	11%
427	10%	Oxycodone	2,773	11%
211	5%	Hydromorphone	1,190	5%
147	3%	Tramadol	971	4%
61	1%	Nubain	595	2%
148	3%	NSAID/OTC	541	2%
70	2%	Butorphanol	536	2%
104	2%	Meperidine	306	1%
105	2%	Propoxyphene	239	1%
10	0%	Opioid	88	0%
19	0%	Fentanyl	60	0%
2	0%	Dihydrocodeine	23	0%
2	0%	Oxymorphone	21	0%
11	0%	Other	20	0%
4479	100%		25,461	100%

Note: Some patients may have more than one recorded opiate allergy

Four drugs triggered 96% of opiate alerts: Hydromorphone (34%), Oxycodone (25%), Morphine (19%), and Hydrocodone (18%).

Table 3 Drug Orders Triggering Alerts (all ages)

Drug Order Group	#	%
Hydromorphone	8,599	34%
Oxycodone	6,395	25%
Morphine	4,916	19%
Hydrocodone	4,590	18%
Tramadol	435	2%
Fentanyl	354	1%
Belladonna-opium suppository	82	0%
Acetaminophen w Codeine	34	0%
Meperidine	20	0%
Codeine	16	0%
Propoxyphene (Darvon)	10	0%
Methadone (Dolophine)	10	0%
Total	25,461	100%

Reaction and Severity

Severity (mild, moderate, severe) was missing for 74.3% of the reactions. Of those, 16% had no reaction or severity. Overall, alerts with severity coded or described are 2.5% mild, 6.5% moderate, and 16.8% severe.

Non-Allergic/Low Severity Reactions (NALS)

First alerts were stratified into 3 allergy reaction/severity classes as previously described. Non-allergic/low severity (NALS) GI reactions accounted for 15.4% of the first alerts. The override rate for the patient's first alert for all opiate alerts was 89%. Removing the GI NALS alerts reduced the first alert volume for opiates by 15% (425/2767) but did not significantly change the 89% overall alert override rate.

Influence of patient, prescriber and reaction/severity factors on override

Patient characteristics, reason for admission, prescriber role and reaction/severity group on opiate alert overrides were analyzed . The Generalized Estimating Equations (GEE) procedure (SPSS v20) was used to extend the generalized linear model to allow for analysis of clustering and repeated measurements. Clustering can happen when a physician may treat primarily a specialty population like orthopedics or cancer that may unduly influence outcomes based on opiate use. Repeated measurements over time may reflect the fatigue that occurs when a provider has a lot of alerts over a short period of time. The dependent variable of interest was the override response (yes, no), so a binary logistic model was selected, with "no override" as the reference category. There were 697 providers (89% Physicians, 11% APN) who had from 1-28 first opiate allergy alerts (mode=1, median= 3).

For inpatients of all ages, 53% had opiates ordered (n=153026 orders). Inpatient opiate allergy prevalence was 9.1 % (2767/30321 patients). This number is understated as opiate allergic patients without opiates ordered were not captured. For those with opiate allergy alerts, the mean age was 54 years, 81.2% white and 68.6% female. For all admissions, medical discharges were higher (56%) than procedure related ones (44%). However, for the patient's first visit, procedural admissions were higher (59.8%). This reversal was expected as chronic medical conditions are more likely to have readmissions. Physicians received 91.1% of the first alerts and overrode 90% of them. Advance Practice Nurses (APN) received 15% of the first alerts and overrode 80% of them. A summary of the study sample characteristics for the patients first opiate allergy alert are shown in Table 4 below.

Table 4 First Opiate Allergy Alert Response

			Override = No		Override = Yes	
	N	%	N	%	N	%
All	2767	(100%)	304	(11%)	2463	(89%)
Gender						
Female	1900	69%	222	(12%)	1678	(88%)
Male	867	31%	82	(9%)	785	(91%)
Race						
Black	302	11%	30	(10%)	272	(90%)
Other	80	3%	11	(14%)	69	(86%)
White	2385	86%	263	(11%)	2122	(89%)
Reason for Admission						
Medical	1112	40%	111	(10%)	1001	(90%)
Procedural	1655	60%	193	(12%)	1462	(88%)
Allergy Reaction/Severity						
Unknown	699	25%	78	(11%)	621	(89%)
NALS	425	15%	36	(8%)	389	(92%)
Not NALS	1643	59%	190	(12%)	1453	(88%)
Prescriber Role*						
APN	246	9%	49	(20%)	197	(80%)
Physician	2521	91%	255	(10%)	2266	(90%)
			Mean	(SD)	Mean	(SD)
Patient Age in Years			54.7	16.7	54.5	16.4

**Significant p=.001* *NALS = Non-Allergic/Low Severity alert*

Summary

Over half of all discharges had opiates ordered during their stay. Of those, patients with recorded opiate allergies (9.1%) triggered 25461 CPOE opiate allergy alerts. The largest sub-group of alerts was triggered by gastrointestinal (GI) "allergies" such as nausea and constipation. Removing non-allergic, low severity GI reactions from the opiate alert pool reduced the first alert volume by 15% and the overall alert volume by 22%.

Opiate allergy override rate was 93% for all admissions and re-admissions. It was 89% for the first admission's alert. In the GEE analyses, provider role was the most significant variable in predicting alert overrides. Advanced practice nurses (APNs) were generally less likely to override the patient's first opiate alert as compared to physicians (GEE, β=-.793, p=.001). However, override rates remained high, with 80% for APN's and 90% for physicians. Other factors were not statistically significant in predicting override.

The study revealed two related phenomena in the evaluation of patient opiate allergy alerts. The first being that there was a high prevalence of opiate allergies recorded and alerted on for a reportedly rare occurrence. Particularly problematic were recalled allergies to codeine, which trigger alerts to any opiate order. Findings suggest that this is due, in part, from inappropriately broad allergy definitions and use of the allergy data collection field to alert the prescriber to other non-allergy clinical concerns.

The second was the high volume of common opiate side effects recorded as allergic reactions. Of particular importance was the high volume of overridden opiate allergy alerts triggered by non-allergic/low severity gastrointestinal reactions. Eliminating this group of triggers, would have resulted in a 9% reduction in opiate allergy alerts at the study site. It also opens for discussion that proactive management of drug side effects may be needed.

In this study, 16% of the opiate alerts had no reaction documented, and 74.3% of the alerts had no severity assigned. The absence of reaction detail has been identified as a significant hindrance in provider evaluation of alerts. Data intake screens must be designed to make it easier to capture accurate data, rather than "guess" or ignore these data fields.

This study has several limitations. The data is from a university medical center with advanced capacity to build and continuously tune its CPOE system. The alert trajectory is complex and involves many systems and disciplines which may differ across institutions. Those systems include the allergy data collection tools, allergy processing algorithms, and CPOE alert displays. Limiting the study to opiates may limit its generalizability to other drugs.

Discussion

Drug allergy alerting was one of the earliest and supposedly simplest forms of CPOE CDS, yet still has unacceptably high override rates. While alert logic may be simple, the complex environment in which it exists is not. This study highlights unintended consequences that can occur when computerized clinical decision support logic is not tightly aligned with the electronic medical record data upon which it depends. Much of our recorded healthcare data is influenced by human factors such as memory, personal experience, and varied interpretation by patients, clinicians and support staff.

The volume of side effects, recorded as allergies, should not be ignored or completely attributed to vague allergy definitions, human error or the influences of data entry screen design. Moreover, this

reflects the need for better symptom management of opiates and all medications that are known to cause distressing side effects. Increased education and engagement of patients in understanding the use and effects of opiates, accurate reporting and partnering with their providers in comfort management is fundamental.

This research also highlights the need for continuous feedback and analysis of alerts in clinical practice, to identify opportunities to improve systems and reveal emerging healthcare patterns as more people use computers to record, deliver and improve care. The high use opiates over time may be introducing new hypersensitivities in susceptible individuals that have yet to be fully revealed. More than half of all inpatients had opiates ordered. Of these, 9% of the patients triggered over 25,000 opiate alerts based on how their allergy history was recorded and how often opiates were re-ordered for the visit. This count could be tripled when you consider that the same alert that may be presented to the pharmacist on dispensing, and the nurse on administering.

References

Bates, D. W., Teich, J., & Lee J (1999). The impact of computerized physician order entry on medication error prevention. *Journal of the American Medical Informatics Association, 6,* 313-321.

HIMSS (2010). The Basics - Frequently Asked Questions on Meaningful Use and the American Recovery & Reinvestment Act of 2009. http://www.himss.org [On-line]. Available: http://www.himss.org/content/files/BasicFactsAboutMeaningfulUseARRA.pdf

Hsieh, T. C., Kuperman, G. J., Jaggi, T., Hojnowski-Diaz, P., Fiskio, J., Williams, D. H. et al. (2004). Characteristics and Consequences of Drug Allergy Alert Overrides in a Computerized Physician Order Entry System. *Journal of the American Medical Informatics Association, 11,* 482-491.

Hunteman, L., Ward, L., Read, D., Jolly, M., & Heckman, M. (2009). Analysis of allergy alerts within a computerized prescriber-order-entry system. *American journal of health-system pharmacy, 66,* 373

Johansson I (2005). The revised allergy nomenclature A sharp tool that must not be blunted. *Allergy Clin Immunol Int - J World Allergy Org, 17,* 128-130

Kohn, L., Corrigan J, & Donaldson, M. (1999). *To Err Is Human: Building a Safer Health System*. Washington, D.C.: National Academy Press.

Kuperman, G., Gandhi TK, & Bates DW (2003). Effective drug-allergy checking: methodological and operational issues. *Journal of biomedical Informatics,* 70-79.

Sittig, D. F., Krall, M., Kaalaas-Sittig, J., & Ash, J. S. (2005). Emotional Aspects of Computer-based Provider Order Entry: A Qualitative Study. *Journal of the American Medical Informatics Association, 12,* 561-567.

van der Sijs, H., Aarts, J., van Gelder, T., Berg, M., & Vulto, A. (2008a). Turning of frequently overridden drug alerts: limited opportunities for doing it safely. Journal of the American Medical Informatics Association Epub ahead of print[April]. Ref Type: Abstract

van der Sijs, H. (2009). Drug Safety Alerting in Computerized Physician Order Entry Unraveling and Counteracting Alert Fatigue. PhD Erasmus University of Rotterdam.

van der Sijs, H., Aarts, J., Vulto, A., & Berg, M. (2006). Overriding of drug safety alerts in computerized physician order entry. *Journal of the American Medical Informatics Association, 13,* 138-147.

Application of Bayesian Logistic Regression to Mining Biomedical Data

Viji R. Avali PhD[1], Gregory F. Cooper MD, PhD[1,2,3], and Vanathi Gopalakrishnan PhD[1,2,3]
[1]Department of Biomedical Informatics, [2]Intelligent Systems Program, [3]Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA

Abstract

Mining high dimensional biomedical data with existing classifiers is challenging and the predictions are often inaccurate. We investigated the use of Bayesian Logistic Regression (B-LR) for mining such data to predict and classify various disease conditions. The analysis was done on twelve biomedical datasets with binary class variables and the performance of B-LR was compared to those from other popular classifiers on these datasets with 10-fold cross validation using the WEKA data mining toolkit. The statistical significance of the results was analyzed by paired two tailed t-tests and non-parametric Wilcoxon signed-rank tests. We observed overall that B-LR with non-informative Gaussian priors performed on par with other classifiers in terms of accuracy, balanced accuracy and AUC. These results suggest that it is worthwhile to explore the application of B-LR to predictive modeling tasks in bioinformatics using informative biological prior probabilities. With informative prior probabilities, we conjecture that the performance of B-LR will improve.

Introduction

Biomedical data tend to have many variables and a scarcity of samples. Mining such high dimensional data with existing classifiers is challenging and the predictions are often inaccurate. Logistic regression (LR) is often applied in making predictions. However, it is difficult to include prior biological knowledge into the analysis when using LR. Almost all biomedical domains have associated domain knowledge. It would be helpful to be able to include such additional knowledge when building predictive models. For example, if a predictor variable is already known as a biomarker for a disease, it will be prudent to use this information when trying to come up with a model for classification and prediction for that disease. Prior knowledge can be incorporated into Bayesian Logistic Regression (B-LR) and the method is computationally efficient. B-LR has been applied successfully in text categorization [1], in integrating early physiological responses to predict later illness severity in preterm infants[2], and in early prediction of the response of breast tumors to neoadjuvant chemotherapy [3]. But in both [2] and [3], the number of predictor variables is small. We want to study the performance of B-LR on classifying high dimensional data and compare its performance to other existing classifiers. This paper uses a B-LR implementation that is readily available in the WEKA data mining environment[4]. Our goal is to understand the extent to which B-LR performs on par with other classifiers in WEKA according to the following performance measures – accuracy, balanced accuracy (i.e., average of sensitivity and specificity), and the area under the ROC curve (AUC).

Background

Linear logistic regression is a probabilistic classification model used for predicting a target variable depending on one or more predictor variables. It can give accurate predictions, but it often does not handle high dimensional data well. One way to overcome the shortcomings of LR is to apply a Bayesian approach with a prior probability distribution over predictor variables. In Bayesian analyses, the three steps involved are (1) specifying prior probabilities for the parameters, (2) determining the marginal likelihood of the data, (3) and using Bayes theorem to determine the posterior distribution of the parameters. B-LR captures the nonlinear relationships between the predictor variables and the outcome variable using Bayesian modeling. In B-LR, the equation for calculating the posterior probability of a sample belonging to a specific class is generated by the traditional logistic function:

$$P(Class|a_1, a_2, a_3, \dots, a_n) = \frac{1}{(1 + \exp\,(b + w_0 * c + \sum_{i=1}^{n} w_i * f(a_i)))} \quad (1)$$

Where, $'a_i'$ denotes the predictor variables, 'c' is the prior log odds ratio $\left(c = log\ \frac{P(\text{class}=0)}{P(\text{class}=1)}\right)$, the bias '$b$' and weights w_0 and w_i are learned from the training data, and the i^{th} attribute a_i is used to calculate the feature $f(a_i)$, using $f(a_i) = \log\left(\frac{P(a_i \mid \text{class}=0)}{P(a_i \mid \text{class}=1)}\right)$ (for binary class outcome variables). In the Bayesian approach to logistic regression, a univariate Gaussian prior with a mean '0' and a variance of 'σ_i' over the weights is commonly used. By

using a mean of '0', we assert our prior belief that the weights are close to zero. The values of σ_i are positive, with small values indicating our confidence in the values of the weights and larger values indicating the lack there of. Though this Gaussian prior favors weights with values close to zero, it does not favor the values exactly being zero. Maximum a posteriori (MAP) estimate of these weight values is similar to ridge regression for the logistic model.

The B-LR implementation in WEKA is based on [1] and has Gaussian parameter priors and Laplace parameter priors as the two options. Domain knowledge related to the datasets can be incorporated by specifying a prior, thereby defining a distribution over the values of the weights. Since the WEKA implementation of B-LR has Gaussian and Laplace priors as the two options available, we used only these non-informative priors in our analysis.

Experimental Method

Twelve datasets with binary class variables were chosen. Eleven are publicly available and one is a private dataset collected in the LungSPORE project[5]. The LungSPORE dataset contains as yet unpublished data that was collected to validate the results of an earlier study [5]. This study identified a panel of ten serum biomarkers that distinguished lung cancer from controls and have the potential to aid in the early detection of lung cancer and more accurate interpretation of indeterminate pulmonary nodules detected by CT screening.

Table 1. Details of the 12 datasets that were analyzed.
G/P indicates if the data is Genomic or Proteomic. P/D shows whether the data is Prognostic (P) or Diagnostic (D). The number of variables (Original) gives the total variables in the original dataset. The number of variables (PAIFE) gives the total number of variables after processing the dataset through our irrelevant feature elimination algorithm 'PAIFE'. The Sample (Class1, Class2) gives the total number of samples and class distribution, and 'Reference', the relevant reference to the dataset.

ID	G/P	P/D	Number of variables (Original)	Number of variables (PAIFE)	Sample(Class1,Class2)	Outcome variable	Reference
1	G	D	6584	1972	61(40,21)	Colon Cancer	Alon et al [6]
2	G	P	5372	858	86(69,17)	Lung Cancer	Beer et al. [7]
3	P	D	70	15	205(66,139)	Lung Cancer	Bigbee, et al. [5]
4	G	D	7129	2288	72(47,25)	Leukemia	Golub, et al. [8]
5	G	D	7464	1880	36(18,18)	Breast Cancer	Hedenfalk et al. [9]
6	G	P	7129	699	60(20,40)	Hepatocellular carcinoma	Iizuka et al. [10]
7	G	P	7399	1084	240(138,102)	Lymphoma	Rosenwald, et al. [11]
8	G	D	7129	1927	77(58,19)	Lymphoma	Shipp, et al. [12]
9	G	P	24481	4251	78(44,34)	Breast cancer	Van't Veer, et al. [13]
10	G	D	7039	1230	39(35,4)	Ovarian Cancer	Welch, et al. [14]
11	G	P	12625	1166	249(201,48)	Leukemia	Yeoh, et al. [15]
12	P	D	16	12	583(184,401)	Lung cancer	LungSPORE (unpublished)

For high dimensional biomedical data, the presence of uninformative variables in the dataset introduces noise and adversely affects the performance of classifiers. As a preprocessing step, we used our in-house developed algorithm called 'Partitioning based adaptive irrelevant feature eliminator' (PAIFE) to remove presumptive non-informative features from the datasets[16]. PAIFE evaluates predictor variable – outcome variable relationships over not only a whole dataset, but also the partitioned subsets and is effective in identifying variables whose relevance to the outcome are conditional on certain other variables. In experiments with synthetic datasets, PAIFE had outperformed other state-of-the-art feature selection methods in retaining relevant features and eliminating irrelevant ones [16]. PAIFE successfully removed irrelevant features when tested on proteomic and genomic datasets and the models developed from the PAIFE processed datasets performed either better or on par with the models built without any processing.

The PAIFE processed datasets were then normalized using the unsupervised attribute filter 'normalize' in WEKA's preprocessing step. The following methods were applied with 10-fold cross validation on each of these PAIFE processed, normalized datasets: B-LR with both Gaussian and Laplace priors, B-LR with Gaussian priors and cross-validation-based hyperparameter selection [1], C4.5 (J48 in WEKA) [17], naïve Bayes [18], simple logistic regression [19, 20], ridge logistic regression [21], CART (SimpleCart in WEKA) [22], Random Forest [23], and SVM (SMO in WEKA[24]), as implemented in WEKA 3.6.10 [4]. Simple logistic regression (LR_{simple}) in WEKA is the linear logistic regression and ridge logistic regression (LR_{ridge}) is the logistic regression model with a ridge estimator. For all the classifiers, except B-LR with Gaussian priors and hyperparameter selection based on cross validation (CV), default parameter values in WEKA were used. That classifier was chosen by selecting the 'CV based hyperparameter' option in WEKA's B-LR classifier.

The statistical significance of the results was analyzed by paired two-tailed t-test and by non-parametric Wilcoxon paired-samples signed ranks test. We used the alpha value of 0.05 for significance testing.

Results

In our performance analysis, we compared the accuracy, balanced accuracy (BACC) and percentage AUC values of each classifier.

Table 2: Comparison of the accuracy (percentage) of the classifiers from 10-fold cross validation.

ID	B-LR_{GP1}	B-LR_{GP2}	B-LR_{LP}	J48	NB	LR_{simple}	LR_{ridge}	CART	SVM	RF
1	96.72	96.72	96.72	98.36	96.72	98.36	96.72	**100.0**	98.36	88.52
2	93.02	93.02	88.37	66.28	84.88	81.40	86.05	76.74	**93.02**	79.07
3	78.05	74.63	73.17	80.98	78.54	82.93	**83.90**	79.02	80.49	80.49
4	98.61	97.22	98.61	84.72	98.61	90.28	94.44	83.33	**98.61**	94.44
5	97.22	97.22	94.44	91.67	**100.0**	94.44	97.22	94.44	97.22	91.67
6	**91.67**	**91.67**	75.00	55.00	88.33	71.67	78.33	65.00	**91.67**	78.33
7	70.00	**71.67**	70.42	60.00	65.83	67.50	65.42	57.08	69.17	65.42
8	**97.40**	**97.40**	**97.40**	70.13	89.61	94.81	**97.40**	68.83	**97.40**	85.71
9	96.15	85.90	89.74	82.05	83.33	93.59	93.59	80.77	**98.72**	92.31
10	**100.0**	**100.0**	**100.0**	92.31	97.44	97.44	**100.0**	87.18	**100.0**	97.44
11	89.16	85.14	82.33	72.69	79.92	80.72	86.75	80.72	**89.96**	81.12
12	91.21	92.43	80.37	93.46	90.59	95.50	95.50	94.89	92.43	**96.11**
Avg	91.60	90.25	87.21	78.97	87.82	87.39	89.61	80.67	**92.25**	85.89
s.d.	8.61	8.82	10.19	13.49	9.61	9.92	9.65	12.16	8.70	8.98

Table 3: Comparison of BACC (percentage) values of the classifiers from10-fold cross validation.

ID	B-LR_{GP1}	B-LR_{GP2}	B-LR_{LP}	J48	NB	LR_{simple}	LR_{ridge}	CART	SVM	RF
1	96.37	96.37	96.37	97.73	96.37	98.78	95.65	**100.0**	97.73	90.21
2	92.73	92.73	**93.67**	48.04	76.42	70.34	85.90	39.76	92.73	40.00
3	82.65	83.38	**85.82**	78.32	76.89	82.08	83.38	76.45	82.70	77.79
4	**98.96**	96.94	**98.96**	82.95	**98.96**	89.05	93.33	81.48	**98.96**	96.08
5	97.37	97.37	95.00	92.86	**100.00**	95.00	97.37	94.44	97.37	92.86
6	**92.41**	**92.41**	74.12	52.29	90.00	67.78	75.64	55.39	91.08	75.67
7	69.28	**71.03**	69.92	59.25	65.50	66.80	64.71	54.68	68.46	64.60
8	96.51	96.51	96.51	62.34	83.78	93.01	**98.33**	58.08	96.51	84.74
9	96.26	87.47	89.88	81.75	84.64	94.13	93.38	81.18	**98.89**	93.13
10	**100.0**	**100.0**	**100.0**	78.53	98.61	90.00	**100.0**	44.74	**100.0**	98.61
11	87.88	**92.23**	78.79	58.79	68.20	67.15	82.02	40.36	86.99	73.98
12	92.93	93.25	87.56	92.94	90.37	95.61	95.50	94.78	94.17	**95.65**
Avg	91.95	91.64	88.88	73.82	85.81	84.15	88.77	68.45	**92.13**	81.94
s.d,	8.27	7.59	9.54	16.30	11.49	12.08	10.25	21.22	8.73	16.24

Table 4: AUC (percentage) values for the classifiers from 10-fold cross validation.

ID	B-LR_{GP1}	B-LR_{GP2}	B-LR_{LP}	J48	NB	LR_{simple}	LR_{ridge}	CART	SVM	RF
1	96.40	96.40	96.40	98.70	96.30	98.40	99.50	**100.00**	98.70	97.90
2	84.60	84.60	70.60	49.40	85.80	80.60	**88.50**	39.70	84.60	79.50
3	67.10	61.00	58.30	79.00	85.80	85.90	86.40	78.80	71.70	**88.10**
4	98.00	96.90	98.00	81.40	98.70	95.40	**99.30**	79.00	98.00	96.70
5	97.20	97.20	94.40	91.70	**100.00**	93.20	98.60	94.40	97.20	91.70
6	88.80	88.80	66.30	52.20	86.70	70.50	84.70	53.60	**90.00**	87.80
7	68.80	70.50	68.60	62.40	72.30	**74.70**	71.70	54.40	68.50	69.20
8	96.50	96.50	96.50	62.20	88.60	**99.50**	97.50	61.40	96.50	93.30
9	95.90	84.50	89.20	84.40	86.40	97.20	**99.50**	75.50	98.50	95.40
10	**100.00**	**100.00**	**100.00**	84.60	87.50	99.30	**100.00**	44.60	**100.00**	**100.00**
11	75.00	61.50	55.80	61.50	73.10	69.20	**81.60**	48.10	78.70	74.40
12	88.70	90.70	73.90	92.20	96.70	**98.90**	98.80	93.70	90.10	98.60
Avg	88.08	85.72	80.67	74.98	88.16	88.57	**92.18**	68.60	89.38	89.38
s.d.	11.26	13.41	15.95	15.95	8.59	11.35	8.98	20.11	10.63	9.64

In tables 2, 3, 4, 5, and 6, we use B-LR_{GP1} to indicate B-LR with Gaussian priors, B-LR_{GP2} to indicate B-LR with Gaussian priors with cross-validation based hyperparameter selection, B-LR_{LP} for B-LR with Laplace priors, NB for Naïve Bayes, RF for Random Forest, 'Avg' for 'Average', 's.d.' for 'standard deviation'.

Table 2 shows the accuracy of all the classifiers on the different datasets. The bold value on each row indicates the classifier with the highest accuracy for that dataset.

Table 3 gives the balanced accuracy for the different classifiers with the bold numbers indicating the classifier with the maximum BACC value for a specific dataset. Table 4 shows the percentage of AUC values for each of the classifiers for all the datasets.

We evaluated the statistical significance of these performance measures using the paired two-tailed t-test and the non-parametric Wilcoxon signed ranks test. Tables 5 and 6 show the results. The captions of the tables explain the contents of their cells.

Table 5: Comparison of the classifiers using a paired two-tailed t-test.
The numbers shown are p-values; the values below 0.05 are shown in bold. The value in a parenthesis is the mean performance of B-LR minus the mean performance of the listed classifier, expressed as a percentage. The values underlined are those in which the p-value is less than 0.05 and B-LR performed better.

B-LR with Gaussian priors versus each of the following:	Accuracy	BACC	AUC
B-LR_{GP2}	0.18 (1.35)	0.73(0.31)	0.14(2.37)
B-LR_{LP}	**0.01**(4.39)	0.10(3.06)	**0.01**(7.42)
J48	**0.01**(12.63)	**0.00**(18.13)	**0.02**(13.11)
NB	**0.02**(3.78)	**0.01**(6.13)	0.98(-0.08)
LR_{simple}	0.06(4.22)	**0.02**(7.8)	0.86(-0.48)
LR_{ridge}	0.20(1.99)	0.06(3.18)	**0.04**(-4.09)
CART	**0.00**(10.93)	**0.00**(23.5)	**0.01**(19.48)
SVM	0.06(-0.65)	0.57(-0.19)	**0.02**(-1.29)
RF	**0.01**(5.71)	**0.04**(10.00)	0.55(-1.3)

Table 6: Comparison of the classifiers using the Wilcoxon paired-samples signed ranks test. The numbers shown are p-values; the values below 0.05 are shown in bold. The value in a parenthesis is the mean performance of B-LR_{GP1} minus the mean performance of the listed classifier, expressed as a percentage. The values underlined are those in which the p-value is less than 0.05 and B-LR performed better.

B-LR_{GP1} versus each of the following:	Accuracy	BACC	AUC
B-LR_{GP2}	0.22 (1.35)	1(0.31)	0.31(2.37)
B-LR_{LP}	**0.02**(4.39)	0.15(3.06)	**0.01**(7.42)
J48	**0.01**(12.63)	**0.00(18.13)**	**0.02**(13.11)
NB	**0.02**(3.78)	**0.01(6.13)**	1.00(-0.08)
LR_{simple}	0.07(4.22)	**0.02(7.8)**	0.89(-0.48)
LR_{ridge}	0.31(1.99)	0.06(3.18)	0.02(-4.09)
CART	**0.01**(10.93)	**0.00**(23.5)	**0.02**(19.48)
SVM	0.09(-0.65)	0.58(-0.19)	**0.03(-1.29)**
RF	**0.01**(5.71)	**0.00**(10.00)	0.70(-1.3)

Discussion

From table 2 of the results, we can see that B-LR with Gaussian prior had the highest accuracy for three of the datasets and an average accuracy of 91.60%. Though SVM has outperformed B-LR with Gaussian prior with an average accuracy of 92.25%, the difference between the two values is very small (0.65%) and the standard deviation is 8.61 and 8.70, for B-LR and SVM respectively. B-LR with Gaussian prior has the maximum BACC value for three of the datasets and an average of 91.95%. SVM has the highest average BACC value of 92.13% leading B-LR with Gaussian prior by 0.18% (Table 3).

It is interesting to observe from table 4 that LR (with ridge estimator) has the highest AUC value of 92.18% when its accuracy and BACC measure were about 2% and 3% behind those of B-LR with Gaussian prior. In comparing the performance using paired two-tailed t-test and the non-parametric Wilcoxon signed ranks test, on accuracy, B-LR with Gaussian priors (B-LR_{GP1}) performed statistically significantly better ($p \leq 0.02$) than B-LR with Laplace priors, J48, Naïve Bayes, CART, and Random Forests. Only SVM had a higher accuracy, which was higher by 0.65% ($p = 0.06$). On BACC, B-LR_{GP1} performed statistically significantly ($p \leq 0.04$) better than J48, Naïve Bayes, LR_{simple}, CART, and Random Forest. No method had a statistically significantly better performance than B-LR_{GP1}, according to the BACC measure. On AUC, B-LR_{GP1} performed statistically significantly better ($p \leq 0.04$) than B-LR with Laplace prior, J48, and CART. LR_{ridge} and SVM had higher AUCs than B-LR_{GP1}, with LR_{ridge} being higher by 4.09% ($p = 0.04$) and SVM being higher by 1.29% ($p = 0.02$).

B-LR_{GP1} was also among the fastest methods, with an average time of 0.13 sec to build the model compared to LR_{ridge}'s average time of 13.45 sec.

In this study, we limited ourselves by using a single type of feature selection method, PAIFE. It would be important to learn the impact of the choice of feature selection method. In the future, we will examine other state-of-the-art feature selection methods and compare its performance to PAIFE. We would also like to observe the results without feature selection. Genkin et al. [1], observed that lasso logistic regression was effective on high dimensional data

analysis problems, it would therefore be interesting to observe the performance of lasso logistic regression on such higher-dimensional and noisier data.

Conclusion

The results from this study provide support that B-LR with a Gaussian prior performs well compared to a set of classifiers that include those often applied in bioinformatics. It provides researchers with an additional classifier from which they can choose when analyzing high dimensional data.

With these promising preliminary results, the next step will be to use biological domain knowledge to develop informative priors to use in B-LR, and then repeat the evaluation of its predictive performance. In this study, we analyzed only 12 datasets with binary class variables. We plan to extend our analysis to more datasets and datasets with multinomial class variables. Our future analysis will also evaluate the performance of B-LR when changing the parameter options in WEKA's B-LR classifier. For example, we used the default value of 100 for the number of iterations. We plan to have our own implementation of B-LR with options to choose the type of informative priors depending on the application domain.

Acknowledgements

The authors gratefully acknowledge the following grants from the National Library of Medicine at the National Institutes of Health: R01-LM010950 and 5T15 LM007059-26. VG was funded in part by grants R01GM100387 and P50CA090440 from the National Institutes of Health. GFC was funded in part by NIH grant R01LM010020 and NSF grant IIS0911032. The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the National Science Foundation.

This project used the UPCI Cancer Biomarkers Facility that is supported in part by award P30CA047904. We thank Dr. William L. Bigbee and his laboratory for the recently produced dataset from the Lung Cancer SPORE project (supported by NCI grant number: P50CA090440) that was also analyzed in this paper. The authors thank Jeya B. Balasubramanian, MS for processing the datasets through PAIFE and for performing the Wilcoxon test analysis.

References

1. Genkin, A., Lewis, D., and Madigan, D., *Large-Scale Bayesian Logistic Regression for Text Categorization.* Technometrics, 2007: p. 291-304.
2. Saria S, R., AK, Gould J, Koller D, Penn AA., *Integration of early physiological responses predicts later illness severity in preterm infants.* Science Translational Medicine, 2010. **2**(48): p. 48-65.
3. Mani S, C.Y., Arlinghaus LR, Li X, Chakravarthy AB, Bhave SR, Welch EB, Levy MA, Yankeelov TE. *Early Prediction of the Response of Breast Tumors to Neoadjuvant Chemotherapy using Quantitative MRI and Machine Learning*. in *AMIA Anuual symposiam*. 2011.
4. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., *The WEKA Data Mining Software: An Update.* SIGKDD Explorations, 2009. **11**(1).
5. Bigbee, W.L., Gopalakrishnan, V., Weissfeld, J. L., Wilson, D. O., Dacic, S., Lokshin, A. E., & Siegfried, and J. M., *A multiplexed serum biomarker immunoassay panel discriminates clinical lung cancer patients from high-risk individuals found to be cancer-free by CT screening.* Journal of Thoracic Oncology, 2012(Apr;7(4)): p. 698-708.
6. Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., & Levine, A. J., *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays.* Proceedings of the National Academy of Sciences of the United States of America, 1999(96(12)): p. 6745-6750.
7. Beer, D.G., Kardia, S. L. R., Huang, C.-C., Giordano, T. J., Levin, A. M., Misek, D. E., … Hanash, S, *Gene-expression profiles predict survival of patients with lung adenocarcinoma.* Nature Medicine, 2002(8): p. 816–824.

8. Golub, T.R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., … Lander, E. S., *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.* Science,, 1999(286): p. 531–537.
9. Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., … Sauter, G. , *Geneexpression profiles in hereditary breast cancer.* The New England Journal of Medicine, 2001(344): p. 1-6.
10. Iizuka, N., Oka, M., Yamada-Okabe, H., Nishida, M., Maeda, Y., Mori, N., … Hamamoto, Y., *Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection.* Lancet, 2003(361): p. 923–929.
11. Rosenwald, A., et al, *The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma.* New England Journal of Medicine, 2002(346(25)): p. 1937-47.
12. Shipp, M.A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C. T., … Golub, T. R, *Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning.* Nature Medicine, 2002(8): p. 68-74.
13. Veer, L.v.t., Dai, H., & Vijver, M. Van De., *Gene expression profiling predicts clinical outcome of breast cancer.* Nature, 2002.
14. Welsh, J.B., Zarrinkar, P. P., Sapinoso, L. M., Kern, S. G., Behling, C. A., Monk, B. J., ... & Hampton, G. and J. M., *Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer.* Proceedings of the National Academy of Sciences of the United States of America, 2001(98(3)): p. 1176-1181.
15. Yeoh, E., et al., *Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling.* Cancer Cell, 2002(1(2)): p. 133-43.
16. Liu, G., Kong, L., & Gopalakrishnan, V. *A Partitioning Based Adaptive Method for Robust Removal of Irrelevant Features from High-dimensional Biomedical Datasets*. in *AMIA Joint Summits on Translational Science*. 2012.
17. Quinlan, R., *C4.5: Programs for Machine Learning*. 1993: Morgan Kaufmann Publishers Inc.
18. John, G.H., Langley P. *Estimating Continuous Distributions in Bayesian Classifiers. In: Eleventh Conference on Uncertainty in Artificial Intelligence*. in *Conference on Uncertainty in Artificial Intelligence*. 1995. Morgan Kaufmann.
19. Landwehr, N., Hall, M., and Frank, E., *Logistic Model Trees.*, in *Machine Learning*. 2005, Springer-Verlag. p. 241-252.
20. Sumner, M., Frank, E. , and Hall, M.A. *Speeding up logistic model tree induction*. in *9th European Conference on Principles and Practice of Knowledge Discovery in Databases*. 2005. Springer.
21. le Cessie, S., van Houwelingen, J.C. , *Ridge Estimators in Logistic Regression.* Applied Statistics, 1992: p. 191-201.
22. Breiman, L., et al., *Classification and regression trees*. 1984: CRC press.
23. Breiman, L., *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.
24. Platt, J., *Fast Training of Support Vector Machines using Sequential Minimal Optimization.* Advances in Kernel Methods - Support Vector Learning, 1998.

Cognitive design of a digital desk for the emergency room setting

Magnus Bang, PhD[1], Erik Prytz, PhD[2], Jonas Rybing, MSc[1], Toomas Timpka, MD PhD[1]
[1]Department of Computer and Information Science, Linköping University, Sweden
[2] SICS East Swedish ICT AB, Linköping, Sweden

Abstract

Digital desk technology has a still mainly unexplored potential to support the everyday work of collaborating clinicians. This paper presents ER Desk – a digital desk that was designed to specifically support a team of healthcare professionals working in an emergency room setting. The underlying design requirements were elicited in a comprehensive distributed cognition study of paper-based practices in an emergency room of a middle-sized Swedish hospital. We present the user interface and visualization requirements for digital desks for small clinical emergency room teams. Moreover, we discuss key design issues more generally with a focus on supporting team awareness, cognition, and collaborative routines of healthcare personnel working in clinical environments such as emergency rooms and intensive care units.

Introduction

Healthcare settings are highly collaborative work environments where the combined efforts of different professional groups are required to bring efficient and safe care. However, this basic requirement is not necessarily recognized in the design of the modern electronic patient-record systems (EMRs). It is currently common to employ desktop computers (EMR workstations) and replace paper forms with digital media when attempting to make the clinical work practices more rational and effective. Researchers have criticized this approach to computerization as it may impose unnatural and inefficient ways of working on the clinicians. This may, in fact add to the workload of the healthcare professionals rather than reduce it.[1,2,3] EMR systems have been deployed at all hospitals in Sweden. As a result, the clinical documentation processes have changed from a flexible and nomadic paper-based practice to a stationary and individualized process. One problem when substituting paper forms for digital media is that it deprives clinicians from informal information sharing possibilities that could facilitate the workflow and aid them in their situational awareness. Another disadvantage of the workstations is that they use small screens, which make them unsuitable for clinical real-time, face-to-face collaborations. Overall, the screens and interaction devices of such systems are clearly designed to be used by one person at a time.

An overview of the patients under treatment, the triage order, and the means to flexibly divide labor and share information are important qualities for healthcare teams at hospitals worldwide. However, adequate user interfaces for co-located care coordination and planning actions are still missing. Even in computerized healthcare settings is important team-oriented information still scribbled on large whiteboards or represented by stacking paper forms in special ways to visualize important care actions and prioritizations. We believe that using carefully designed interactive surfaces, such as digital desks with multi-touch functionality, is a promising approach to address the above-discussed issues. Such digital desks have primarily been developed in research environments[4] but are now starting to emerge on the commercial market.[5] However, there has been little work presented on the topic by researchers working in healthcare-related domains. Most efforts have been on medical imaging, where both augmented environment approaches[6] and multi-touch solutions have been presented[7].

The aim of this study was to examine how workflow processes uncovered in ethnographic workplace studies can be transformed into representations and systems functionality in a digital healthcare context. We present the ER Desk – a digital desk that was designed specifically to support a small team of clinicians working tightly in an emergency room setting. The foundation for the design was established during a comprehensive workplace study at an emergency room in a middle-sized hospital in Sweden during two periods in 2003 and 2007 respectively. The studies were made from the perspective of distributed cognition[8] – a viewpoint that acknowledges that cognitive processes such as memory, knowledge, and attention are distributed across people and artifacts rather than contained within a single individual. The paper is arranged as follows: first, we provide a brief introduction to the perspective of distributed cognition. Second, the paper offers an in-depth analysis how clinicians worked in the emergency room before a modern EMR system was installed with an analysis of their paper-based collaborative practices from the

perspective of distributed cognition. Third, we present the digital desk and its design components in relation to the analysis. Lastly, we discuss issues that are of importance for interaction designers that wish to develop interfaces for collaborative clinicians. In particular, we discuss matters related to cognition and collaboration in these settings.

Background

Distributed cognition and external tools

Workplace researchers have studied how professionals manage and share large amounts of information and knowledge in the workplace.[9-11] This research emphasizes the role of immediate here-and-now experiences and flexible user interaction techniques. For example, distributing paper forms on a desk to create a large display to facilitate a general overview of the patients under treatment is a common practice even in computerized wards. Having natural meeting places where patients can be discussed and actions coordinated improves learning, creates a general awareness of work-to-do, and simplifies the social distribution of work among the team members.[12,13] Distributed cognition emphasizes the social aspects of cognition and, particularly, how people arrange their physical environments to facilitate communication and to offload cognitively demanding tasks. According to Norman, so-called *cognitive artifacts*[14] – such as physical arrangements and visualizations – are particularly supportive since they aid memory and direct attention to the important tasks. This distributed cognition perspective can be used to generate design requirements to guide digital desk technology for healthcare teams.

Workplace study: Requirements on the digital desks

Our previous studies of teamwork in the emergency room setting have resulted in a set of high-level and low-level requirements for digital desks. Moreover, we have developed basic functional prototypes of digital desks for healthcare teams[16, 17]. The following is a brief synopsis of the clinical collaborative routines at a tabletop when the clinic was entirely paper-based.

Figure 1. The surgical team desk when the emergency department was paper-based in 2003.

Situational overview

The main team coordination artefact was an ordinary desk, centrally placed in the ER, with stacked patient folders (see Figure 1). It was used to visualize the *current state of the clinic and the workflow*. The arrangement of patient folders on this desk portrayed an *overview* of all cases being treated by members of the surgical team. The large display of folders (about 100 by 100 cm) facilitated an *on-the-fly visual evaluation* of the amount of work to be done. From a cognitive perspective, representations such as this arrangement function as an external memory and allow clinicians the use of less demanding, low-level perceptual routines (i.e., seeing the folders on the table) rather than high-level cognitive processes (i.e., keeping a list of cases in working memory) to assess a situation.

Collective clinical urgency ranking

The patient folders were arranged in two rows on the desk to signify the triage order and the work process (again, see Figure 1). Patients that had not been seen by the physician were placed in a row on the right side of the desk. This row portrayed the triage; that is, the prioritization of patients made by the head nurse. On the left side were folders of patients that were at the radiology department or awaiting test results. Much effort was devoted to *keeping the arrangement of folders and forms on the desk updated*. For example, clinicians moved the folders around on the desk to indicate patients' rank and position in the workflow, and they placed special plastic tags on the folders to designate where patients were treated (e.g., at the radiology department). This constant *cognitive activity* created a natural and *shared awareness* of patients with regard to urgency and an ordering of the activities of the team.

Signalling turn-taking

Signalling a clear handover of a task was important. Our study found that the specific placement of a folder on a desk could, for example, signify that the physicians asked the nurse to take a biomedical test. This simple and visible strategy for signalling and turn-taking was highly effective, because it reduced the need for verbal communication, and it also functioned as a persistent external memory of work-to-do.

Cognitive focusing by document order

In addition to the shared affordances, nurses arranged the paper forms within individual patient folders to offload cognitive tasks for the physicians. For example, newly-arrived laboratory results were placed in a folder in such a way that they were easily seen by the doctors, whereas unimportant information and forms were placed further back in the pile. These routines reduced the need for physicians to browse through the documents in search of necessary facts. From the perspective of distributed cognition, this collaborative technique can be seen as distributed cognitive processing and collaborative filtering of information[9] – an approach that changes the nature of subsequent tasks.

Digital desk technology

Wellner[4] proposed the digital desk concept in the early nineties. Basically, digital desks have a large area that can show computer-generated imagery, and they also permit tactile interaction, for example to annotate objects and move them around spatially to create order.[18, 19, 20] Some desk implementations recognize objects that are placed on the tabletop[21] and may also provide tactile feedback when an object is moved.[22] Other, more advanced desk-based interaction techniques are digital clay and sand that provide feedback to the user in real-time[23]. Several desk designs have been presented for various purposes such as education in the school setting,[24] landscape architecture,[23] and music creation[25]. Particularly interesting for this project are advances in multi-touch functionality that allow several individuals to work on the same digital desk simultaneously.[30-32] The devices and the software tools for developing multi-touch collaborative surfaces for digital desks are currently emerging on the market as off-the-shelf products. However, even though research in related domains such as pervasive healthcare and clinical visualization is abundant,[26-29] research on using digital desk technology in healthcare settings is scarce.

Method

Ethnographic research methods were used for a month-long period of qualitative data collection at an emergency room in a medium-sized Swedish hospital[16] from the perspective of distributed cognition.[8] The collected data were analyzed to create a requirements specification for the design of a digital desk graphical user interface (GUI).

Apparatus

A digital desk prototype was constructed using a 52 inch HDMI full-HD plasma screen with a 1920x1080 resolution measuring about 25x45 cm. The multi-touch sensor was a 52 inch IR-touch frame from IRTOUCH Systems. This frame was placed on top of a plate of glass and then placed on the plasma screen. A standard PC with a 3.1 GHZ processor and 4 GB RAM ran the tracking and visualisation software that was written in Delphi (tracking stokes) and Adobe Flex (visualisation). The PC had a wireless physical keyboard that became the main text input device.

Results

The resulting GUI was named ER Desk for ease of reference. The overall design was implemented on a large, multi-touch digital desk and was focused on supporting collaborating teams working in an emergency room setting. The design can be conceptualized as a shared workspace for physicians, nurses and nurse's aids. The overall goal was to allow the patient records – particularly the temporary clinical information and documents about the patients under treatment in the ward – to be visualised and accessed directly as virtual folders on a digital work surface to provide functional mapping from the paper-based system to the digital.

Overall design

The design principle was to exploit the supportive and flexible components of the practices previously recorded in the workplace studies. The interaction design basically followed the structure of how paper files were employed in the emergency room. Simple drag-and-move actions are used to create order and organize the work area. The urgency ranking approach and layout was an important feature of the paper-based clinic that supported overview of tasks and aided the workflow. To retain parts of this work method, we implemented a *patient strip* based on the concept of flight strips used by air traffic controllers[33]. Each patient is represented by one patient strip (see Figure 2). The strip can be configured to show key patient information, such as the patient identification number, the name of the patient, allergies, and sex of the patient. The information displayed on the strip can, unlike the paper-based strips used in air traffic control, be updated with new information continuously. For example, when the patient has been assigned to the clinic and a nurse has checked the patient, the strip can be updated to contain additional data such as the reason for the visit, blood pressure, heart rate, temperature, allergies, and medications. The central idea behind using strips was to allow clinicians to fairly freely create *their own* spatial representation of the work situation and the workflow using simplified representations of the electronic health care records. Hence, the patient strips can be stacked and arranged on the desk according to the clinical situation. For example the head nurse may convey triage order to the team, highlight prioritized patients, and demonstrate turn-taking of group tasks to facilitate the workflow. As an example, Figure 3 shows patient strips arranged in two rows to designate the workflow analogously to the paper stacks seen in Figure 1. When a patient strip is moved to the central area of the desk it opens to show the basic and shared nursing and physician documentation needed to treat this patient. The patient strip may contain physician orders, referrals, laboratory results, prescriptions, and similar information. These documents are in the current design accessed using tabs within the opened main view.

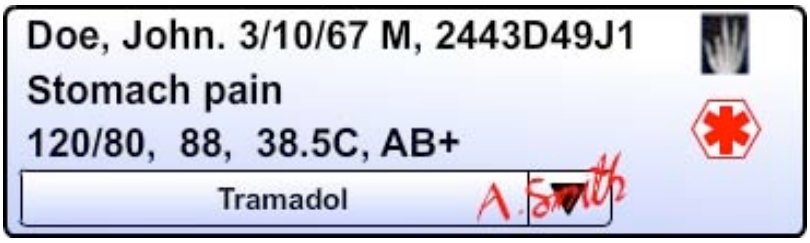


Figure 2. The Patient Strip. The strip holds a canonical set of medical information needed for the clinician to evalutate a situation, look for more information and take appropriate action. Moreover, the strip also says where the patient is currently situated within the hospital (the x-ray icon).

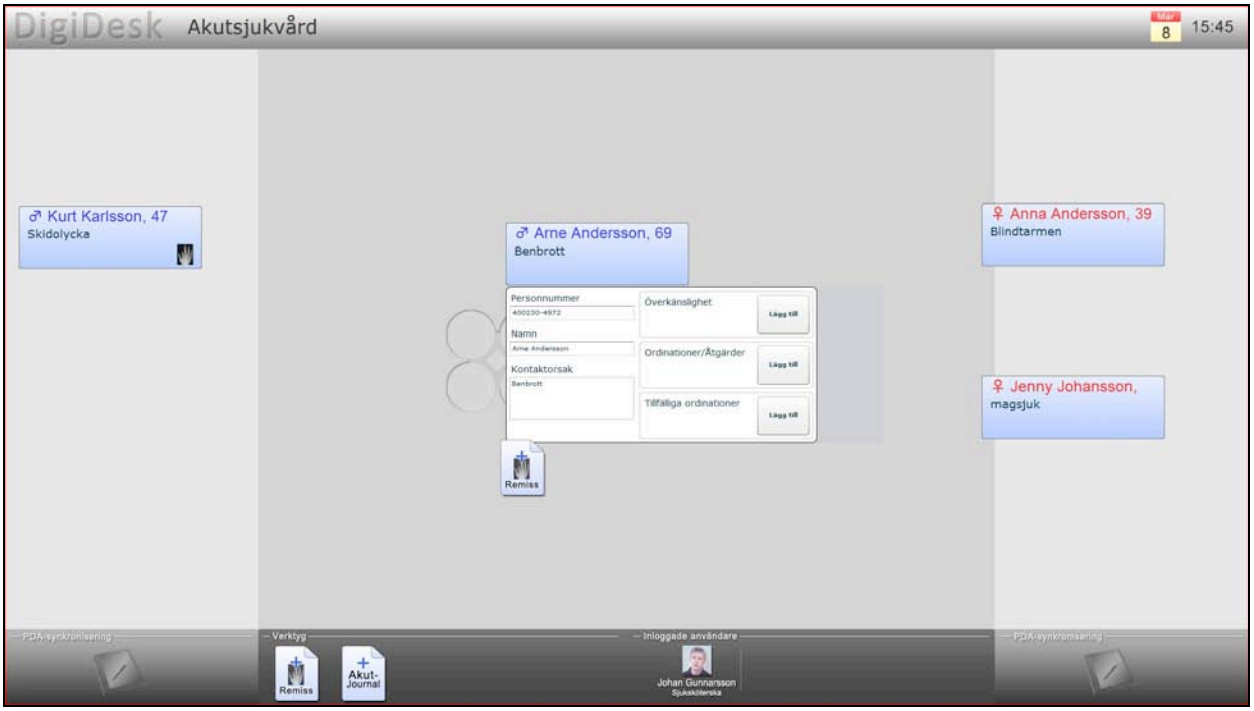


Figure 3. The ER Desk with patient strips. The icons in the dock are used to create new documents, such as a referral by dropping the icon on the activated patient strip. Moreover, the dock shows the active clinicians that are near the desk. Signing a referral or a biomedical test that has been taken is done by dragging and dropping the the personal icon on top of a data set or a document.

Situational overview and collective clinical urgency ranking

By utilizing a large (52") HD screen, a good situational overview of the patients can be obtained by both nurses and physicians based on the *spatial arrangement* of the patient strips. The two columns of digital patient strips at each side of the desk can be seen at a glance by the clinician to evaluate the amount of work to do for the team. For example, the patient strips can be placed in a sequence on the desk to convey the triage order. Each minimized patient strip also shows the most important information in this situation such as the basic biomedical measurements and important warnings (i.e., drug allergies).

Much work in healthcare is data driven, that is, nurses and doctors act on newly arrived information. The clinical workflow can thus be disrupted if new data does not arrive in a timely manner or is not intelligibly visualised to the responsible personnel. The patient strips should therefore be continuously updated and signal when new data, such as a lab result, has been received. In the current design, appearance of new data is highlighted on the patient trip as a flashing icon. In the current design, appearance of new data is highlighted on the patient trip as a flashing icon. Updates can also be tied to specific icons, such as using a flashing radiology icon for new X-ray images and the related diagnoses from the radiologist. A workspace is located at the center of the desk. Strips can be moved from the sides to the center, where they are opened, showing more information. This design maintains the sides for overview but allows the user to seek and display more detailed information for each patient. If one patient must be discussed by the clinical team, the desk can display this patient folder solely along with its main content similarly to opening and displaying the pages of a paper journal. This functionality allows clinicians to create meaning and discuss patient-related issues standing face-to-face or side-by-side.

Turn-taking and user defined cues

The prior workplace study revealed that the processes by which the clinical team hands over tasks and information between team members are intricate. This was often signalled by the spatial placement of the patient folder on the desk. For example, a physician could place a patient folder at a specific place on the desk to signify to the nurse that this patient needs an exam or that the patient has to be sent to the radiology department. This functionality was

mapped to the digital context by *allowing the operators to define and create their own cues* and customize the user interface to their own work process. The approach used in the present design to retain sufficient flexibility to allow such intricate handover processes is to use an open design. This means that the digital patient strips can be placed wherever on the desk, be adjusted or skewed slightly to the right or left to highlight a strip by letting it stick out from the rest. The meaning of these functions are not imposed by the designer but rather left to the clinical team to allow processes and markers to be defined socially in the team at "run-time".

Cognitive focusing by document order

In the paper-based system, each patient folder contained physician orders, referrals, laboratory results, and prescriptions. In the workplace study, we saw that these document were arranged within the folder by nurses so that a physician could find and process the most relevant and pressing information without effort. Thus, in a digital context, collaborative ordering of information within an electronic healthcare record must be allowed. This can be achieved by using tabs within an open electronic patient strip, such that each tab corresponds to individual documents. The team could rearrange these tabs so that they can create order and highlight the most relevant information in each situation.

Discussion

This study set out to examine how workflow processes uncovered in ethnographic workplace studies can be transformed into representations and systems functionality in a digital healthcare context. A large-surface digital desk was selected as the technical medium rather than a traditional computer workstation to facilitate overview, shared situational awareness, and to retain the workflow from the paper-based context. In particular, the current work outlines the design used to capture the situational overview, collective clinical urgency ranking, turn-taking, and cognitive focusing processes defined by the workplace study. Rather than using the PC desktop metaphor as a way to visualize workflow (e.g., Windows or MacOS), the ER Desk design relied on the workflow (i.e., cognitive functions) as revealed by the previous ethnographic study of one specific ER. This metaphor was augmented using the concept of patient strips, inspired by air traffic control. The design process highlighted two major design problems when developing a multi-touch desk for a healthcare team; how one should appropriately represent *a set of patients* on the one hand and *the individual patient* on the other. The general principle outlined by Shneiderman and colleagues[34] provides some guidance for this problem. They suggest that it is important to have overview first, and then have so-called "zoom and filter functionality" to obtain details-on-demand.[34]

Visualising patient sets

The visualization of patient sets can be said to represent a *clinic-overview* where all patients under treatment are visualised. The overview was, in our case, provided by a set of patient strips. The advantage of using strips is that it readily maps onto the work plan in a sequential process. This view and layout is common at hospitals today. From a distributed cognition perspective it functions as an external memory and could – given a social contract on the meaning of the spatial layout – provide cues on what to do next.[15] Having an open design allows clinicians to create meaning from functions in their social context. For example, clinicians may choose to convey important informal messages just by shifting the alignment of strips slightly because this cue yet has no a priori ascribed meaning. However, an open design that relies on a social contract could – particularly when there are many clinicians working at the clinic – be difficult to use for clinicians that work part-time. Novice users of the system may find it difficult to remember parts of the turn-taking system and the meaning of the spatial ordering of strips.

There are other approaches to metaphor and representation in the emergency room case, and choosing one or another would affect the in situ cognition and collaboration. For example, a strictly spatial view could be used as the underlying interaction metaphor, such as showing a map over the clinic and its beds with patient records laid out as the actual patients. Such representations cannot easily convey triage order but readily maps the location of the patient with the clinic. Nevertheless, the choice of underlying metaphor for the GUI representation should be anchored in ethnographic studies of current practices, which also involves the actual users early in the design process.

Visualizing individual patient data

Visualizing individual patient data is in the present design implemented as a basic *diagnosis view* where the vital signs, referrals, and lab results are shown. There has been some research on the visualization of patient data,[27, 28]

though these representations seem to take a limited view on patient visualisation. In the case of a digital desk design the question is more complex. The design problem, from a distributed cognition standpoint, is to find a way to visualise "*the big picture*" at the same time as providing detailed information. The physician needs to have *cognitive cues* and the exact information to take decisions, and *low-level cues* to support where to look. Also, the creation of cues may be distributed across members of the clinical team, which may support search, cognition and attention implicitly. Hence, it is important that designers carefully acknowledge these intricate cognitive processes and provide (open) designs that support distributed cognition.

Conclusion and Future Work

This paper presented a digital desk design developed to support a team of collaborating clinicians working in an emergency room setting. Specifically, the interaction design focussed on supporting face-to-face collaborative actions as well as decreasing the cognitive load on the clinicians. Future work includes studies on how to visualise single patients and sets of biomedical data appropriately, as well as evaluation studies of the design concepts outlined in the current paper. Further research is also needed on how to translate locally discovered workflow metaphors in a digital domain from one clinic to another, and how open design concepts are used and developed through shared, social practices in the workplace.

References

1. Berg M. Accumulating and Coordinating: Occasions for Information Technology in Medical Work. Computer Supported Cooperative Work. 1999;8:373–401.
2. Sellen A, Harper R. The Myth of the Paperless Office. Cambridge: The MIT Press; 2001.
3. Heath C, Luff P. Technology in Action. New York: Cambridge University Press; 2000.
4. Wellner P. The DigitalDesk calculator: tangible manipulation on a desk top display. UIST 91. Proceedings of the 4th annual ACM symposium on User interface software and technology; 1991 Nov 11-13; South Carolina, USA. New York: ACM Press; 1991. P. 27-33.
5. Microsoft Corporation. Microsoft PixelSense [Internet]. 2014. Avaliable from: http://www.microsoft.com/en-us/pixelsense/default.aspx
6. Brainlab. Brainlab [Internet]. 2014. Avaliable from: https://www.brainlab.com
7. Koehring A, Foo JL, Miyano G, Lobe T, Winer E. Framework for Interactive Visualization of Digital Medical Images. Journal of Laparoendoscopic & Advanced Surgical Techniques. 2008 Sep;18(5): 697-706.
8. Hutchins E. Cognition in the Wild. Cambridge: MIT Press; 1995.
9. Malone TM. How Do People Organize Their Desks? Implications of the Design of Office Information Systems. ACM Transactions on Office Information Systems. 1983 Jan;1(1):99–112.
10. Suchman L. Plans and Situated Actions: The problem of human-machine communication. New York: Cambridge University Press; 1987.
11. Lave J, Chaiklin, S. (eds.). Understanding Practice: Perspectives on Activity and Context, Cambridge: University of Cambridge Press; 1993.
12. Dourish P, Bellotti, V. Awareness and Coordination in Shared Workspaces. Proceedings of the ACM Conference on Computer-Supported Cooperative Work CSCW'92 (Toronto, Ontario), 107-114. New York: ACM Press 1992.
13. Latour B. Visualization and cognition: thinking with eyes and hands. Knowledge and Society. Studies in the Sociology of Past and Present 1986;6:1–40.
14. Norman D. Cognitive Artifacts. In: J.M. Carroll (ed.), Designing Interaction: Psychology at the Human-Computer Interface. Cambridge University Press, Cambridge, MA, 1991, pp 17–38.
15. Kirsh D. The Context of work. Human Computer Interaction 2001;16:305–322.
16. Bang M, Timpka T. Cognitive Tools in Medical Teamwork: The Spatial Arrangement of Patient Records. Methods of Information in Medicine 2003;42:331-336.
17. Bang M, Timpka T. Ubiquitous computing to support co-located clinical teams: Using the semiotics of physical objects in system design. International Journal of Medical Informatics 2007;76:58-64.
18. Arai T, Machii K, Kuzunuki S, Shojima H. Interactive-DESK: A Computer-Augmented Desk which Responds to Operations on Real Objects. In: Katz IR, Mack RL, Marks L, Rosson MB, Nielsen J, editors. Companion of the Conference on Human Factors in Computing Systems 1995.
19. Scott SD, Grant KD, Mandryk RL. System Guidelines for Co-located, Collaborative Work on a Tabletop Display. In: Kutti K. et al. Proceedings of the Eighth European Conference on Computer Supported Cooperative Work. 2003 14-18 Helsinki, Finland. 2003 ACM Press.
20. Rekimoto J. SmartSkin: an infrastructure for freehand manipulation on interactive surfaces, Proceedings of the SIGCHI conference on Human factors in computing systems; April 20-25, 2002, Minneapolis (MI), USA. ACM Press 2002.
21. Ullmer B, Ishii H. Emerging Frameworks for Tangible User Interfaces. IBM Systems Journal 2000;39:915–931.

22. Noma H et al. The proactive desk: a new haptic display system for a digital desk using a 2-DOF linear induction motor. Presence: Teleoperators and Virtual Environments 2004;13(2): 146-163.
23. Ishii H. Bringing Clay and Sand into Digital Design — Continuous Tangible user Interfaces. BT Technology Journal 2004;22(4):287 – 299.
24. Koike H et al., Interactive textbook and interactive Venn diagram: natural and intuitive interfaces on augmented desk system. Proceedings of the SIGCHI conference on Human factors in computing systems CHI '00. ACM Press 2000.
25. Jordà, S, Geiger G, Alonso M, Kaltenbrunner. The reacTable: Exploring the Synergy between Live Music Performance and Tabletop Tangible Interfaces. Proceedings of the first international conference on "Tangible and Embedded Interaction" (TEI07). Baton Rouge, Louisiana.
26. Bardram J, Baldus H, Favela J. Pervasive Computing in Hospitals. In Pervasive Healthcare: Research and Applications of Pervasive Computing in Healthcare. CRC Press, 2006.
27. Powsner SM, Tufte ER. Graphical summary of patient status. The Lancet 1994;344:386-389.
28. Plaisant C, Milash B, Rose A, Widoff S, Shneiderman B. LifeLines: Visualizing Personal Histories. In: Proceedings of the ACM CHI 96 Human Factors in Computing Systems Conference (CHI 96); 1996 April 14-18; Vancouver, Canada. 1996 ACM Press.
29. Bardram J, Hansen T, Soegaard M. Large Interactive Displays in Hospitals - Motivation, Examples, and Challenges. In Proceedings of the CHI 2006 Workshop on `Information Visualization and Interaction Techniques for Collaboration Across Multiple Displays; 2006 April 22-27 Montreal, Canada. 2006 ACM Press.
30. Dietz P, Leigh D. DiamondTouch: A Multi-User Touch Technology. In: Proceedings of the 14th Annual ACM Symposium on User Interface Software and Technology (UIST 2001); 2001 November 11-14, Orlando, Florida. 2001 ACM Press.
31. Rekimoto, J. SmartSkin: An Infrastructure for Freehand Manipulation on Interactive Surfaces. Proceedings of the ACM CHI 2001 Human Factors in Computing Systems Conference (CHI 2001); 2001 Seattle, Washington. ACM Press 2001.
32. Shen. C, Vernier, F, Forlines C, Ringel M. DiamondSpin: An Extensible Toolkit for Around-the-Table Interaction. Conference on Human Factors in Computing Systems (CHI 2004); 2004 April 24-29; Vienna, Austria.
33. Mackay W. Is paper safer? The role of paper flight strips in air traffic control. ACM Transactions on Computer-Human Interaction. ACM Transactions on Computer-Human Interaction 1999;6: 311–40.
34. Shneiderman B. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In: Proceedings of the 1996 IEEE Symposium on Visual Languages; 1996 September 3 – 6; Boulder (CO). 1996 IEEE Computer Society.

Learning to Identify Treatment Relations in Clinical Text

Cosmin A. Bejan, PhD[1] and Joshua C. Denny, MD, MS[1,2]

[1]Department of Biomedical Informatics, Vanderbilt University, Nashville, TN;
[2]Department of Medicine, Vanderbilt University, Nashville, TN

Abstract

In clinical notes, physicians commonly describe reasons why certain treatments are given. However, this information is not typically available in a computable form. We describe a supervised learning system that is able to predict whether or not a treatment relation exists between any two medical concepts mentioned in clinical notes. To train our prediction model, we manually annotated 958 treatment relations in sentences selected from 6,864 discharge summaries. The features used to indicate the existence of a treatment relation between two medical concepts consisted of lexical and semantic information associated with the two concepts as well as information derived from the MEDication Indication (MEDI) resource and SemRep. The best F1-measure results of our supervised learning system (84.90) were significantly better than the F1-measure results achieved by SemRep (72.34).

Introduction

Discovering treatment relations from clinical text is a fundamental task in clinical information extraction and has various applications in medical research. For instance, the availability of treatment relations could enable a more comprehensive understanding of a patient's treatment course,[1,2] improve adverse reaction detection,[3] and assess healthcare quality.[4,5] However, such relations are not stored in a structured format in most electronic medical records (EMRs), but rather they are encoded in narrative patient reports. Therefore, natural language processing technologies need to be employed to facilitate the automatic extraction of treatment relations from clinical text.

A treatment relation is a relation holding between two medical concepts in which one of the concepts (e.g., a medication or procedure) is a treatment for the other concept (e.g., a disease). The treatment relation between the concepts highlighted in (1), for instance, is defined by a medication(*levofloxacin*) as a treatment administered for a disease(*pneumonia*). The relation in (2), holding between a procedure(*surgery*) and a disease(*lumbar stenosis*), represents another common category of treatment relation. On the other hand, the relations between the emphasized concepts in (3) and (4) do not constitute treatment relations. As observed in (3), in spite of a medication being prescribed for a specific medical problem, the treatment did not cure or improve the medical condition of the corresponding patient. Similarly in (4), although *lorazepam* can be indicated for *nausea*, the context in which these concepts are described invalidate the existence of a treatment relation. Due to this observation, the only treatment relations in (4) are between *morphine PCA* and *pain*, *IV lorazepam* and *anxiety*, and *IV reglan* and *nausea*. The remaining pair combinations of the concepts from this example constitute non-treatment relations.

(1) *She switched to* [***levofloxacin***] *for treatment of* [***pneumonia***] *as evidenced on CT.*

(2) *He had a previous* [***surgery***] *for* [***lumbar stenosis***].

(3) *She was treated with* [***morphine***]*, which did not control her* [***pain***].

(4) *She will be discharged with her morphine PCA for pain, scheduled* [***IV lorazepam***] *for anxiety, and scheduled IV reglan for* [***nausea***].

Our approach for identifying treatment relations in clinical text is based on (a) the idea of exploring the contextual information in which medical concepts are described and (b) the idea of using predefined medication-indication pairs. For this purpose, we used MEDication Indication (MEDI),[6] a large database of medication-indication pairs. Another resource we employed in the feature extraction phase of our machine learning framework is SemRep,[7] which uses linguistic rules to identify treatment relations in text. For example, a rule of the form *X for treatment of Y* can be applied to find the treatment relation in (1).

The goal of our study was not only to build a system for accurately extracting treatment relations from clinical notes, but also to use this system for expanding MEDI with new medication-indication pairs.

Related Work

SemRep is a publicly available biomedical information extraction tool which was developed at the U.S. National Library of Medicine (http://semrep.nlm.nih.gov) and has been used by a number of investigators. Given a text document, this system analyzes each sentence from the document and identifies multiple types of semantic relations between the concepts described in the sentence. Examples of relations that SemRep is able to extract are DIAGNOSES, CAUSES, LOCATION_OF, ISA, TREATS, PREVENTS, etc. The system relies on MetaMap[8] to extract the Unified Medical Language System (UMLS) concepts from text and on linguistic and semantic rules specific to each relation. Its output for each sentence consists of a list of all the UMLS concepts mentioned in the sentence followed by the semantic relations that exist between these concepts. SemRep has been used in a wide range of applications in biomedical informatics including automatic summarization and literature based discovery.[9,10] While SemRep was primarily designed for processing documents from the biomedical research literature, only a few studies involving this system have been performed on clinical documents. In one of these studies, drug-disorder co-occurrences were computed from a large collection of clinical notes to improve the SemRep performance on extracting treatment relations from Medline citations.[11] Another study focused on how the semantic relations extracted by SemRep from Medline abstracts can guide the process of labeling concept associations from clinical text.[12]

One of the first machine learning systems developed to extract treatment relations from clinical text is described in Roberts et al.[13] In this work, the evaluation was performed on a small set of 77 oncology narratives, which was manually annotated with 7 categories of semantic relations. The feature set comprises various lexical, syntactic, and semantic features designed to capture different aspects of the relation arguments. Using a classification framework based on support vector machines (SVMs), the system achieved an average F1-measure of 72 over the 7 relation categories. An SVM-based framework was also developed by Uzuner et al.[14] to identify treatment relations defined for a more specific scope. To represent treatment relations, the authors of this study utilized sematic categories of concepts and the assertion values associated with these concepts. Examples of relation categories consist of present disease-treatment, possible disease-treatment, and possible symptom-treatment. Using a rich set of features, the SVM-based relation classifier recognized 84% and 72% of the relations annotated in two different corpora. Furthermore, due to its importance, the task of treatment relation extraction was part of the 2010 Informatics for Integrating Biology and the Bedside (i2b2)/Veteran's Affairs (VA) challenge.[15] Examples of treatment relations devised for this competition include relations in which the corresponding treatment (a) has cured or improved a medical problem, (b) has worsened a medical problem, (c) has caused a medical problem, (d) has been administered for a medical problem, and (e) has not been administered because of a medical problem. The concept pairs that occurred in the same sentence and did not fit this criteria were not assigned a relationship. The majority of the systems solving the 2010 i2b2/VA task on relation extraction relied on supervised machine learning approaches.[16–19]

In our preliminary studies,[20] we have implemented a simple algorithm using MEDI and have shown that it is a reliable method on assessing the validity of treatment relations identified in clinical notes. Like SemRep, MEDI is also publicly available (http://knowledgemap.mc.vanderbilt.edu/research/). It was developed by aggregating RxNorm, Side Effect Resource (SIDER) 2,[21] MedlinePlus, and Wikipedia and was designed to capture both on-label and off-label (e.g., absent in the Food and Drug Administration's approved drug labels) uses of medications. While RxNorm and SIDER 2 store the medication and indication information in a structured format, MedlinePlus and Wikipedia encode this information in narrative text. Therefore, further processing of the documents from MedlinePlus and Wikipedia was performed including the use of KnowledgeMap Concept Indexer[22,23] and custom-developed section rules to identify the text expressions describing indications and to map them into the UMLS database. The current version of MEDI contains 3,112 medications and 63,343 medication-indication pairs.

Method

We implemented a supervised learning framework that is able to predict whether or not two concepts co-occurring within the same sentence are in a treatment relation. To capture the relationship between the two concepts, we extracted various features based on the lexical information surrounding the concepts, on the semantic properties associated with the two concepts, and on the information derived from both MEDI and SemRep.

MEDI-based treatment relation extraction

A simple algorithm for treatment relation extraction is based on the assumption that any two concepts that co-occur within the same sentence and match a medication-indication pair in MEDI are likely to be in a treatment relation.

Table 1 The set of features for predicting whether two UMLS concepts are in a treatment relation.

Feature	Description
f_1:*semrep*	Boolean feature that is true whether SemRep indicates a treatment relation between the two concepts.
f_2:*medi*	Boolean feature that is true whether there is a match between the two concepts and a medication-indication pair in MEDI.
f_3:*semrep or medi*	Boolean feature that is true whether *f1* is true or f_2 is true.
f_4:*unigram*	The lowercased word unigrams surrounding the two concepts.
f_5:*bigram*	The lowercased word bigrams surrounding the two concepts.
f_6:*expression*	The lowercased word expression describing each of the two concepts.
f_7:*cui*	The concept unique identifiers (CUIs) of the two concepts.
f_8:*sem type*	The semantic types associated with the two concepts.

Despite the fact that this assumption does not take into account the context in which the two concepts are mentioned, our review of clinical documents before this study revealed that it holds true for the majority of the cases.

To increase the coverage of MEDI, we expanded the initial set of medication-indication pairs by using ontology relationships from the RxNorm database. For instance, because the medications in MEDI are mapped to generic ingredients,[6] the resource contains pairs involving RxCUI#1000082 (a*lcaftadine*), but it does not include pairs with medications containing alcaftadine ingredients such as RxCUI#1000083 (a*lcaftadine 2.5 MG/ML*) or brand medication names of alcaftadine as, e.g., RxCUI#1000086 (*lastacaft*). The relations we used to perform this expansion are *has_ingredient* and *tradename_of* from MRREL.

Features for predicting treatment relations

To learn a prediction model that is able to differentiate between pairs of UMLS concepts in treatment relations and the ones not belonging in such relations, we extracted the set of features described in Table 1. In this table, each feature was designed to capture a specific property associated with a pair of concepts. For instance, the lowercased word bigrams surrounding the emphasized concepts in (1) are "*switched to*", "*for treatment*", "*treatment of*", and "*as evidenced*". Also in (1), the semantic types extracted by MetaMap for *levofloxacin* are 'antibiotic' and 'organic chemical', and the semantic type for *pneumonia* is 'disease or syndrome'. As observed, since a concept can be associated with multiple semantic types in the UMLS Metathesaurus, the *sem type* feature can have multiple values for each concept. Of note, before the feature extraction phase, we assumed that the medical concepts were already identified in text and mapped to concepts in the UMLS Metathesaurus using SemRep.

In addition to the features listed in Table 1, we investigated the contribution of several other features including more word n-gram features and the version of f_4, f_5, and f_6 without lowercasing their corresponding textual expressions. We also extracted the concept preferred names and the semantic group(s) in which the semantic type(s) of each concept belongs to. None of these features were able to improve the overall performance of our prediction system.

Evaluation

To train our supervised learning system, we first constructed a dataset annotated with treatment and non-treatment relations. Based on this dataset, we then evaluated the performance results of our system and compared them against the results achieved by SemRep.

Dataset

For creating the dataset with annotated treatment relations, we randomly selected 6,864 discharge summaries from the Vanderbilt Synthetic Derivative, a de-identified version of the electronic medical record. In the data processing phase, we first split the content of each report into sentences using the OpenNLP sentence detector (http://opennlp.apache.org/) and removed the duplicate sentences. The output generated by this process consisted of 290,911 sentences. We then parsed these sentences with the current version of SemRep, v1.5, which identified 943,306 UMLS concepts (~3.2 concepts/sentence), and 3,386 treatment relations in 2,841 sentences. Next, we ran the MEDI algorithm over the same concepts extracted by SemRep in the previous step. Since SemRep is designed to identify treatment relations at sentence level, we constrained the algorithm based on MEDI to match any pair of concepts mentioned within the same sentence. As a result, 3,716 MEDI relations were obtained.

Table 2 Results for extracting treatment relations from text.

System	TP	FP	FN	TN	P	R	F
SemRep	625	145	333	9483	81.17	65.24	72.34
Our system	790	113	168	9515	87.49*	82.46*	84.90*

*p<0.001; statistically significant differences in performance between our system and SemRep.
F, F1-measure; FN, false negatives; FP, false positives; P, precision; R, recall, TN, true negatives, TP, true positives.

In the manual annotation phase, two reviewers examined 620 sentences in which the MEDI algorithm and SemRep identified at least one relation. We decided on this set of sentences to cover as many SemRep and MEDI relations as possible and, at the same time, to minimize the annotation effort. However, since all these sentences contain relations identified by both SemRep and the MEDI algorithm, one limitation of this selection is that the evaluation of the two systems may result in overestimating their corresponding recall values. Blinded from the results extracted by SemRep and the MEDI algorithm, the annotation process consisted of manually linking pairs of concepts that represent treatment relations inside every sentence. Once a sentence was annotated, the remaining combinations of concept pairs in the sentence were automatically marked as non-treatment relations. During this process, the reviewers performed a double annotation on 25% of the data obtaining a percentage agreement of 97.9, with the Cohen's Kappa value of 0.86. In the manual annotated dataset, the disagreements were adjudicated by an experienced clinical expert. For annotation, the BRAT annotation tool[24] was employed resulting in 958 and 9,628 treatment and non-treatment relations, respectively.

Results

We evaluated our machine learning framework on the manual annotated dataset using a 5-fold cross validation scheme. For performing the classification of treatment and non-treatment relations, we employed LIBSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm/), an implementation of the SVM algorithm. In all experiments, we used the radial basis function as the kernel function of choice. Since our interest was in detecting the treatment relations as accurately as possible, we measured the performance results of our system in terms of precision, recall, and F1-measure.

In Table 2, we report the results obtained from comparing the manual annotated relations with the relations extracted by SemRep and our system. Our system results were obtained by aggregating the results over the test folds during cross validation. As observed, our system managed to achieve significantly higher results than the SemRep results. To measure the statistically significant differences in performance between the two systems, we employed a randomization test based on stratified shuffling.[25]

To get a better insight of the features extracted for identifying treatment relations, we performed several feature ablation studies. The findings of these studies are listed in Tables 3(a) and 3(b). The experiments in Table 3(a) show the contribution of each feature to the overall performance of our system. As indicated, the largest drop in performance is caused by the *sem type* feature. This behavior is expected since there is a clear pattern in the semantic types corresponding to most of the concepts participating in treatment relations. For instance, the medication-disease and medication-medication relation types represent strong indicators for treatment and non-treatment relations, respectively. However, to properly estimate the contribution of SemRep and the MEDI algorithm to the overall performance of our system, we also performed experiments using the All – {*semrep*, *semrep or medi*} and All – {*medi*, *semrep or medi*} feature configurations. This is because the *semrep or medi* feature is highly correlated with the *semrep* and *medi* features. For instance, if *medi* is true for a given concept pair, the *semrep or medi* feature is also true for the same concept pair. Using the All – {*semrep*, *semrep or medi*} configuration, our system achieved a recall of 78.39 and an F1-measure of 82.03. When All – {*medi*, *semrep or medi*} was employed, the recall and F1-measure values dropped to 73.49 and 79.82, respectively.

The experiments from Table 3(b) show how our system performed when only one feature from Table 1 was selected to differentiate between treatment and non-treatment relations. The best performing experiments in this table are the ones employing the results of the MEDI algorithm (i.e., *medi* and *semrep or medi*). The results of these experiments are also statistically significant from the results of the *semrep* experiment (p<0.001). Also interestingly, the machine learning framework using only the *sem type* feature was able to obtain better results than SemRep (74.08 vs. 72.34). Nevertheless, the difference in performance was not statistically significant. Of note, the machine learning framework using only the *semrep* feature was able to find the same separation of the relations as the SemRep

Table 3 Feature ablation studies for treatment relation extraction.

(a)

Features	P	R	F
All – *semrep*	85.12	82.99	84.04
All – *medi*	84.95	83.09	84.01
All – *semrep or medi*	87.88	81.00	84.30
All – *unigram*	86.75	82.05	84.33
All – *bigram*	86.74	81.94	84.27
All – *expression*	87.58	81.73	84.56
All – *sem type*	85.94	81.00	83.40
All – *cui*	87.00	81.00	83.89

(b)

Features	P	R	F
semrep	81.17	65.24	72.34
medi	80.48	76.62	78.50
semrep or medi	73.61	85.91	79.29
unigram	63.56	39.87	49.01
bigram	77.89	16.18	26.79
expression	74.27	39.77	51.80
sem type	68.78	80.27	74.08
cui	68.80	49.48	57.56

All, the entire set of features from Table 1; F, F1-measure; P, precision; R, recall.

system. As observed, the SemRep results in Table 2 are identical with the results of the *semrep* experiment in Table 3(b). Similarly, the MEDI algorithm achieved the same performance results as the results of the *medi* experiment.

Error Analysis

The cases when the MEDI algorithm was not able to find an exact match for a given concept represented some of the most frequent false negative examples of our system. For instance, the relation *sucralfate→heartburn* does not have a corresponding medication-symptom pair in MEDI despite the fact that related concepts such as *esophagitis*, *burn of esophagus*, and *esophageal reflux* are included in the list of indications for *sucralfate* in this resource. Likewise, the relation *unasyn→pneumonia* cannot be matched by the algorithm although a medication-indication pair between *unasyn* and a more general concept of *pneumonia*, *communicable diseases*, exists in MEDI. Furthermore, from the false positive examples we analyzed, many of them occurred in complex sentences in which the context is critical in determining the relationship between concepts. The emphasized concepts in (4), e.g., represent a false positive instance because they correspond to a medication-indication pair in MEDI.

Discussion

Our experiments indicate that a machine learning framework is a successful approach for capturing treatment relations in clinical text. By incorporating various sources of lexical and semantic information associated with relation concepts as well as information extracted from a knowledge base of medication-indication pairs, our system managed to improve the performance results over SemRep, a widely used rule-based system in information extraction applications. The major improvements in recall over SemRep due to the information derived from MEDI confirmed our assumption that a valid medication-indication pair expressed in a sentence corresponds to a treatment relation in the majority of cases. Furthermore, as observed from the results of our experiments, the most significant decrease in recall and F1-measure is achieved when discarding all MEDI related features (i.e., using the All – {*medi*, *semrep or medi*} feature configuration). Despite the fact that SemRep was not particularly implemented for the clinical domain, our experiments from Tables 2 and 3(b) showed that this system is able to extract treatment relations with high precision. This study also adds to the relatively few evaluations of SemRep on clinical text, demonstrating that it does work in this domain also.

It is worth mentioning that both the MEDI algorithm and SemRep are not able to identify all treatment relation types. For instance, since the first argument of the treatment relation in (1) is a procedure, the concept pair from this example will not have a corresponding match in MEDI. Similarly, in SemRep, the types of the relations extracted from text are constrained to match the types of their corresponding relations in the UMLS Semantic Network.[7,26,27] As illustrated in Figure 1, only 9% of relations are identified by the two algorithms from the total number of relations extracted over the entire collection of 6,864 discharge summaries. The top 3 most frequent types of these relations are listed in Table 4. In this table, each relation type is described using the UMLS semantic types associated with the relation arguments. Not surprisingly, the most frequent relation type identified by both algorithms is the one abbreviated as orch,phsu→sosy, which represents the generic type of medication *treats* disease. On the other hand, the next two most frequent SemRep relation types (i.e., topp→dsyn and topp→podg) have procedure as first argument and therefore, they were not found among the MEDI relation types. From the types identified by the MEDI algorithm in Table 4, clnd→sosy was not found in the relation types extracted by SemRep.

Figure 1 The connection between the relations identified by the MEDI algorithm and SemRep.

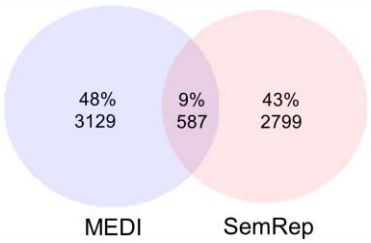


Table 4 Top 3 most frequent relation types identified by the MEDI algorithm and SemRep. The semantic types of the relation arguments are abbreviated as returned by MetaMap.

SemRep relation types	Freq.	MEDI relation types	Freq.
orch,phsu→sosy	425	orch,phsu→sosy	1582
topp→dsyn	324	orch,phsu→dsyn	616
topp→podg	264	clnd→sosy	270

clnd, clinical drug; dsyn, disease or syndrome; orch, organic chemical; phsu, pharmacologic substance; podg, patient or disabled group; sosy, sign or symptom; topp, therapeutic or preventive procedure.

In future research, we plan to improve our machine learning framework by implementing additional features that are able to better differentiate treatment relations from non-treatment relations. For instance, features using the structural and syntactic information derived from constituent and dependency trees could better capture the contextual properties between two medical concepts.[13,14,28] For this task, assertion classification[29] can also be investigated to better detect the relations whose corresponding treatment did not improve or cure a medical condition. Other technologies that may improve treatment relation extraction are statistical feature selection[30] and learning methods for imbalanced data.[31] Moreover, we intend to run our system over a large collection of clinical notes that could enable the discovery of new medication-indication pairs. Examples of pairs not in MEDI that our machine learning system was able to extract included valid relations such as *ethambutol→infection*, *lidocaine→stump pain*, *famvir→oral ulcers*, *levaquin→pyuria*, and *vesicare→bladder spasm*.

Conclusion

In this paper, we described a supervised learning system for identifying treatment relations in clinical notes. Our system successfully integrated various types of information which lead to achieving significantly better performance results than SemRep. One relevant source of information which had a major impact in boosting our system's recall is MEDI. As we empirically proved, MEDI is a broad and reliable resource on assessing the validity of treatment relations. We believe that future information extraction systems in the clinical domain should rely on a knowledge base of medication-indication pairs to accurately identify treatment relations in text. We plan to further improve this task and to assess its usability in various clinical applications.

Acknowledgements

This work was supported by NLM/NIH grants 5 T15 LM007450-12 and 1 R01 LM010685. The dataset used for the analyses described were obtained from Vanderbilt University Medical Center's Synthetic Derivative which is supported by institutional funding and by the Vanderbilt CTSA grant ULTR000445 from NCATS/NIH.

References

1 Cebul RD, Love TE, Jain AK, *et al.* Electronic health records and quality of diabetes care. *N Engl J Med* 2011;**365**:825–33.

2 Ghitza UE, Sparenborg S, Tai B. Improving drug abuse treatment delivery through adoption of harmonized electronic health record systems. *Subst Abuse Rehabil* 2011;**2011**:125–31.

3 Liu M, Wu Y, Chen Y, *et al.* Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs. *J Am Med Inform Assoc* 2012;**19**:28–35.

4 Roth CP, Lim YW, Pevnick JM, *et al.* The challenge of measuring quality of care from the electronic health record. *Am J Med Qual* 2009;**24**:385–94.

5 Roth MT, Weinberger M, Campbell WH. Measuring the quality of medication use in older adults. *J Am Geriatr Soc* 2009;**57**:1096–102.

6 Wei WQ, Cronin RM, Xu H, *et al.* Development and evaluation of an ensemble resource linking medications to their indications. *J Am Med Inform Assoc* 2013;**20**:954–61.

7 Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *J Biomed Inform* 2003;**36**:462–77.

8 Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: *Proc AMIA Symp*. 2001. 17–21.

9 Fiszman M, Rindflesch TC, Kilicoglu H. Abstraction summarization for managing the biomedical research literature. In: *HLT-NAACL Workshop on Computational Lexical Semantics*. 2004. 76–83.

10 Hristovski D, Friedman C, Rindflesch TC, *et al.* Exploiting Semantic Relations for Literature-Based Discovery. In: *Proc AMIA Symp*. 2006. 349–53.

11 Rindflesch TC, Pakhomov SV, Fiszman M, *et al.* Medical Facts to Support Inferencing in Natural Language Processing. In: *Proc AMIA Symp*. 2005. 634–8.

12 Liu Y, Bill R, Fiszman M, *et al.* Using SemRep to label semantic relations extracted from clinical text. In: *Proc AMIA Symp*. 2012. 587–95.

13 Roberts A, Gaizauskas R, Hepple M, *et al.* Mining clinical relationships from patient narratives. *BMC Bioinformatics* 2008;**9 (Suppl. 11)**.

14 Uzuner Ö, Mailoa J, Ryan R, *et al.* Semantic relations for problem-oriented medical records. *Artif Intell Med* 2010;**50**:63–73.

15 Uzuner Ö, South BR, Shen S, *et al.* 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc* 2011;**18**:552–6.

16 Rink B, Harabagiu S, Roberts K. Automatic extraction of relations between medical concepts in clinical texts. *J Am Med Inform Assoc* 2011;**18**:594–600.

17 De Bruijn B, Cherry C, Kiritchenko S, *et al.* Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *J Am Med Inform Assoc* 2011;**18**:557–62.

18 Minard AL, Ligozat AL, Ben Abacha A, *et al.* Hybrid methods for improving information access in clinical documents: concept, assertion, and relation identification. *J Am Med Inform Assoc* 2011;**18**:588–93.

19 Patrick JD, Nguyen DHM, Wang Y, *et al.* A knowledge discovery and reuse pipeline for information extraction in clinical notes. *J Am Med Inform Assoc* 2011;**18**:574–9.

20 Bejan CA, Wei WQ, Denny JC. Using SemRep and a medication indication resource to extract treatment relations from clinical notes. In: *AMIA Jt Summits Transl Sci Proc*. 2014.

21 Kuhn M, Campillos M, Letunic I, *et al.* A side effect resource to capture phenotypic effects of drugs. *Mol Syst Biol* 2010;**6**.

22 Denny JC, Smithers JD, Miller RA, *et al.* 'Understanding' medical school curriculum content using KnowledgeMap. *J Am Med Inform Assoc* 2003;**10**:351–62.

23 Denny JC, Spickard A 3rd, Miller RA, *et al.* Identifying UMLS concepts from ECG Impressions using KnowledgeMap. In: *Proc AMIA Symp*. 2005. 196–200.

24 Stenetorp P, Pyysalo S, Topić G, *et al.* brat: a Web-based Tool for NLP-Assisted Text Annotation. In: *Demonstrations Session at EACL*. 2012. 102–7.

25 Noreen EW. *Computer-Intensive Methods for Testing Hypotheses: An Introduction*. 1st ed. New York: : John Wiley & Sons 1989.

26 Rindflesch TC, Fiszman M, Libbus B. Semantic interpretation for the biomedical research literature. In: Chen H, Fuller SS, Friedman C, *et al.*, eds. *Medical Informatics: Knowledge Management and Data Mining in Biomedicine*. 2005. 399–422.

27 Ahlers CB, Fiszman M, Demner-Fushman D, *et al.* Extracting semantic predications from Medline citations for pharmacogenomics. In: *Pac Symp Biocomput*. 2007. 209–20.

28 GuoDong Z, Jian S, Jie Z, *et al.* Exploring various knowledge in relation extraction. In: *Annual Meeting on Association for Computational Linguistics*. 2005. 427–34.

29 Bejan CA, Vanderwende L, Xia F, *et al.* Assertion modeling and its role in clinical phenotype identification. *J Biomed Inform* 2013;**46**:68–74.

30 Bejan CA, Xia F, Vanderwende L, *et al.* Pneumonia identification using statistical feature selection. *J Am Med Inform Assoc* 2012;**19**:817–23.

31 He H, Garcia EA. Learning from Imbalanced Data. *IEEE Trans Knowl Data Eng* 2009;**21**:1263–84.

An Exploration of the Potential Reach of Smartphones in Diabetes

Katherine S Blondon, MD, PhD[1], Paul L Hebert, PhD[1], James D Ralston, MD, MPH[1]
[1]University of Washington, Seattle, WA

Abstract

Although smartphones bear potential to improve diabetes self-management, the reach of smartphones in diabetic populations remains uncertain. Using survey data from the Pew Research Center, we compared smartphone use in individuals with and without diabetes, and determined factors associated with smartphone use among those with diabetes. Of the 2989 adults surveyed, 1360 were smartphone users, and 332 individuals had diabetes. Compared to individuals without diabetes, adults with diabetes were less likely to be smartphone users (relative risk of 0.43, 95% CI 0.31 to 0.54) even after adjusting for age, race, ethnicity and socioeconomic status (adjusted RR of 0.78, 95%CI 0.57-0.98). Among individuals with diabetes, high income, younger age and online health information seeking were associated with higher smartphone use. While smartphones can reach subgroups for diabetes care and prevention (racial/ethnic minorities, newly diagnosed individuals), studies are needed to understand this current difference in smartphone use.

Introduction

Smartphones have shown potential in improving diabetes self-management. The rapid adoption of smartphones[1] and the rising prevalence of diabetes[2] have led to the development of a large array of supportive applications aimed at improving diabetes self-management in daily tasks and choices: they help learn about diabetes and its management, track health parameters with graphs and data sharing options, and allow calendaring for daily tasks or annual screening procedures[3, 4]. Seven out of ten American adults track a health parameter such as weight, exercise or sleep[5], and one in five smartphone users having a health app on their device[6]. Although treatment goals are well established[7], less than one fifth of patients achieve HbA1c, blood pressure and LDL-cholesterol goals[8]. A meta-analysis of 22 trials assessing the effect of mobile phone interventions on glycemic control showed a reduction of HbA1c of 0.5% over a median of 6 months' follow-up duration[9]. A randomized controlled trial with a smartphone application that offers automated clinical coaching based on patient-reported data and that supports data-sharing with care providers, showed a 1.2% greater decrease in HbA1c with web- and mobile-based tools compared to usual care over a year (1.9% in web- and mobile-based group vs 0.7% in usual care group, $p<0.001$)[10]. Early studies suggest promise of effective support for diabetes self-management with smartphone applications.

The uptake of smartphone applications depends on consumer access to the devices, as well as on human-computer interaction issues including sufficient functionality and usability[11]. Although studies have characterized who has diabetes and smartphone users, we do not have a clear understanding of the socio-demographic characteristics associated with smartphone use among individuals with diabetes. The exponential growth in smartphone adoption has reached more than half of US adults, who are from all demographic groups[1, 12]. The diffusion of innovation theory[13] claims that age predicts the adoption of technologies, with the younger population adopting new technologies before the older population. Although higher income and higher education are strong predictors of all-age smartphone adoption, younger adults seem to be less affected by these variables. Asians and African Americans are more likely to be smartphone users than Caucasians, and African-Americans are also more likely to have health applications on their phone than other racial/ethnic groups (15% of African Americans vs 7% of whites and 11% of Latinos). This difference is particularly relevant, as these groups also have a higher prevalence of diabetes (11.8% of Hispanics, 12.6% of non-Hispanic Blacks compared to 8.3% of the U.S. population)[14, 15], with lower glycemic control in African Americans and Hispanics. Diabetes prevalence is also higher with increasing age (13.7% in the 45-64 year old group, and 26.9% in those over 65 years old)[15].

The aim of this study is to identify the reach of use smartphones among individuals with diabetes and to examine the socio-demographic profile of smartphone users with diabetes. Although many smartphone and tablet applications to support diabetes self-management are available, characteristics of smartphone users in the targeted population remain unclear. These characteristics are important, if developers desire to tailor mobile apps to the users' needs.

We hypothesized that compared to individuals without diabetes, individuals with diabetes will be less likely to use smartphones, due to older age and lower socioeconomic status (SES)[15]. We also hypothesized that individuals with diabetes would be more likely to use smartphones if they have a higher level of education, younger age, non-White

race and high income[16]. Finally, as many of the current diabetes application features involve data tracking and interpretation[3, 4], we hypothesized that smartphone use among diabetics would be associated with higher self-monitoring. Understanding who uses and does not use smartphones among patients with diabetes can inform future diffusion and design of diabetes applications.

Methods

Population

This cross-sectional study uses a survey dataset of the Pew Research Center collected between August 7 and September 6, 2012. This data was collected initially to describe the use of mobile technology for health[17]. Researchers conducted a nationwide survey in the U.S. using phone interviews in English and Spanish of 3,014 adults >18 years old (1,808 landline and 1,206 mobile users). Participant selection was by random digit dial of both landline and cellphone numbers. For landline sample, the interviewers followed a systematic respondent selection technique that closely mirrors the population in terms of age and gender when combined with cellphone sample. Both samples were oversampled for Black and Hispanic respondents. Response rates were 11.5% for landlines and 6.6% for cell phones (total of 58,848 landlines and 32,129 cell phones dialed). The survey weights supplied with the survey[17] were used to generate population level estimates. The margin of sampling error was ±2.4 percentage points for the complete set of weighted data. This survey was conducted by Princeton Survey Research Associates International and sponsored by the Pew Research Center's Internet & American Life Project and the California HealthCare Foundation. This current study was conducted with the permission of the Pew Research Center's Internet & American Life project.

Survey

This phone survey was created to study how cell phones, particularly smartphones, are used to look for health information. It included questions about self-reported health conditions (overall health and common chronic diseases), tracking behaviors (in particular of weight, diet, exercise, and blood pressure), use of social networks, and searching for online health information about specific diseases or treatments, heath insurance, or more general food and drug safety information. The complete interview (26 questions) can be found online[18].

Analysis

The primary analysis (Table 3) compared the use of smartphones between diabetic and non-diabetic participants using unadjusted and adjusted weighted logistic regression models. Smartphone users were defined as any user of a smartphone, including those who had both a landline and a smartphone. Non-smartphone users were all other participants, and included feature phone (cell phone that is not a smartphone) users and landline-only users. The covariates in the logistic regression models were defined a priori, with an age-adjusted model and the full model (age, measures of socio-economic status, race, ethnicity). Age was hypothesized to be the main confounder of smartphone use and diabetes. The other demographic confounders are based on results from market studies of smartphone use. In these models, age, race, income and education were categorical variables, as presented in tables 1 and 2.

The secondary analysis (Table 4) used a multivariate logistic regression model among diabetics to estimate predictors of smartphone use. All variables of interest were chosen a priori by the investigators and entered together in the model: age (categorical), race and ethnicity, education and income (all binary). For this secondary analysis, we used a binary variable combining race and ethnicity, as well as binary variables for education and income to avoid having too few events per variable. Education was defined using a threshold of high school completion. The cutoff for income level of $50,000 per year was based on the annual median household income in the U.S. of $50,502 in 2011[19]. The second model for this analysis also included online health information seeking as a proxy for patient engagement, and health tracking (both binary).

All analyses and characteristics of the sample were computed from the weighted sample after exclusion of missing diabetes or smartphone values (N=2989). The percentage of complete cases was 78%. Missing covariate data were infrequent (≤1.8%) other than for income (18.5%). Missing data were multiple-imputed with 25 imputed datasets using imputation by chained-equations[20]. The imputation model included the covariates used in all our analysis (with dependent variables), as well as three auxiliary variables: state, use of urgent care in the past 12 months and self-reported health. Categorical variables were compared using Chi-Square tests. P values from regression models were derived from Wald tests with robust standard errors. A p-value <0.05 determined statistical significance. Residual confounding was assessed by testing the effect modification of diabetes by age group. As smartphone use

is not a rare event, odds ratios do not approximate relative risks. To avoid misinterpretation of the results, we presented our results after conversion to relative risks using the margins function[2121212120] of Stata 11 (Stata Corporation, Texas).

Results

We describe the diabetic and non-diabetic populations in our sample (Table 1), with weighted descriptive analyses. Among the 2991 respondents, 332 had diabetes (weighted proportion of 11.1%). Individuals with diabetes were significantly older than non-diabetic participants (mean age of 59.9y vs 45.5y, p<.001, Table 1). They also had more comorbidities (hypertension, congestive heart failure and other chronic diseases such as asthma or cancer). Gender and ethnicity were not different, but more participants with diabetes were of Black race than participants without diabetes (16.9% vs 12.7%). Socio-economic status was lower in the diabetic group, with fewer insured participants, lower educational attainment, a lower income and a higher prevalence of unemployment. Compared with non-diabetics, individuals with diabetes had a significantly higher feature phone use (50.6% vs 38.3%, p<.001 Table 1) and a lower smartphone use (20.7% vs 48.6%, p<.001). About half of the patients with diabetes used email and Internet, compared with over three quarters of individuals without diabetes. Moreover, five out of six participants with diabetes tracked at least one health parameter (diet, weight, carbohydrates, etc.), compared with two thirds of non-diabetics (p<.001). Individuals with diabetes had a lower use of health applications on smartphones (3.8% vs 10.0%, p=.002).

In Table 2, we report the characteristics of smartphone users in our sample (both diabetic and non-diabetic individuals included) compared with non-smartphone users. This latter group includes individuals who use feature phones and/or landlines. The sample comprised 1360 smartphone users (weighted proportion 45.5%) and 1629 non-smartphone users (1185 feature phone users and 444 individuals with only landlines, weighted proportions 39.6% and 14.8% respectively). The landline only population was older, with more comorbidities: they had the highest prevalence of diabetes (21.4%), compared with feature phone users (14.2%) and smartphone users (5.0%). They had a higher proportion of Caucasians (81.2%) and Hispanics (16.1%) than the feature phone and smartphone users, and had lower educational attainment. They had lower use of email and Internet. Gender was not significantly different among these groups.

Table 1. Weighted comparison of individuals with and without diabetes in the study population

	Diabetes	**No diabetes**
Total N	332 (11.1%)	2657 (88.9%)
Mean age (SD, 95% CI)*	59.9y (1.1, 57.8-62.0)	45.5y (0.4, 44.7-46.4)
18-35 y	8.8%	34.6%
36-50 y	16.8%	28.2%
51-64 y	35.4%	21.4%
65-80 y	31.5%	11.4%
>80 y	7.5%	4.4%
Male	53.5%	50.9%
Comorbidities:		
Hypertension*	67.1%	20.1%
Heart disease*	27.5%	4.9%
Other chronic disease*	82.9%	38.0%
Race		
White	70.2%	74.0%
Black	16.9%	12.7%
Asian	1.6%	3.0%
Other race	11.3%	10.3%
Ethnicity		
Hispanic	15.8%	13.2%
Health insurance*		
Uninsured	11.2%	18.5%
Medicaid	11.2%	7.4%
Medicare	24.3%	7.1%
Private insurance	50.8%	64.4%
Attained education*		
No high school	20.7%	10.4%
High school	62.1%	59.2%
College or higher	17.2%	30.4%
Annual income*		
<30,000$	60.2%	36.4%
30,000-99,999$	36.0%	47.1%
≥100,000$	3.8%	16.5%
Employed*	30.3%	58.0%
Feature phone users*	50.6%	38.3%
Smartphone users*	20.7%	48.6%
Use of Internet*	53.4%	81.6%
Use of email*	47.6%	75.4%
Tracks any health parameter*	84.3%	64.9%
Use of health app on smartphone*	3.9%	10.05

*p≤.001

Table 2. Weighted comparison of smartphone users and non-smartphone users (landline or feature phones users)

	Smartphone users	**Not smartphone users**
Total N	1360 (45.5%)	1629 (54.5%)
Landline only (% total population)	-	444 (14.8%)
Diabetes (%)*	5.0%	16.1%
Mean age (SD, 95% CI)*	38.9 y (0.5, 38.0-39.8)	54.0 y(0.6, 52.8-55.2)
18-35 y	46.5%	19.4%
36-50 y	32.9%	21.9%
51-64 y	16.0%	28.7%
65-80 y	3.8%	21.9%
>80 y	0.7%	8.1%
Male	50.1%	52.1%
Comorbidities:		
Hypertension*	14.7%	34.1%
Heart disease*	3.1%	11.0%
Other chronic disease*	30.3%	53.5%
Race*		
White	69.5%	77.0%
Black	14.0%	12.5%
Asian	5.1%	1.0%
Other race	11.4%	9.5%
Ethnicity		
Hispanic	14.7%	12.6%
Health insurance*		
Uninsured	17.7%	17.6%
Medicaid	5.4%	9.9%
Medicare	2.6%	14.3%
Private insurance	72.2%	55.2%
Attained education*		
No high school	5.5%	16.5%
High school	56.0%	62.4%
College or higher	38.4%	21.1%
Annual income*		
<30,000$	27.7%	48.5%
30,000-99,999$	49.6%	42.8%
≥100,000$	22.7%	8.7%
Employed*	72.3%	40.4%
Use of internet*	97.7%	62.4%
Use of email*	92.3%	55.7%
Tracks any health parameter	67.9%	66.3%

*p<.001.
(The smartphone user group includes users of both smartphone and landline phones)

Compared to White participants, Asian participants had a significantly higher use of smartphones (Table 2). The smartphone population also had significantly higher income, employment rate and higher educational attainment than the non-smartphone population. The proportion of uninsured did not differ with smartphone use. Finally, use of emails and Internet was almost universal in the smartphone group (92% and 98% of individuals, respectively), compared to less than two thirds of the non-smartphone group. Two thirds of the participants tracked some health parameter, regardless of the type of phone technology.

The results of the primary analysis comparing the use of smartphone between diabetic and non-diabetic participants are presented in Table 3. In the unadjusted analysis, individuals with diabetes were less likely to use smartphones compared with those without diabetes (RR 0.43, 95% CI 0.31 to 0.54, p<.001). After adjusting for age, individuals with diabetes were still less likely to be smartphone users compared with those without diabetes (RR 0.58, 95%CI 0.40 to 0.75, p<.001). In the full model that adjusted for race, ethnicity, income and education level (potential confounders), participants with diabetes remained significantly less likely to be smartphone users compared with those without diabetes (RR 0.78, 95% 0.57-0.98, p=0.05). In this multivariate model, we also observed that Blacks, Asians and Hispanics were more likely to use a smartphone than Caucasians and non-Hispanics, respectively. We also found strong evidence that a higher income and education attainment was positively associated with smartphone use. There was no significant residual confounding by age in the adjusted analysis.

The results of the secondary analysis of predictors of smartphone use among individuals with diabetes are shown in Table 4. In the multivariate model 1, younger age and higher income were strongly associated with smartphone use, whereas race/ethnicity and education were not. Model 2 further explored online health seeking behavior and health tracking behavior. Individuals who sought health information online were more likely to be smartphone users (RR 3.68, 95% CI 1.06-6.30, p<.001). The individuals who tracked health parameters, however, were *less* likely to be smartphone users (RR of 0.62, 95%CI 0.36-0.88, p=0.04).

Table 3. Unadjusted and adjusted RR comparing individuals with diabetes to individuals without diabetes for smartphone use (N=2989)

	Unadjusted			**Age-adjusted**			**Full model**		
	RR	(95%CI)	p-value	RR	(95%CI)	p-value	RR	(95%CI)	p-value
Diabetes	0.43	(0.31-0.54)	<.001	0.58	(0.40-0.75)	<.001	0.78	(0.57-0.98)	0.05
Age						<.001			<.001
18-35 y				(Ref)			(Ref)		
36-50 y				0.81	(0.70-0.92)		0.82	(0.74-0.90)\	
51-64 y				0.45	(0.38-0.53)		0.48	(0.39-0.57)	
65-80 y				0.38	(0.13-0.23)		0.24	(0.17-0.31)	
>80 y				0.09	(0.04-0.15)		0.10	(0.04-0.17)	
Race									<.001
White							(Ref)		
Black							1.30	(1.04-1.56)	
Asian							2.10	(1.45-2.75)	
Other race							1.07	(0.75-1.39)	
Hispanic							1.30	(1.02-1.57)	0.02
Annual income:									<.001
<30,000$							(Ref)		
30,000-99,999$							1.65	(1.31-1.99)	
≥100,000$							2.63	(1.98-3.28)	
Attained education:									<.001
No high school							(Ref)		
High school							1.82	(1.14-2.49)	
College or higher							2.52	(1.53-3.51)	

Table 4. Multivariate predictors of smartphone use among individuals with diabetes (N=332)

Covariates	**Model 1**			**Model 2**		
	RR	**95%CI**	**p-value**	**RR**	**95%CI**	**p-value**
Age 18-35 y	(Ref)		<.001	(Ref)		<.001
Age 36-50 y	0.37	(0.15-0.60)		0.50	(0.18-0.83)	
Age 51-64 y	0.28	(0.12-0.44)		0.39	(0.15-0.63)	
Age >65 y	0.13	(0.04-0.22)		0.26	(0.08-0.44)	
Non-White or Hispanic	1.42	(0.77-2.06)	0.13	1.41	(0.82-2.00)	0.10
High school education	1.22	(0.51-1.93)	0.50	0.96	(0.45-1.47)	0.88
Annual income > 50,000$	3.09	(1.35-4.84)	<.001	2.34	(1.13-3.54)	0.002
Seeks health information online				3.68	(1.06-6.30)	<.001
Tracks any health parameter				0.62	(0.36-0.88)	0.04

Discussion

In our nationwide sample, we found that individuals with diabetes were less likely to be smartphone users compared with individuals without diabetes, even after adjusting for potential confounding by age, SES, race and ethnicity.

The pattern of lower use among the older population is consistent with market studies on smartphone adoption[1, 22]. Although early evidence supports the long-term effectiveness of mobile technologies in diabetes self-management with improved HbA1c values after 6 months[9] and 12 months[10], our results suggest that these technologies might not be appropriate for all individuals, and efforts to improve standard care in diabetes self-management should continue to include more traditional contacts with patients including in-person and standard telephone communications.

Understanding lower smartphone use among individuals with diabetes is important, as healthcare delivery systems are likely to move towards a higher use of smartphone applications for diabetes self-management. Diabetes, its long-term complications and related comorbidities can lead to physical and cognitive impairments, such as lower dexterity from neuropathy or visual impairments. All these unmeasured factors are barriers to smartphone use, and are all accentuated by older age. Despite the slow increase in smartphone uptake among older adults, the very rapid uptake among younger adults may accentuate the age gap in smartphone use. The trend towards larger screens of newer devices (phablets, mini-tablets and tablets) and improved usability are promising approaches to address visibility and dexterity impairments related to diabetes and age[23, 24]. Finally, care-providers might also have a role to play in diffusing diabetes applications to patients who already use smartphones[25], as they already provide guidance for websites and online communities to their patients[26]. Our findings support further research on understanding the differences in the use of smartphone use among patient with diabetes.

The 51 to 64 year-old age group may benefit the most from smartphone-based interventions for diabetes, as it is the second fastest growing age group for smartphone uptake[12]. We found that one in six individuals in this age group currently used smartphones. With over a million individuals from this age group newly diagnosed with diabetes each year in the U.S.[2], the potential reach of smartphone-based interventions should not be overlooked, particularly for the early stages of disease[27]. Smartphone interventions have the potential to lower HbA1c values effectively, (e.g., mean HbA1c reduction of 1.9% in the smartphone group vs 0.7% in the usual care group (P=0.001) over 12 months)[10]. Any 1% reduction in HbA1c is associated with significant reduction of the risk for myocardial infarction (14%) and stroke (12%)[28]. Furthermore, not only does the benefit of HbA1c reduction increase over time, the reduction in all-cause mortality is greater when the HbA1c reduction occurs early in the disease[29].

As smartphone use is highest among young adults, it may offer unique opportunities for early diabetes self-management or diabetes prevention. Most diabetes applications provide tracking tools to monitor health parameters, and can guide early disease management, making these tools particularly useful at early stages of disease)[3, 27]. In addition, the reach of smartphones among racial/ethnic minorities might allow early prevention by supporting behavior change in this subpopulation with higher prevalence of early type 2 diabetes, in particular among adolescents.

In contrast to the age gap, smartphones have greater potential to help bridge the typical “digital divide” in Internet and computer access in racial and ethnic groups. The higher smartphone use by Blacks, Asians and Hispanics compared to Whites and non-Hispanics, reflect the importance of race and ethnicity for smartphone use[1, 16]. This finding has two important implications: (1) smartphones might be a new approach to improve access to high quality care for diabetes in racial/ethnic minority groups, and (2) future diabetes applications need to take cultural differences into consideration in their design. For instance, only a few diabetes applications currently offer the option of Spanish. Also, most applications or websites do not include options for cultural preferences in their food plans. One possible design implication could be to integrate more culturally adapted nutrition facts in smartphone applications to facilitate the adoption and adherence to diabetes-friendly food plans in this racial/ethnic minority groups.

Contrary to our hypothesis of smartphone use for health tracking, we found that individuals were less likely to be a smartphone user if they monitored their own health. Nearly two thirds of our participants tracked some health parameter, regardless of phone type. Individuals with diabetes predominantly have type 2 diabetes and do not use insulin, and therefore do not require highly intensive health tracking. In a prior study, patients seemed to categorize many of their health results as normal or abnormal, rather than recall the exact values, and therefore are content to check with little or not tracking[27, 30]. They tended to use traditional tracking methods (pen and paper, or websites) according to a recent survey[26]. Smartphones offer a wide range of features, and are not primarily adopted for health tracking purposes. This underlines the gap in assessing overall smartphone use rather than the use of diabetes-related applications. Furthermore, there is a delay in the diffusion of diabetes applications among smartphone users, and the use of pervasive technologies in smartphones to effortlessly track health.

Strengths and Limitations

One of the strengths of this paper is its secondary use of a dataset collected to study mobile health. It uses a random sample of landline and cellphone users from the national U.S. population, without targeting any disease in particular. This helps avoid the bias related to successful diabetes self-management. The survey was also conducted in two languages, English and Spanish, which facilitate participation from the rapidly growing Hispanic population.

Although the prevalence of diabetes in this sample (11.1% after weighting) is comparable to the prevalence of diabetes among U.S adults, a limitation of this study is its use of self-reported diabetes, as undiagnosed diabetes is estimated to be about a third of cases in the United States[2]. This limitation could contribute to the lack of association seen between smartphone use and race/ethnicity among individuals with diabetes. Furthermore, the survey did not differentiate between type 1 and type 2 diabetes. Yet type 2 is more prevalent among older adults and in lower SES populations[15]. The low survey response rate for both the landline and cellphone samples may also limit the generalizability of our findings. Moreover, this dataset only provided information about self-reported diabetes, without information about type of disease. Future studies should explore the association between disease duration, severity (HbA1c and comorbidities) and type of treatment (insulin use) and smartphone use.

Implications

Smartphones have generated great interest for patient empowerment, and bear potential in improving diabetes self-management in particular. Although the exponential growth of smartphone adoption now affects all age groups, this growth remains moderate among older individuals with lower income. Individuals with diabetes are older and have lower income than the general population: they therefore remain less likely to have a smartphone than individuals without diabetes. So although it is important for clinicians to be familiar with smartphone tools for diabetes, especially for subgroups of younger, newly diagnosed individuals or those from racial/ethnic minorities, we underline the importance of pursuing efforts to improve traditional diabetes care (in-person visits and phone calls) at this time. Further research is needed to understand this gap in smartphone adoption in the diabetic population, both in terms of access to the devices as well as for usability and design (font and screen sizes, for example).

Conclusion

This study is, to the best of our knowledge, the first report on the characteristics of smartphone users in a population of patients with diabetes. Compared to individuals without diabetes, smartphone use remains significantly less likely among individuals with diabetes. Since smartphones have the potential to improve self-management support for diabetes care, further research is needed to better understand this gap in smartphone use, in particular to address diabetes- and age-related differences. Our findings also emphasize the potential of smartphones to help prevent diabetes in younger adults, and to improve access to care for racial and ethnic minorities.

Acknowledgements

We acknowledge the Pew Internet & American Life Project, who granted us permission to use their dataset for this study.

References

1. Nielsenwire. America's New Mobile Majority: a Look at Smartphone Owners in the U.S. 2012 [cited 2012 6/1/2012]; Available from: http://www.nielsen.com/us/en/newswire/2012/who-owns-smartphones-in-the-us.html.
2. (CDC) CfDCaP. 2011 National Diabetes Fact Sheet. 2011 [6/15/2012]; Available from: http://www.cdc.gov/diabetes/pubs/factsheet11/fastfacts.htm.
3. Chomutare T, Fernandez-Luque L, Arsand E, Hartvigsen G. Features of mobile diabetes applications: review of the literature and analysis of current applications compared against evidence-based guidelines. Journal of medical Internet research. 2011;13(3):e65. Epub 2011/10/08.
4. El-Gayar O, Timsina P, Nawar N, Eid W. Mobile applications for diabetes self-management: status and potential. Journal of diabetes science and technology. 2013;7(1):247-62. Epub 2013/02/27.
5. Fox S, Duggan M. Mobile Health 2012. 2012.
6. Fox S. The e is for Engagement. Pew Internet; 2012; Available from: http://www.pewinternet.org/Commentary/2012/October/The-e-is-for-engagement.aspx.
7. Standards of medical care in diabetes--2014. Diabetes Care. 2014;37 Suppl 1:S14-80. Epub 2013/12/21.

8. Vouri SM, Shaw RF, Waterbury NV, Egge JA, Alexander B. Prevalence of achievement of A1c, blood pressure, and cholesterol (ABC) goal in veterans with diabetes. Journal of managed care pharmacy : JMCP. 2011;17(4):304-12. Epub 2011/05/04.
9. Liang X, Wang Q, Yang X, Cao J, Chen J, Mo X, et al. Effect of mobile phone intervention for diabetes on glycaemic control: a meta-analysis. Diabetic medicine : a journal of the British Diabetic Association. 2011;28(4):455-63. Epub 2011/03/12.
10. Quinn CC, Shardell MD, Terrin ML, Barr EA, Ballew SH, Gruber-Baldini AL. Cluster-randomized trial of a mobile phone personalized behavioral intervention for blood glucose control. Diabetes Care. 2011;34(9):1934-42. Epub 2011/07/27.
11. LeRouge C, Wickramasinghe N. A review of user-centered design for diabetes-related consumer health informatics technologies. Journal of diabetes science and technology. 2013;7(4):1039-56. Epub 2013/08/06.
12. Nielsen. Generation App: 62% of Mobile Users 25-34 own Smartphones. 2011 [12/6/2011]; Available from: http://blog.nielsen.com/nielsenwire/online_mobile/generation-app-62-of-mobile-users-25-34-own-smartphones/.
13. Rogers EM. Diffusion of innovations. 4th ed. New York, London: Free Press ; Collier Macmillan; 1995. xix, 453 p. p.
14. Lee JW, Brancati FL, Yeh HC. Trends in the prevalence of type 2 diabetes in Asians versus whites: results from the United States National Health Interview Survey, 1997-2008. Diabetes Care. 2011;34(2):353-7. Epub 2011/01/11.
15. Diabetes statistics. American Diabetes Association; 2012; Available from: http://www.diabetes.org/diabetes-basics/diabetes-statistics/.
16. Pew Research Center FS. Mobile Health 2010. 2010 Oct. 19, 2010. Report No.
17. Pew Research Center FS, Duggan M. Mobile Health 2012. 2012 Nov.8, 2012. Report No.
18. Pew Research Center FS, Duggan M. Explore survey questions 2012; Available from: http://www.pewinternet.org/Static-Pages/Data-Tools/Explore-Survey-Questions/Roper-Center.aspx?item=%7B96693466-DAF7-48DA-BA09-9F8C2A35DDC9%7D.
19. Noss A. Household Income for States: 2010 and 2011. United States Census Bureau, 2012.
20. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Statistics in medicine. 2011;30(4):377-99. Epub 2011/01/13.
21. How can I get margins for a multiply imputed survey logit model? [May 15, 2013]; Available from: http://www.ats.ucla.edu/stat/stata/faq/mi_svy_logit.htm.
22. Nielsen. Survey: New U.S. Smartphone Growth by Age and Income. 2012 [cited 2012 3/16/2012]; Available from: http://blog.nielsen.com/nielsenwire/?p=30950.
23. Cheung A, Janssen A, Amft O, Wouters EF, Spruit MA. Usability of digital media in patients with copd: a pilot study. International journal of technology assessment in health care. 2013;29(2):162-5. Epub 2013/04/05.
24. Chun YJ, Patterson PE. A usability gap between older adults and younger adults on interface design of an Internet-based telemedicine system. Work. 2012;41 Suppl 1:349-52. Epub 2012/02/10.
25. Drinkwater D. 3 in 4 physicians are using tablets; some are even prescribing apps. TabTimes; 2013 [May 5,2013]; Available from: http://tabtimes.com/news/ittech-stats-research/2013/04/19/3-4-physicians-are-using-tablets-some-are-even-prescribing.
26. Research M. New study reveals that physicians embrace patient self-tracking. 2013 [April 21, 2013]; Available from: http://manhattanresearch.com/News-and-Events/Press-Releases/physicians-embrace-patient-self-tracking.
27. Blondon K, Klasnja, P., Kendall, L., Pratt, W. Long-Term Engagement with Health-Management Technology: a Dynamic Process in Diabetes. (forthcoming). 2013.
28. Stratton IM, Adler AI, Neil HA, Matthews DR, Manley SE, Cull CA, et al. Association of glycaemia with macrovascular and microvascular complications of type 2 diabetes (UKPDS 35): prospective observational study. Bmj. 2000;321(7258):405-12. Epub 2000/08/11.
29. Vinall P. Explaining the UKPDS legacy effect: The Goodwin Group International; 2012. Available from: http://www.mdconferencexpress.com.
30. Guirguis LM, Kieser MA, Chewning BA, Kanous NL. Recall of A1C, blood pressure, and cholesterol levels among community pharmacy patients with diabetes. Journal of the American Pharmacists Association : JAPhA. 2007;47(1):29-34. Epub 2007/03/07.

Analyzing U.S. prescription lists with RxNorm and the ATC/DDD Index

Olivier Bodenreider, MD, PhD, Laritza M. Rodriguez, MD, PhD
Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Bethesda, MD

Abstract

***Objectives**: To evaluate the suitability of the ATC/DDD Index (Anatomical Therapeutic Chemical (ATC) Classification System/Defined Daily Dose) for analyzing prescription lists in the U.S. **Methods**: We mapped RxNorm clinical drugs to ATC. We used this mapping to classify a large set of prescription drugs with ATC and compared the prescribed daily dose to the defined daily dose (DDD) in ATC. **Results**: 64% of the 11,422 clinical drugs could be precisely mapped to ATC. 97% of the 87,001 RxNorm codes from the prescription dataset could be classified with ATC, and 97% of the prescribed daily doses could be assessed. **Conclusions**: Although the mapping of RxNorm ingredients to ATC appears to be largely incomplete, the most frequently prescribed drugs in the prescription dataset we analyzed were covered. This study demonstrates the feasibility of using ATC in conjunction with RxNorm for analyzing U.S. prescription datasets for drug classification and assessment of the prescribed daily doses.*

1. Introduction

Medication errors have been identified as a significant cause of mortality in hospitalized patients [1] and medication safety remains an important issue today [2]. Medication dose errors are a specific category of medication errors [3]. Large variations can be observed in prescribed doses, some of which correspond to medication dose errors, including tenfold medication dose errors [4]. One strategy for reducing medication errors, including dose errors, is to use Computerized Physician Order Entry (CPOE) systems offering clinical decision support [5]. The information used for clinical decision support in CPOEs generally comes from proprietary drug knowledge bases.

The Anatomical Therapeutic Chemical (ATC) classification of drugs is widely available and provides basic information such as drug classification and defined daily doses. This information may be insufficient to fully assist prescription, but can be used for analyzing a prescription dataset retrospectively. One typical use of ATC is to measure drug utilization for pharmaco-epidemiology purposes. However, most of the published studies leveraging the classification and defined daily doses features of ATC have been performed in Europe (e.g., [6-8]).

ATC was recently integrated in RxNorm. While RxNorm only integrates the terminological features of ATC, and not its defined daily doses and routes of administration, this integration already facilitates the analysis of prescription lists indexed with RxNorm identifiers, by providing a reliable entry point into ATC.

The objective of this study is to assess the suitability of the ATC/DDD Index (Anatomical Therapeutic Chemical (ATC) Classification System/Defined Daily Dose) for analyzing prescription lists in the U.S. More specifically, we propose to analyze drug classification based on ATC groupings and to compare the prescribed daily dose to the defined daily dose in ATC for a large prescription dataset from Surescripts. To our knowledge, this study is the first application of ATC to the analysis of a U.S. prescription dataset.

2. Background

This investigation leverages ATC, RxNorm and a large prescription list obtained from Surescripts.

2.1. ATC

The Anatomical Therapeutic Chemical (ATC) classification [http://www.whocc.no/atc_ddd_index/], a system developed by the World Health Organization (WHO) Collaborating Centre for Drug Statistics Methodology, is recommended for worldwide use to compile drug utilization statistics. The system includes drug classifications at 5 levels; anatomical, therapeutic, pharmacological, chemical and drugs or ingredients. Also included are defined daily doses (DDDs) and administration routes assigned to most drugs in accordance to the therapeutic and pharmacological groups.

For example, as shown in Figure 1, drugs from the "Digitalis glycosides" 4th-level group are included in the anatomical group "Cardiovascular system". The route of administration (Adm.R) and the defined daily dose (DDD) are listed for each of the four 5th-level drugs.

C CARDIOVASCULAR SYSTEM
C01 CARDIAC THERAPY
C01A CARDIAC GLYCOSIDES
C01AA Digitalis glycosides

ATC code	Name	DDD	U	Adm.R	Note
C01AA01	acetyldigitoxin	0.2	mg	O	
C01AA02	acetyldigoxin	0.5	mg	O	
C01AA03	digitalis leaves	0.1	g	O	
C01AA04	digitoxin	0.1	mg	O	
		0.1	mg	P	
C01AA05	digoxin	0.25	mg	O	
		0.25	mg	P	

Figure 1. Drugs from the Digitalis glycosides 4th-level group in ATC (partial screenshot from the ATC website)

The active ingredients in the classification include a wide range of chemical entities used in a variety of countries. New ingredients are not included in the ATC system until they are approved for pharmaceutical use in at least one country. Only herbal medicinal products approved by regulatory authorities are included in the classification.

The active moieties are classified according to the main therapeutic use of the main ingredient. Since an ingredient can have therapeutic applications on different anatomical sections, ATC assigns a different code to the same ingredient in different anatomical sections. For example, the beta-blocker *timolol* has different codes when used as a cardiovascular drug (*C07AA06*) and as a treatment for glaucoma (*S01ED01*).

The defined daily dose (DDD) is the assumed average maintenance dose per day for a drug used for its main indication in adults. The DDD is calculated based on adult weight of 70 kg. The DDD can be an average of doses from different countries and might the reflection of the more commonly used strengths. The DDD is not necessarily the prescribed daily dose, as the latter depends on individual patient characteristics such as age, weight and pharmacokinetic considerations. Topical products, sera, vaccines, antineoplastic agents, allergen extracts, anesthetics and contrast media are not assigned a DDD.

The 2014 edition of ATC used in this study contains 4580 5th-level ATC drugs, of which 3904 correspond to single-active moieties (as opposed to combinations).

2.2. RxNorm

RxNorm is a standardized nomenclature for clinical drugs compiled from 12 drug source vocabularies, and maintained by the National Library of Medicine (NLM). A clinical drug is defined as a pharmaceutical product with therapeutic or diagnostic properties available to patients. A clinical drug includes the ingredient(s), strength or concentration, and dose form appropriate for the intended administration route (e.g., *Thyroglobulin 32 MG Oral Tablet*). The February 2014 edition of RxNorm is used in this study.

Base ingredients are the active moieties of clinical drugs (e.g., *amoxicillin*). RxNorm also covers their various salts, esters and complexes (e.g., *amoxicillin trihydrate*), referred to as "precise ingredients" in RxNorm parlance. Unlike RxNorm, ATC represents mostly base ingredients and does not distinguish between base and precise ingredients. Single-ingredient drugs have a unique chemical component, and multiple-ingredient drugs have two or more (e.g., *Amoxicillin 250 MG / Clavulanate 125 MG Oral Tablet*). While drug combinations are precisely defined in RxNorm, they are often unspecified in ATC (e.g., *meprobamate, combinations*).

Dose forms are administration vehicles, such as pills, tablets, syringes and lotions. Dose form groups (DFGs) are grouping of dose forms (DFs). For example, the DFG *Oral Product* includes the DFs *Oral Tablet, Oral Capsule, Chewable Tablet*, etc. RxNorm DFGs roughly correspond to administration routes in ATC.

RxNorm identifies a subset of drugs intended to be an approximation of the prescription drugs currently marketed in the U.S. We refer to this subset as the "prescribable subset" of RxNorm drugs, and use it as to restrict our analysis to the most clinically significant drugs.

Ingredient-level mapping between RxNorm and ATC. Since August 2013, ATC is a source vocabulary in RxNorm, which provides a mapping between RxNorm ingredients and 5th-level ATC drugs. Of the 3166 RxNorm single ingredients (base and precise), 1552 (49%) mapped to a 5th-level ATC drug, corresponding to 1991 ATC codes and 1554 distinct ATC drug names. These ingredients in common represent 51% of the 3904 5th-level ATC drugs, ignoring drug combinations.

Since the scopes of RxNorm and ATC are slightly different, the mapping is not expected to be complete. For example, RxNorm includes several hundred allergenic extracts (e.g., *papaya allergenic extract 50 MG/ML Injectable Solution*) that are out of the scope of ATC. Conversely, diagnostic and therapeutic radiopharmaceuticals (e.g., *technetium (^{99m}Tc) bicisate*) are present in ATC (under *V09* and *V10*), but out of the scope of the prescribable subset of RxNorm.

In contrast to RxNorm, in which each ingredient is represented only once, ATC can have multiple codes for the same active moiety, depending on the anatomical system or therapeutic domain in which it is used. As a consequence, there will often be multiple ATC mappings for a given RxNorm ingredient. For example, the RxNorm ingredient *Ketoconazole* (*6135*) maps to the following ATC codes for this drug: *D01AC08* (from the *ANTIFUNGALS FOR DERMATOLOGICAL USE* group), *G01AF11* (from the *GYNECOLOGICAL ANTIINFECTIVES AND ANTISEPTICS* group) and *J02AB02* (from the *ANTIMYCOTICS FOR SYSTEMIC USE* group).

2.3. Surescripts dataset

The prescription drug list is a de-identified list comprised of 102,709 clinical drugs dispensed to emergency room patients over a period of three months in 2011 at Suburban Hospital in Bethesda, Maryland. Each drug includes an anonymized prescription identifier, clinical drug name, drug form, strength, prescribed amount, and the intake duration. This prescription list was annotated with RxNorm identifiers for clinical drugs. When updated against the February 2014 version of RxNorm, 99,576 drugs were valid (or could be mapped to a valid code), while 3133 were obsolete. Of these, we only investigate the 87,001 drug codes corresponding to single-ingredient drugs from the prescribable subset of RxNorm.

2.4. Related work

Many studies have been published reporting on drug utilization based on ATC for various classes of drugs, including antibiotics [9, 10], cardiovascular drugs [8], and anti-depressants [11], or across classes [7]. Some studies specifically compare prescribed daily doses to defined daily doses in ATC for anti-epileptic drugs [6] and for several classes of anti-hypertensive drugs [12]. One characteristic of most of these studies is that they were performed in Europe, where ATC is more widely used than in the U.S.

More recently, ATC has also been used as a terminological reference for drugs. For example, ATC has been used to support the detection of adverse events in the EU-ADR project [13]. Additionally, ATC has been used as a reference in research projects where drug classes were predicted by integrating chemical-chemical interactions and similarities [14] or through text mining [15]. In earlier work, we compared and contrasted ATC with the National Drug File-Reference Terminology (NDF-RT) developed by the U.S. Department of Veterans Affairs (VA) Veterans Health Administration [16].

The specific contribution of our work is the application of ATC in combination with RxNorm, the standard drug vocabulary in the U.S. While many pharmaco-epidemiology studies leveraging ATC have been published in Europe, to the best of our knowledge, this study is the first analysis of a prescription dataset in the U.S. with ATC and RxNorm.

3. Methods

In ATC, a 5th-level code is assigned not to an ingredient, but to an ingredient for a specific therapeutic intent. For example, the beta-blocker *timolol* can be used orally or parenterally as a cardiovascular drug (*C07AA06*) and in eye drops as a treatment for glaucoma (*S01ED01*). Moreover, a defined daily dose (DDD) is assigned not to an ingredient, but to an ingredient with a specific route of administration. For example, the DDD for *acetylsalicylic*

acid is 3 g for oral forms, but 1 g when administered parenterally. As a consequence, for the purpose of finding the DDD, the mapping of RxNorm clinical drugs to ATC requires a match for both the ingredient and the route of administration. Our approach to comparing the prescribed daily dose to the defined daily dose is depicted in Figure 2. While the ingredient-level mapping is provided by RxNorm, we had to create a mapping for the routes of administration in order to relate RxNorm clinical drugs to their appropriate ATC 5th-level code. We then mapped clinical drugs from the prescription dataset to RxNorm and computed the prescribed daily dose for comparison to the corresponding defined daily dose in ATC.

In this investigation, we restrict the scope of the mapping to single-ingredient clinical drugs in RxNorm and ATC, because combination drugs are often underspecified in ATC. For example, the ATC 5th-level code *N05BC51* corresponds to *meprobamate, combinations*, and is distinct from the single-ingredient category for *meprobamate* (*N05BC01*), but without specifying which ingredients can be associated with *meprobamate* or what the DDD for *meprobamate* is in this case. Moreover, since our goal is to analyze prescription lists, we restrict the mapping to clinical drugs from the prescribable subset of RxNorm. Finally, since the prescribed daily dose is not explicitly mentioned in the Surescripts data, we further restrict the comparison of daily doses to oral solid dose forms of clinical drugs, for which we can rely on RxNorm to extract the quantity per prescription dose (i.e., pill).

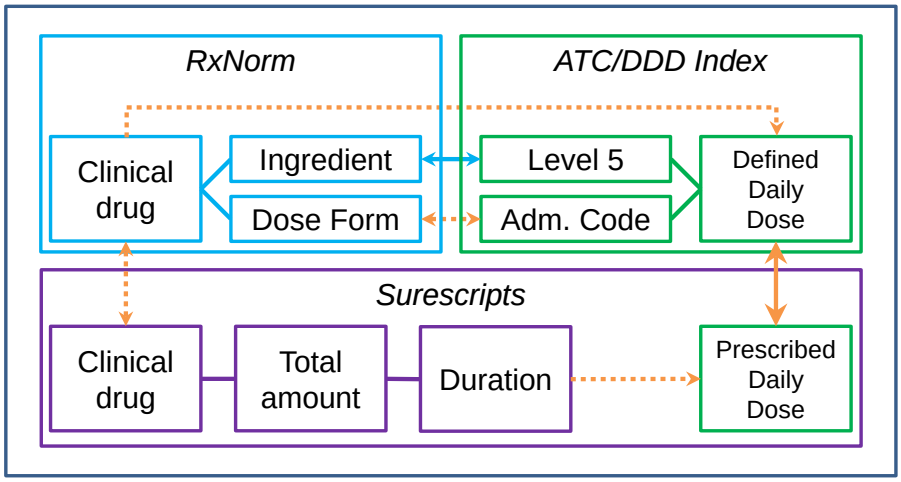


Figure 2. Overview of the methods for comparing the prescribed daily dose to the defined daily dose

3.1. Mapping RxNorm clinical drugs to ATC

In order to support our use cases of drug classification and assessment of prescribed daily doses, the mapping of RxNorm clinical drugs to ATC must account for both the ingredient and the route of administration. More specifically, we define a mapping between a clinical drug in RxNorm and an ATC 5th-level drug when the following two conditions are met.

1. The ingredient (active moiety or salt ingredient) of the clinical drug in RxNorm maps a 5th-level drug in ATC. We use the ingredient-level mapping provided in RxNorm.
2. The dose form group for the clinical drug in RxNorm and the administration code (or one of the administration codes, if multiple) of the 5th-level drug in ATC are compatible. (i.e., are associated through the same administration route), as defined below.

For example, as illustrated in Figure 3, the RxNorm clinical drug *Amoxicillin 25 MG/ML Oral Suspension* (*313797*) maps to the ATC code *J01CA04* (*amoxicillin*), because the ingredient of the RxNorm clinical drug, *amoxicillin*, maps to this ATC code, and the dose form group of the RxNorm clinical drug, *Oral Product*, matches one of the routes of administration for the ATC code *J01CA04*, *O*, through the administration route *oral*. In contrast, despite the fact that both drugs have the same ingredient, *butoconazole*, we failed to map *5000 MG Butoconazole nitrate 20 MG/ML Prefilled Applicator* (*890780*) to *G01AF15*, because the dose form group of RxNorm drug, *Prefilled Applicator Product*, is not listed as compatible with the vaginal route, *V*, listed for this drug in ATC. Finally, some

RxNorm drugs have no mapping to ATC because their ingredient is simply not present in ATC (e.g., *oregano allergenic extract 50 MG/ML Injectable Solution*).

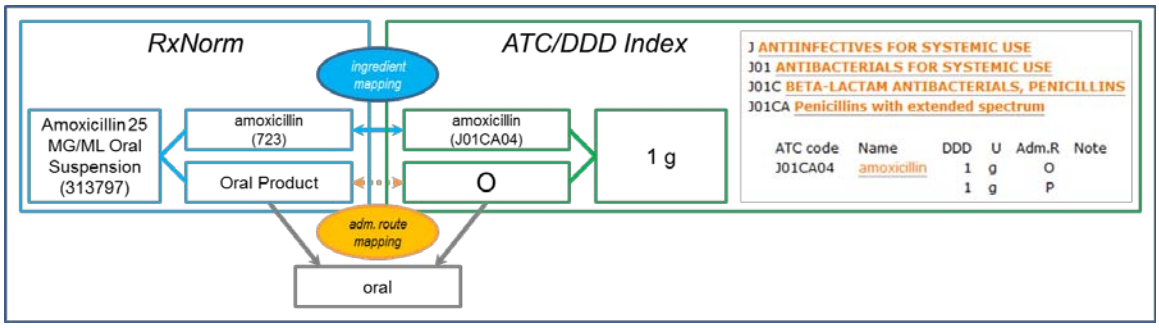


Figure 3. Mapping between RxNorm and ATC through both the ingredient and the route of administration

Mapping routes of administration between RxNorm and ATC. In order to map routes of administration between RxNorm and ATC, we harmonized the dose form groups in RxNorm and the administration codes in ATC. We also assigned administration codes to ATC drugs when they were missing.

Harmonization of administration codes between RxNorm and ATC. RxNorm and ATC have different ways of representing routes of administration. In RxNorm, the route is expressed through the dose form group (DFG), but RxNorm DFGs actually represent the dose form (e.g., *Pill*), the route (e.g., *Ophthalmic Product*) or a mix of both (e.g., *Oral Gel Product*). Many clinical drugs are associated with multiple DFGs, typically one for the dose form and one for the route. For example, *Ketoconazole 200 MG Oral Tablet* (*197853*) is associated with both *Pill* and *Oral Product*. Of the 45 DFGs in RxNorm, 22 represent dose forms exclusively, but some of them are indicative of topical products nonetheless (e.g. *Shampoo Product*).

ATC assigns administration codes to the drugs in scope for the defined daily dose (e.g., *O* for oral, *N* for nasal, etc.). In addition to the 22 administration codes, ATC defines 10 coarser administration routes. Although ATC does not provide a correspondence between administration codes and administration routes, this correspondence is usually trivial to establish. Missing from the list of ATC administration routes are entries for the routes of ophthalmologic, otic, stomatologic and other topical products, for which ATC typically does not provide DDDs.

We extended the list of 10 administration routes from ATC with *ophthalmologic*, *otic*, *stomatologic* and *topical*, adding *urethral*, as it exists as an administration code. We mapped all relevant DFGs from RxNorm to the extended list of 15 administration routes derived from ATC.

Assignment of missing administration codes in ATC. As mentioned earlier, one issue for mapping drugs between RxNorm and ATC is that ATC assigns administration codes only to a subset of its drug entities, as required for the DDD. In practice, drugs for which no DDD is asserted are also missing an administration code. These drugs typically include topical products and systemic drugs for which there are large inter-individual dose variations (sera, vaccines, antineoplastic agents, allergen extracts, general and local anesthetics and contrast media).

For these drugs, we used various strategies to semi-automatically infer an administration code when it was missing. More specifically, we manually created rules to assign administration codes to sets of drugs based on various characteristics, so that missing administration codes could be automatically inferred at the level of individual drugs. These rules were applied in the following order.

1. The authors assigned one or more administration code manually, not to individual drugs, but to specific ATC groups (at various levels), based on clinical knowledge (e.g., the group *Enemas* (*A06AG*) was associated with the administration route *rectal*).
2. The authors assigned one or more administration code manually based on specific expressions found in the labels of ATC groups (e.g., ATC drugs from groups, whose label contain the expression “systemic” were associated with the administration routes *oral* and *parenteral*).

3. All administration codes found in the drugs for a given 4th-level ATC group were propagated to the drugs with missing administration codes in the same group (e.g., the drug *alogliptin* (*A10BH04*) "inherits" the administration code *oral* from the other drugs in the group *Dipeptidyl peptidase 4 (DPP-4) inhibitors* (*A10BH*), namely *sitagliptin*, *vildagliptin*, *saxagliptin*, and *linagliptin*).
4. The administration code *oral* was assigned by default to any digestive drug (e.g., the digestive drug *tilactase* (*A09AA04*) was assigned the administration code *oral*).
5. The administration code *topical* was assigned by default to any drug that has not been assigned one in the previous steps (e.g., the drug *tetracycline* (*D06AA04*) was assigned the administration code *topical*).

3.2. Analysis of the Surescripts prescription dataset

The Surescripts prescription dataset is coded to RxNorm clinical drugs and we use the mapping to ATC in order to be able to classify the prescription drugs with ATC groups and to compare the prescribed daily doses to the defined daily doses listed in ATC.

Assessing coverage. The proportion of RxNorm clinical drugs from the Surescripts dataset to which we can associate a specific ATC 5th-level code (i.e., accounting for both the ingredient and the route of administration) assesses the coverage of RxNorm clinical drugs in ATC.

Classifying prescription drugs. Through the mapping to ATC we extract the ATC classification of the drugs for characterizing the prescription list. For example, the RxNorm clinical drug *sitagliptin 50 MG Oral Tablet* (*665042*) maps to the 5th-level ATC drug *A10BH01*, classified under the diabetes drugs in the ATC level-1 group *A*.

Assessing prescribed daily doses. We also compare the prescribed daily doses to the defined daily doses listed in ATC. In the Surescripts dataset, the prescribed daily dose is not explicitly provided. From the total number of prescription doses and duration of the prescription, we can calculate the number of prescription doses for a day. We then use RxNorm to get the quantity in each clinical drug. For example, a prescription of 45 doses *Clonazepam 0.5 MG Oral Tablet* for 30 days yields a prescribed daily dose of .75 mg (.5 * 45 / 30). Of note, for comparability between drugs, RxNorm normalizes the strength of solutions per milliliter, of inhalers per "puff" for metered-dose inhalers, and of topical creams and gels to mg/mg. As a consequence, the normalized quantity reflected in RxNorm often does not correspond to the prescribed dose. For this reason, we restrict the analysis of prescribed daily dose to oral solid drug form drugs from the Surescripts dataset. We also ignore from the dataset RxNorm drugs for which ATC provides more than one DDD for a given route of administration.

3.3. Implementation

From a technical perspective, this investigation can be thought of as a data integration project. The datasets to be integrated include RxNorm, ATC, the ingredient mapping between RxNorm and ATC, and the mapping of both RxNorm dose form groups and ATC administration codes to administration routes, as well as the prescription dataset. Semantic Web technologies are known to provide support for data integration. Here we converted all the datasets to the Resource Description Format (RDF triples) and loaded them into the triple store Virtuoso. The query language for RDF, SPARQL, also provides support for writing production rules (of the "if … then" type). We created production rules in order to infer the missing administration codes. We also created rules to infer the mapping of clinical drugs to ATC. Finally, we queried the integrated dataset in order to export the prescribed and defined daily doses for each prescription for statistical analysis.

4. Results

4.1. Mapping RxNorm clinical drugs to ATC

Of the 11,422 single-ingredient clinical drugs from the prescribable subset of RxNorm, 7748 (68%) had an ingredient mapping to ATC, and 7260 (64%) had both an ingredient and an administration route mapping. In other words, a mapping between a clinical drug in RxNorm and a drug in ATC (at the 5th level) for a particular administration route was found for 64% of the clinical drugs in RxNorm.

These RxNorm clinical drugs mapped to 1912 unique ATC codes (96% of the 1991 ATC codes to which an ingredient mapping was found) and 1479 unique drug names (95% of the 1554 ATC drug names to which an

ingredient mapping was found), corresponding to 49% of the 3904 ATC codes for single active moieties (and 44% of the drug names).

Harmonization of administration codes between RxNorm and ATC. The correspondence between RxNorm DFGs, ATC administration codes and the extended administration routes is shown in Table 1. Each of the 22 dose form groups from RxNorm and each of the 24 administration codes from ATC (including the four codes we created) is mapped to one of the 15 administration routes (extended list). As a result, each dose form group from RxNorm can be associated with at least one administration code from ATC.

Table 1. Correspondence between RxNorm dose form groups and ATC administration codes through then extended list of administration routes derived from ATC.

RxNorm Dose Form Group	Route of administration	ATC Administration Code
Drug Implant Product	*implant*	implant s.c. implant
Inhalant Product	*inhalation*	Inhal Inhal.solution Inhal.powder Inh.aerosol Inhal.aerosol Instill.sol.
Nasal Product	*nasal*	N
Oral Product	*oral*	O Chewing gum oral aerosol
Ophthalmic Product	*ophthalmic*	lamella [ophthalmic] *
Otic Product	*otic*	[otic] *
Injectable Product	*parenteral*	P
Rectal Product	*rectal*	R
Sublingual Product	*sublingual/buccal*	SL
Buccal Product Dental Product Oral Cream Product Oral Foam Product Oral Gel Product Oral Ointment Product Oral Paste Product	*stomatologic*	[stomatologic] *
Transdermal Product	*transdermal*	TD TD patch
Intraperitoneal Product Irrigation Product Mucosal Product Prefilled Applicator Product Shampoo Product Soap Product Topical Product	*topical*	intravesical ointment [topical] *
Urethral Product	*urethral*	urethral
Vaginal Product	*vaginal*	V

** added to the original ATC administration codes for mapping purposes*

Assignment of missing administration codes in ATC

Of the 3904 5th-level codes in ATC for single active moieties, 2059 (53%) are missing an administration code. The distribution of the number of ATC codes for which administration codes were generated is listed in Table 2, by type of technique. Since the rules for ophthalmic, otic, stomatologic and rectal products were allowed to generate administration codes even when one had been asserted by ATC, the total number of ATC drugs for which administration codes were generated is slightly higher than the number of ATC codes with missing administration codes.

Table 2. Number of ATC codes for which administration codes were generated automatically, by type of technique.

Expression in ATC group label	#
Administration code inferred from ATC group	725
Adminisstration code inferred from expressions found in the labels of ATC	232
Administration code inferred from drugs from the same ATC group	492
Oral admistration code inferred by default (digestive drugs)	23
Topical administration code inferred by default (remaining drugs)	643
Total	2115

4.2. Analysis of the Surescripts prescription dataset

Assessing coverage. Of the 87,001 RxNorm codes from this Surescripts dataset (restricted to single-ingredient drugs from the prescribable subset of RxNorm), 84,380 (97%) mapped to at least one code in ATC (through both the ingredient and the route). Moreover, of the 1695 distinct RxNorm clinical drugs found in the Surescripts dataset, 1606 (95%) were found in ATC.

Classifying prescription drugs. Using the mapping to ATC, we classified the 84,380 prescriptions from the Surescripts set against the top-level categories in ATC, resulting into 86,578 ATC codes. The distribution of Surescripts drugs by top-level ATC groups is shown in Figure 4. The top categories are cardiovascular and nervous system drugs. Of note, some RxNorm clinical drugs map to more than one code in ATC (e.g., drugs with multiple therapeutic uses for the same route of administration, such as *clonidine hydrochloride 0.3 MG Oral Tablet*, used orally as both as an antihypertensive drug (*C02AC01*) and an antimigraine agent (*N02CX02*)).

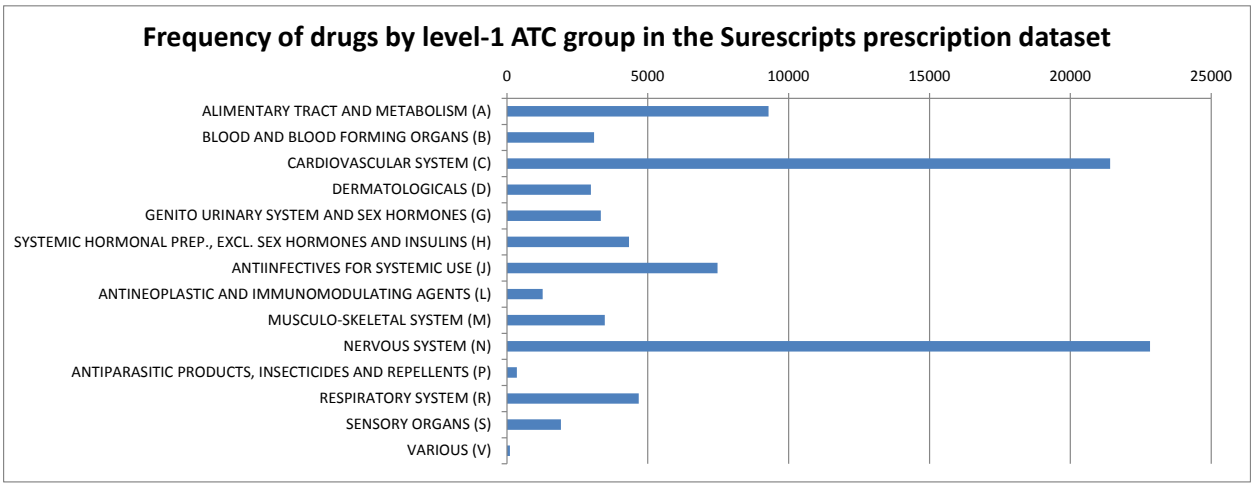


Figure 4. Distribution of Surescripts drugs by top-level ATC groups

Assessing prescribed daily doses. Of the 72,360 RxNorm clinical drugs corresponding to oral solid dose forms in the Surescripts dataset, 70,394 (97%) could be associated with a defined daily dose in ATC, of which 1932 were associated with more than one DDD (and were ignored from the comparison). For the remaining 68,462 drugs, we compared the prescribed and defined daily doses. The distribution of the ratios of the prescribed daily doses (PDDs) to the defined daily doses (DDDs) is plotted in Figure 5 (using a logarithmic scale, because of the amplitude of the variation among the ratios). Overall, the PDD exactly matches the DDD in 28.6% of the prescriptions. The ratio is in a 66%-150% range for 49.5% of the prescriptions, in a 50%-200% range for 76.1%, and in a 33%-300% range for 86.1%. Only 3.4% of the PDDs are beyond 300% of the DDD and 10.4% below 33% of the DDD. The proportions covered by each range are shown in Figure 5.

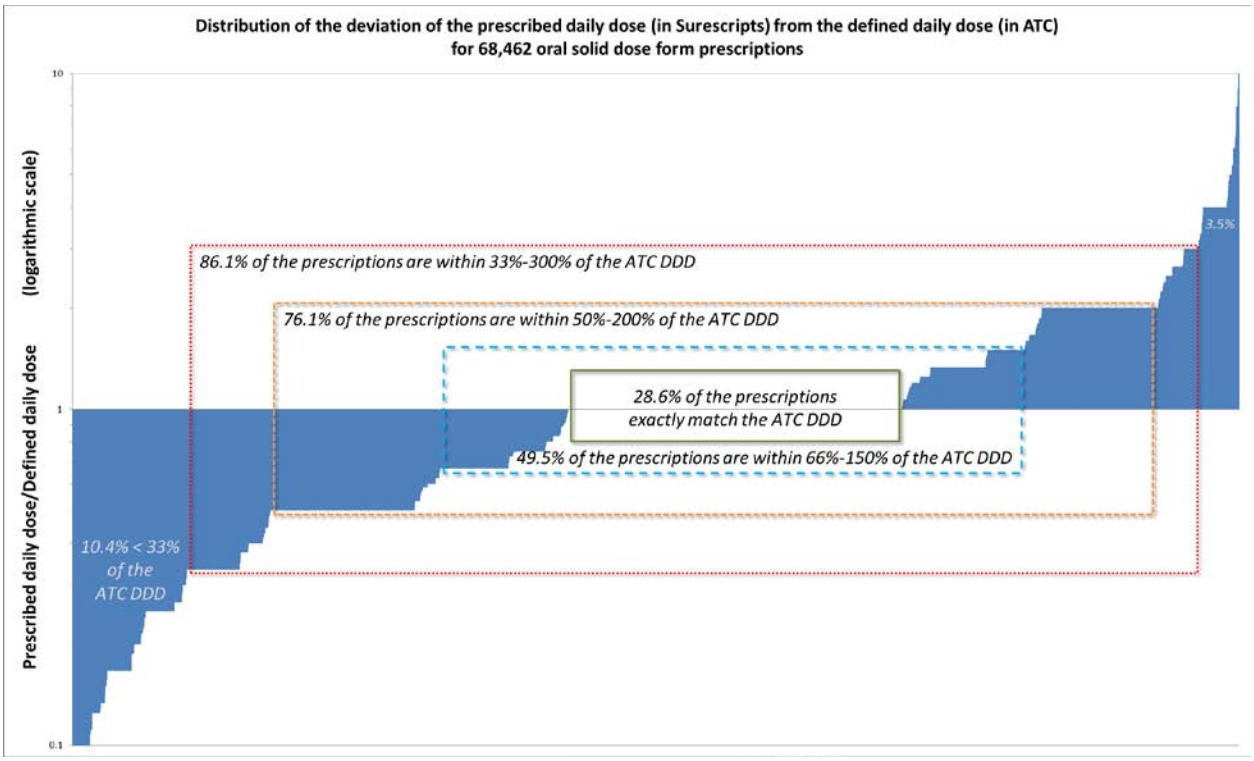


Figure 5. Distribution of the deviation of the prescribed daily dose from the defined daily dose

5. Discussion

Significance. Although the overall coverage for ingredient mapping is limited, we were able to demonstrate that prescription drugs in current use in the U.S. are mapped reliably and in a considerable proportion. This study confirms that the use of ATC in conjunction of RxNorm is a valid strategy for analyzing prescription datasets in the U.S., both from the perspective of classifying drugs and for the comparison of prescribed and defined daily doses. While ATC is routinely used in Europe for pharmaco-epidemiology, our study is the first application of ATC to prescription data in the U.S.

Limitations. The main limitation of our work is the limited size and scope of the prescription dataset, in which the variation of drug ingredients is necessarily limited, even more so in the case of drugs from emergency room patients only. While the proportion of clinical drugs mapped to ATC may be smaller in other datasets, the method of mapping to ATC through RxNorm should be generally applicable, including when drugs are represented with codes from the National Drug Code (NDC) or not coded at all. Another limitation is that the analysis of the prescribed daily doses was restricted to oral solid dose forms, because the prescribed dose was not explicitly mentioned in the Surescripts dataset and could not be reliably extracted from RxNorm. In fact, this issue is being addressed in RxNorm by creating different entities for solution with identical (normalized) concentrations, but different quantities per volume (e.g., 1 mg/ml and 5mg/5ml).

Future work. In addition to the exploration of larger and more diverse prescription datasets, the focus of future work is to refine the administration route assignment for missing routes in the ATC, allowing for more precise mapping. We also would like to combine two aspects of the current work, i.e., drug classification and deviation from the defined daily dose, in order to investigate whether certain classes of drugs tend to be prescribed at higher or lower doses compared to the defined daily dose.

6. Conclusions

Although the mapping of RxNorm ingredients to ATC appears to be largely incomplete, the most frequently prescribed drugs in the prescription dataset we analyzed were covered. This study demonstrates the feasibility of using ATC in conjunction with RxNorm for analyzing U.S. prescription datasets for drug classification and assessment of the prescribed daily doses.

Acknowledgments

This work was supported by the Intramural Research Program of the NIH, National Library of Medicine.

References

1. Medicine Io. To Err is Human: Building a Safer Health System. Washington, DC: National Academies Press; 2000
2. Clancy CM. Ten years after To Err is Human. Am J Med Qual 2009;24(6):525-8.
3. Keers RN, Williams SD, Cooke J, Ashcroft DM. Prevalence and nature of medication administration errors in health care settings: a systematic review of direct observational evidence. Ann Pharmacother 2013;47(2):237-56.
4. Lesar TS. Tenfold medication dose prescribing errors. Ann Pharmacother 2002;36(12):1833-9.
5. Kaushal R, Shojania KG, Bates DW. Effects of computerized physician order entry and clinical decision support systems on medication safety: a systematic review. Arch Intern Med 2003;163(12):1409-16.
6. Koristkova B, Grundmann M, Brozmanova H. Differences between prescribed daily doses and defined daily doses of antiepileptics--therapeutic drug monitoring as a marker of the quality of the treatment. Int J Clin Pharmacol Ther 2006;44(9):438-42.
7. Ravera S, Hummel SA, Stolk P, Heerdink RE, de Jong-van den Berg LT, de Gier JJ. The use of driving impairing medicines: a European survey. Eur J Clin Pharmacol 2009;65(11):1139-47.
8. Stimac D, Polic-Vizintin M, Skes M, Cattunar A, Cerovic R, Stojanovic D. Utilization of cardiovascular drugs in Zagreb 2001-2005. Acta Cardiol 2010;65(2):193-201.
9. Bozkurt F, Kaya S, Tekin R, Gulsun S, Deveci O, Dayan S, et al. Analysis of antimicrobial consumption and cost in a teaching hospital. J Infect Public Health 2013
10. Xu J, Duan X, Wu H, Zhou Q. Surveillance and correlation of antimicrobial usage and resistance of Pseudomonas aeruginosa: a hospital population-based study. PLoS One 2013;8(11):e78604.
11. Viola R, Benko R, Nagy G, Soos G. National trend of antidepressant consumption and its impact on suicide rate in Hungary. Pharmacoepidemiol Drug Saf 2008;17(4):401-5.
12. Grimmsmann T, Himmel W. Discrepancies between prescribed and defined daily doses: a matter of patients or drug classes? Eur J Clin Pharmacol 2011;67(8):847-54.
13. Avillach P, Dufour JC, Diallo G, Salvo F, Joubert M, Thiessard F, et al. Design and validation of an automated method to detect known adverse drug reactions in MEDLINE: a contribution from the EU-ADR project. Journal of the American Medical Informatics Association : JAMIA 2012
14. Chen L, Zeng WM, Cai YD, Feng KY, Chou KC. Predicting Anatomical Therapeutic Chemical (ATC) classification of drugs by integrating chemical-chemical interactions and similarities. PloS one 2012;7(4):e35254.
15. Gurulingappa H, Kolarik C, Hofmann-Apitius M, Fluck J. Concept-based semi-automatic classification of drugs. Journal of chemical information and modeling 2009;49(8):1986-92.
16. Winnenburg R, Bodenreider O. Exploring pharmacoepidemiologic groupings of drugs from a clinical perspective. Stud Health Technol Inform (Proc Medinfo) 2013;192:827-831.

Data Quality and Interoperability Challenges for eHealth Exchange Participants: Observations from the Department of Veterans Affairs' Virtual Lifetime Electronic Record Health Pilot Phase

Nathan Botts, PhD[1], Omar Bouhaddou, PhD[2], Jamie Bennett[2], Eric Pan, MD, MSc[1], Colene Byrne, PhD[1], Lauren Mercincavage, MHS[1], Lois Olinger, MA[1], Elaine Hunolt, MHSA[2], Theresa Cullen, MD, MS[2]
[1]Westat, Rockville, MD; [2]US Department of Veterans Affairs, Washington, DC

Abstract

Authors studied the United States (U.S.) Department of Veterans Affairs' (VA) Virtual Lifetime Electronic Record (VLER) Health pilot phase relative to two attributes of data quality - the adoption of eHealth Exchange data standards, and clinical content exchanged. The VLER Health pilot was an early effort in testing implementation of eHealth Exchange standards and technology. Testing included evaluation of exchange data from the VLER Health pilot sites partners: VA, U.S. Department of Defense (DoD), and private sector health care organizations. Domains assessed data quality and interoperability as it relates to: 1) conformance with data standards related to the underlying structure of C32 Summary Documents (C32) produced by eHealth Exchange partners; and 2) the types of C32 clinical content exchanged. This analysis identified several standards non-conformance issues in sample C32 files and informed further discourse on the methods needed to effectively monitor Health Information Exchange (HIE) data content and standards conformance.

Introduction

Through the Virtual Lifetime Electronic Record (VLER) Health Exchange initiative, the United States (U.S.) Department of Veteran Affairs (VA) can electronically share parts of Veterans' health records with providers at the U.S. Department of Defense (DoD), and participating private sector health care organizations (exchange partners). As a direct benefit to Veterans, doctors involved with their care would have a more comprehensive and timely record of health information, including services received by the Veteran outside of their purview. VLER Health Exchange leverages the policies and technical standards of the eHealth Exchange (formerly the Nationwide Health Information Network or NwHIN) supported by the non-profit, public-private collaborative called Healtheway [1, 2]. Goals include better informed care providers, improved continuity and timeliness of care, enhanced awareness among all parties, and elimination of gaps in a patient's health record.

VA engaged in a pilot of VLER Health Exchange to comprehensively test and improve health data standards for effective exchange of Veteran health information across the eHealth Exchange [3, 4]; to establish models for nationwide health information exchange (HIE); to identify scalable implementation strategies; and to assess early impacts of VLER Health Exchange. VA selected 12 sites with a strong business case for HIE and sought a diversity of characteristics, such as geographic factors (e.g., rural, urban), populations served, the maturity of the HIE organization, and their sustainability models. Four pilot sites participated in a three-way exchange between VA, DoD, and the private sector, and eight sites participated in two-way exchange between VA and the private sector. The pilot period concluded in October 2012. A more in-depth description of the VLER Health initiative and lessons from the pilot phase are reported elsewhere [5]. After satisfactory completion of the pilot phase, Veterans Health Information Exchange has been expanded nationally, connecting to more and more private sector partners and enabling clinicians at all VA Medical Centers to see externally sourced data.

As the policies and technical systems that serve as the foundation for nationwide HIE continue to mature, and the amount of health data and documents being shared across the eHealth Exchange increases, greater attention is now focused on the quality of data exchanged [6]. Data quality issues impact the way in which the

data reach the intended recipients, are realized in the user interface and, for many user systems, incorporated by the Electronic Health Record (EHR). The exchange of health information between the numerous products and versions of EHR systems requires a standardized method for communicating data that is agreed upon and adopted by entities seeking to share data. During the VLER Health pilot, eHealth Exchange Partners shared data through use of a suite of data transport tools and services called CONNECT. The primary health data standards used to structure the exchanged data included the Health Information Technology Standards Panel (HITSP) C32 and C62 document standards [7].

C32 documents use the Health Level 7 (HL7) Continuity of Care Document (CCD) component to describe the content of a health summary to be created, exchanged, and to summarize a patient's medical status. The content may include administrative (e.g., demographics, insurance) and clinical information (e.g., problem list, medication list, allergies and test results) [8]. C32 content standards are comprised of 17 content modules that represent the underlying clinical data in both narrative and structured forms. In addition to C32 structured documents, C62 documents can be used to incorporate unstructured clinical notes and scanned documents (e.g., text file, PDFs, or images such as a scanned image of an electrocardiogram) [9].

The technical hurdles of matching patient records, exchanging data from different EHR systems and geographies and then properly rendering the data in a manner that can be effectively used by health care providers are significant. Within the data standards themselves, including the C32, issues of optionality and interpretation create differences in the way the standards are implemented and the way data are mapped across systems. Consequently, even the most diligent HIE development and implementation can produce challenges that hinder effective health data exchange.

The VLER Health pilots and HIE initiatives that preceded VLER Health established an important baseline of understanding as it relates to the adoption and implementation of eHealth Exchange health data standards, and the availability and quality of clinical content of shared Veterans within pilot regions [10, 11, 12].

Methods

Effective sharing of Veteran health information across the eHealth Exchange requires export and import of a validated C32 document with health information. Our study methods established a baseline for two attributes of data quality: 1) compliance of exchange partner's C32 to current data standards; and 2) the clinical content being provided within VLER Health documents exchanged. Addressed in the first are issues of data quality that occur at the structural level of the C32 exchanged by partners. The second evaluates the types of content, or richness, of information exchanged.

Validation of C32 Data Standards Compliance

All eHealth Exchange partners must comply with the current eHealth Exchange standards. The eHealth Exchange onboarding process during the pilot phase required exchange partners to be able to produce a compliant C32, but did not include formal compliance and content testing. The National Institute of Standards and Technology (NIST) Clinical Document Architecture (CDA) validator was used to validate potential eHealth Exchange partner C32's for conformance to the HITSP/C32 v2.5 standard. The NIST CDA validator tests the underlying Extensible Mark-up Language (XML) within the C32 to determine whether the schema and data provided conform to the requirements established by the HITSP/C32 v2.5 specification.

For this study, the NIST CDA validation application was downloaded from the NIST website, installed locally and configured with the libraries necessary to check the validity of C32 documents per v2.5 specifications. The HITSP/C32 specification consists of an evolving hierarchy of standards for electronic documentation of health information, including services received by a patient (e.g., Continuity of Care Record (CCR), HL7 CDA, CCD, HITSP C32).

The NIST CDA validator reports the types of non-conformances found in each related section and classifies non-conformance alerts with the current C32 standard into levels of severity that include errors (items of non-conformance), warnings (items that technically conform, but could be better constructed), and notes (general comments and suggestions on implementation). Should a part of the C32 being tested not conform, a report is provided that outlines where and why an error occurred according to the specification (Figure 1).

These errors in the underlying XML of the C32 when realized on the screen can produce gaps in the information and reduces the quality of the data experienced by the care provider (Figure 2).

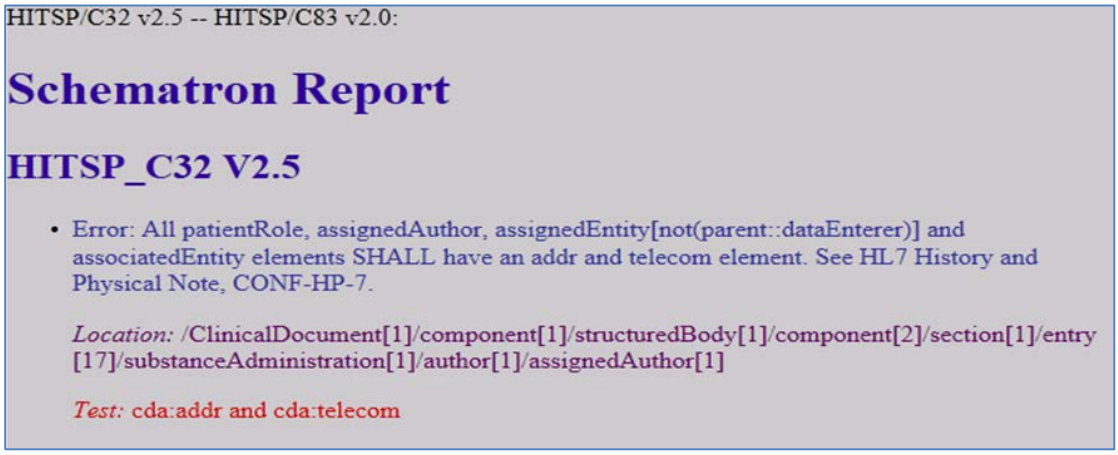


Figure 1. Example of NIST CDA Validation Schematron Report of a sample C32 document with errors

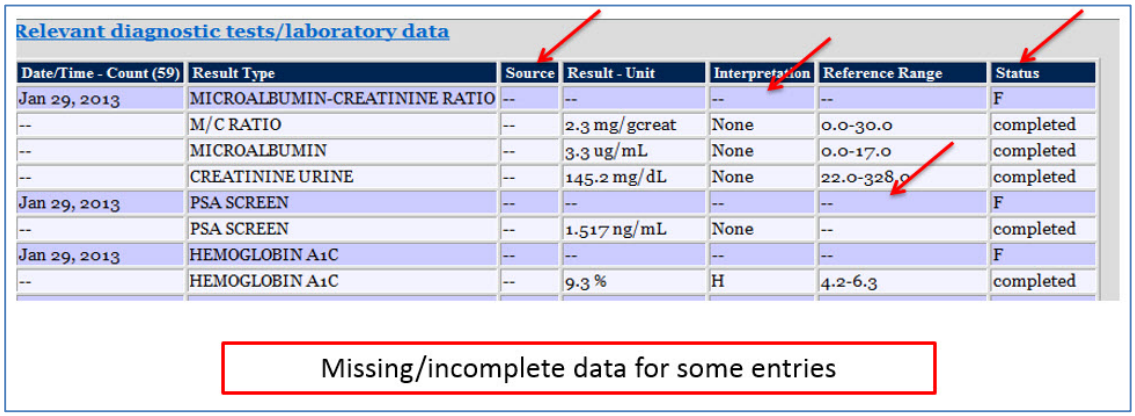


Figure 2. Example of errors in the C32 XML structure when realized in VA VistAWeb EHR

A sample of fourteen populated C32s provided between October 2011 and July 2012 were tested for their conformance to the current C32 standard - one from each of the 12 pilot site private exchange partners' EHRs and one from the VA and DoD systems. Test results recorded when non-conformance to the specification was found and where (e.g., CDA, CCD), the types of non-conformance identified (e.g., missing data element or attribute), and the sections in which the non-conformances were found (e.g., header, problems/conditions, medications). This assessment only considered those classified by the NIST CDA validator as potential errors in the C32 structure.

Evaluation of C32 Data Content Availability

The clinical content of the C32 was also assessed to better understand the type and amount of clinical data available to and from VA providers during the VLER Health pilot. In July 2012, tests of VLER Health pilot clinical content were conducted by VA as a part of an operations and quality assurance initiative. VA partner C32s were analyzed based on data available between October 2011 and June 2012.

Based on a random sample of 250 Veterans with records available for exchange for each VLER Health pilot site partner, C32s were queried using VistAWeb, and the types of content available were analyzed (e.g., medications, laboratory results, procedures). Eight of the twelve pilot site exchange partners, plus VA, were included in the study, providing a total of 2,250 Veteran records for analyses. Partners that were not included in this sub-study were either not in production or lacked a sufficient number of matched patients at the time the study was conducted.

For each retrieved C32 document, the populated modules were recorded. The study measures consisted of: 1) whether a C32 was returned for each Veteran; 2) for retrieved C32s, whether the Veteran had any eHealth Exchange clinical data available in their populated C32; and 3) if so, the types of clinical modules were populated. Testers only examined whether each C32 clinical module (e.g., medications) contained data, but did not review the C32 for other attributes of the data content such as completeness, display issues or data quality.

Results

VLER Health Exchange Partner C32 Validation

Validation analysis of eHealth Exchange partner C32s indicated that six of the VLER Health exchange partners produced conformant validation results. The other eight C32s tested, however, resulted in some level of error being reported when run through the NIST CDA validator. Two of the C32s tested produced over 10 unique errors of non-conformance to the standard (often the same error is repeated multiple times depending on the content present in the C32). Issues encountered were primarily related to undefined attributes or XML pattern errors, problems found within the document header, and missing data elements or required values as defined by the C32 standard.

It is important to note that the majority of issues identified would not necessarily impact the way in which clinical content was reported. Many issues were due to administrative attributes of the clinical document (e.g., proper inclusion of a country code) that are important to broad health information exchange, but may not be critical to a patient's treatment. A person viewing this information in an EHR may not even perceive the impact of these types of errors, but as established by the HITSP C32, these specifications are deemed important and mandatory for proper inclusion, and consequently may result in incompatibilities and errors when shared among C32-compliant software.

Outlined in Table 1 are the categories, definitions, and error types for 103 issues identified across eight of the 14 C32s analyzed. As noted previously, depending on how many records are present in the C32, these issues can be reproduced many times, but represent only one main issue in how the C32 is constructed. The frequency of issues is important in terms of the user experience when viewing the document on their screen.

Table 1. Validation Issues Identified in VLER Health Partner C32 Samples

Issue Categories	Definitions	Percent of total issues (n=103)	Prevalence
Non-conformant Attribute or Pattern	An XML attribute or pattern that is unrecognizable by the rules and requirements provided by the C32 schema	10%	Found at least once in 5 partner C32s, with over 200 different instances found
General Header Constraints	Important details regarding the origins and author of the record may be missing or improperly described	58%	Found at least once in 8 partner C32s, with over 300 different instances found
Missing Elements or Required Values	Required data elements and/or coding missing or improperly represented in the C32	32%	Found at least once in 8 partner C32s, with over 200 different instances found
Source: Based on 14 sample C32s, one from each of the VLER Health exchange partners provided between October 2011 and July 2012.			

Non-conformance issues were most commonly found in the medication, insurance, laboratory result, allergy, and problem modules. This fact may only be because these were the modules reported most frequently for the C32s tested, or because information for some clinical modules are not sent at all by a particular eHealth Exchange partner at the time of the study.

Types of Clinical Content eHealth Exchange Partners are Capable of Sharing

Of the nine eHealth Exchange partners whose C32s were analyzed for clinical content (eight private partners and VA), the majority of C32s (88 percent) included clinical content in addition to basic demographic data. Table 2 provides a breakdown of clinical content data availability among VLER Health exchange partners by C32 module, the average of C32s that were populated with data for that module, and the range of results across exchange partners.

Table 2. Data Availability for VLER Health Partners by C32 Module

C32 Module	Percent of C32s Populated	Range Populated by Partner
Demographics	100%	100%
Providers	86%	71% ~ 99%
Problems	84%	52% ~ 99%
Allergies	74%	13% ~ 99%
Encounters	74%	35% ~ 94%
Medications	63%	<1% ~ 97%
Vital Signs	53%	1% ~ 80%
Laboratory Results	46%	9% ~ 81%
Procedures	42%	9% ~ 61%
Immunizations	35%	2% ~ 63%

Source: Based on a sample of 250 C32s for nine exchange partners (including VA) at the time of the evaluation, plus VA, retrieved July 2012. (n=2,250)

Note: Tests were conducted in July 2012 as a VLER Health quality assurance/operations study. The disclosures are for the period 10/13/2011 - 7/25/2012.

Given the patient matching processes needed to exchange clinical data, basic demographics information was available in 100 percent of the C32s retrieved. This was expected given that demographics information is required for patient matching and exchange of a C32, which were part of the inclusion criteria for this analysis. Provider information was available in 86 percent of the records. Problem list, allergy list, encounters, medications, and vital signs were also available in more than half of the C32s tested. When analyzing the C32s by VLER Health Partners, problem list, allergy list, and medications remain the most commonly available data module; three VLER Health Partners sent these data for more than 75 percent of their Veterans, and another four VLER Health Partners sent these data for at least half of their Veterans.

Subgroup analysis comparing VLER Health Partners shows that problems and allergies were populated the most frequently (seven out of nine partners), followed by medications and laboratory test results (6 out of 9 partners), followed by list of encounters and list of procedures (5 out of 9 partners). Half of VLER Health Partners were able to send at least eight clinical modules, and nearly all VLER Health Partners could send at least five clinical modules, including problem list, allergy list, medication, laboratory test results, and procedures data.

Discussion

The HL7 Consolidated CDA (C-CDA) is the current standard promoted by Meaningful Use 2014, however at the time of the VLER Health pilot the HITSP C32 summary of care record was elected to be the template used to exchange health information through a national-level HIE. The C32 had yet to be implemented and tested in a large-scale, geographically dispersed, production-level exchange. By implementing the C32 standard in a live patient-care environment with multiple exchange partners through the VLER Health program, VA experienced how "optionality" in the specifications can lead to differences of interpretation, uncertainties of implementation, and incompatibilities among compliant software. The NIST CDA validator identified many non-conformance issues in sample C32 files. However, the issues are not reflective of unpracticed adoption of the standards; rather are reflective of the optionality and flexibility in interpretation and implementation built into the standard. These data quality issues related to the implementation of health information standards can have significant implications in effective rendering of the data when sharing data between numerous partners and systems [13].

Although the NIST CDA validation system can help ensure compliance with the C32 specification, and therefore greater compatibility with the receiver's system, it did not yet assist with assessing clinical validity of this information. A small number of C32s were subsequently manually inspected to determine whether they conformed to HITSP specified standard terminologies. Use of clinical coding terminologies (e.g., Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT), Logical Observation Identifiers Names and Codes (LOINC)) helps to ensure semantic interoperability (i.e., meanings of terms and codes translate across systems) and can enhance quality and safety of decision making by clinicians. Results from this informal study suggest that use of clinical coding terminologies was relatively low (less than 30 percent of the terminology coding requirements were met), but further study of the use of data standards and terminologies by provider EHRs would provide more information about this attribute of data quality.

The study of clinical content availability only assessed whether a C32 data domain was present or not and did not assess the data structure, display, or clinical validity aspects that would further inform the usability of the data. As a part of general operations, VLER Health pilot sites compiled clinical content related quality issues (e.g., duplication of data, inconsistencies in formatting and display) based on anecdotes and observations by clinicians and VLER Health staff members when retrieving VLER Health data. The issues identified by VLER Health staff would likely go undetected by the NIST validation system and included problems related to the quality of the data (e.g., little or no clinical data available in the record), inconsistencies in the way the data were presented (e.g., the name of the medication not listed together with the medication allergy, excessive abbreviations, incorrect terminology mapping), and incomplete details that reduced the value of the information (e.g., missing reference ranges for laboratory test results). Further study at this last mile of HIE usability in which the data exchanged is acted upon by the care provider or system who receives the information is critical to understanding future requirements and priorities in health information standards development. VA has expanded its initial data quality assessment activities to include ongoing production validation and scoring of partner CCDs, including the above mentioned data quality aspects.

While significant challenges were encountered in the exchange of VLER Health data during the VLER Health pilot, it is also important to note that eHealth Exchange Partners were reliably providing a core set of valuable clinical content to VA clinicians. Tests confirmed that for sites that are in production and stable, eHealth Exchange data is retrievable for most VA correlated Veterans, and further, that clinical data are available for 91 percent of these correlated Veterans through eHealth Exchange. The VLER Health pilot has supported the VA to better understand the requirements necessary to effectively exchange health information across the eHealth Exchange and to develop a long-term strategy toward better interoperability of content. C32 standard conformance issues, clinical content availability, and the data presentation issues identified during the VLER

Health pilot helped inform VA's work with the Office of the National Coordinator (ONC) Standards and Interoperability Framework in shaping a more robust and interoperable C-CDA standard.

There are two noteworthy efforts to mention that will help construction and testing of more valid and interoperable content among eHealth Exchange partners. The first is the formal requirement that the new eHealth Exchange onboarding body has adopted for content testing. Healtheway guidance explains to implementers what content is required beyond the C32 to reach a meaningful exchange and how this content should be populated and will be tested. For instance, with HITSP, only the Person Information and the Source sections were required, whereas the Healtheway Bridge C32 will require Allergies, Medications, Problems, and Laboratory Results to also be populated [14]. This is further promoted by Meaningful Use Certification 2014. The second is the opportunity to consider alternative technical ways to implement structured document creation. For example, Model-Driven Health Tools (MDHT) is a project that develops structured document creation and evaluation tools that are programmatically derived and enforced from models of the specifications [15].

This study identified the need for improved models of validation and testing of exchange partner data content. Current NIST validation tools include validation guidelines for addressing Meaningful Use requirements for compliant CCDs. As the number of providers exchanging data across the eHealth Exchange increases, "turn-key" tools for validating clinical content for exchange will need to be devised and include strategies for ongoing monitoring and validation of clinical content and standards implementation. Study results have implications related to the complexity of health data interoperability, the evolution of health data exchange standards and its significance to policies such as those embodied in Meaningful Use [16]. In order to effectively comply with Meaningful Use 2014 Certification requirements and beyond, strict rules on validation are needed, and more consistent methods of testing should be established for providers and HIE organizations [17].

Conclusion

The VLER Health pilot served as a live, large-scale, production-level test of eHealth Exchange standards used to exchange clinical data nationally. VA and its partners experienced challenges in validating standards conformance, populating C32 documents, and presenting the data to providers in a consistently accurate, usable fashion. VA continues to work with DoD, Healtheway, ONC, and other stakeholders to leverage the VLER Health pilot experience to create better standards and tools needed by the nation as it moves beyond the barriers of connectedness in health care to the challenges of data integrity, quality, and usability.

References

1. Healtheway. eHealth Exchange. *Healtheway*. March 2014. Available at: http://healthewayinc.org/index.php/exchange. Accessed February 2014.

2. Bouhaddou O, Cromwell T, Davis M, et al. Translating standards into practice: Experience and lessons learned at the Department of Veterans Affairs. *Journal of Biomedical Informatics*. 2012;45(4):813-823.

3. Saef S. The impact of a health information exchange on resource use and Medicare-allowable charges at eleven emergency departments operated by four major hospital systems in a midsized southeastern city: an observational study using clinical estimates. *Annals of Internal Medicine*. 2013;43(4):S97.

4. Ross SE, Radcliff TA, Leblanc WG, Dickinson LM, Libby AM, Nease DEJ. Effects of health information exchange adoption on ambulatory testing rates. *Journal of the American Medical Informatics Association*. 2013;20(6):1137-42.

5. Byrne CM, Mercincavage LM, Bouhaddou O, et al. The Department of Veterans Affairs' (VA) implementation of the Virtual Lifetime Electronic Record (VLER): Findings and lessons learned from health information exchange at 12 sites. *International Journal of Medical Informatics*. August 2014;83(8):537-547.

6. Kuperman GJ, Blair JS, Franck RA, Deveraj S, Low AF, NHIN Core Services Content Workgroup. Developing data content specifications for the nationwide health information network trial implementations. *Journal of the American Medical Informatics Association*. 2013;17(1):6-12.

7. Bouhaddou O, Bennett J, Teal J, et al. Toward a Virtual Lifetime Electronic Record: The Department of Veterans Affairs Experience with the Nationwide Health Information Network. Paper presented at: AMIA 2012 Annual Symposium, 2012; Chicago, IL.

8. Healthcare Information Technology Standards Panel. *C 32 2009A- HITSP summary documents using HL7 continuity of care document (CCD) component* 2009.

9. Healthcare Information Technology Standards Panel. *C62 2009b - unstructured document component* 2009.

10. Frisse ME, King JK, Rice WB, et al. A regional health information exchange: Architecture and implementation. Paper presented at: AMIA Annual Symposium Proceedings, 2008; Washington DC.

11. Biondich PG, Grannis SJ. The Indiana network for patient care: an integrated clinical information system informed by over thirty years of experience. *Journal of Public Health Management and Practice*. November 2004:S81-6.

12. Halamka J, Overhage JM, Riccardi L, Rishel W, Shirky C, Diamond C. Exchanging health information: local distribution, national coordination. *Health Affairs*. 2005;24(5):1170-1179.

13. Williams C, Mostashari F, Mertz K, Hogin E, Atwal P. From the Office of the National Coordinator: The strategy for advancing the exchange of health information. *Health Affairs*. Mar 2012;31(3):527-36.

14. Healtheway. Official eHealth Exchange Technical Errata and Change Log. *Healtheway*. March 2014. Available at: http://healthewayinc.org/index.php/resources/exchange-specifications?highlight=WyJicmlkZ2UiLCJjMzIiLCJicmlkZ2UgYzMyIl0. Accessed March 2014.

15. Model Driven Health Tools. MDHT Project Home. *Open Health Tools*. 2013. Available at: https://www.projects.openhealthtools.org/sf/projects/mdht/.

16. Marcotte L, Seidman J, Trudel K, et al. Achieving meaningful use of health information technology: A guide for physicians to the EHR incentive programs. *Archives of Internal Medicine*. 2012;172(6):731-736.

17. D'Amore JD, Mandel JC, Kreda DA, et al. Are Meaningful Use Stage 2 certified EHRs ready for interoperability? Findings from the SMART C-CDA Collaborative. *Journal of the American Medical Informatics Association*. 2014;Epub ahead of print.

Health information technology: use it well, or don't! Findings from the use of a decision support system for breast cancer management

Jacques Bouaud, PhD[1,2,3,4], Brigitte Blaszka-Jaulerry, MD[5], Laurent Zelek, MD[6], Jean-Philippe Spano, MD, PhD[7,8], Jean-Pierre Lefranc, MD[7,9], Isabelle Cojean-Zelek, MD[10], Axel Durieux, MD[11], Christophe Tournigand, MD, PhD[12], Alexandra Rousseau, PhD[13], Brigitte Séroussi, MD, PhD[3,2,4,14]

[1] **AP-HP, DRCD, Paris, France**
[2] **INSERM, U1142, LIMICS, Paris, France**
[3] **Sorbonne Universités, UPMC Univ Paris 06, UMR_S 1142, LIMICS, Paris, France**
[4] **Université Paris 13, Sorbonne Paris Cité, LIMICS, (UMR_S 1142), Bobigny, France**
[5] **CH Lagny Marne la Vallée, Service de radiothérapie-oncologie, Lagny, France**
[6] **Université Paris 13, Sorbonne Paris Cité, UFR SMBH, Bobigny, France ; AP-HP, Hôpital Avicenne, Service d'oncologie médicale, Bobigny, France**
[7] **Sorbonne Universités, UPMC Univ Paris 06, UFR de médecine, Paris, France**
[8] **AP-HP, Hôpital Pitié-Salpétrière, Service d'oncologie médicale, Paris, France**
[9] **AP-HP, Hôpital Pitié-Salpétrière, Service de chirurgie et cancérologie gynécologique et mammaire, Paris, France**
[10] **Hôpital des Diaconesses, Pôle oncologie médicale, Paris, France**
[11] **Institut de Cancérologie des Peupliers, Paris, France**
[12] **AP-HP, Hôpital St-Antoine, Service d'oncologie médicale, Paris, France**
[13] **AP-HP, Hôpital St-Antoine, URC-EST, Paris, France**
[14] **AP-HP, Hôpital Tenon, Département de santé publique, Paris, France ; APREC, Paris, France**

Abstract

The potential of health information technology is hampered by new types of errors which impact is not totally assessed. OncoDoc2 is a decision support system designed to support treatment decisions of multidisciplinary meetings (MDMs) for breast cancer patients. We evaluated how the way the system was used had an impact on MDM decision compliance with clinical practice guidelines. We distinguished "correct navigations" (N+), "incorrect navigations" (N–), and "missing navigations" (N0), according to the quality of data entry when using OncoDoc2. We collected 557 MDM decisions from three hospitals of Paris area (France) where OncoDoc2 was routinely used. We observed 33.9% N+, 36.8% N–, and 29.3% N0. The compliance rate was significantly different according to the quality of navigations, 94.2%, 80.0%, and 90.2% for N+, N–, and N0 respectively. Surprinsingly, it was better not to use the system (N0) than to use it improperly (N–).

Introduction

Health information technology (HIT) is expected to improve the quality of care, reduce medication errors and adverse events, optimize overall health care utilization, and decrease costs. Electronic health records (EHRs), computerized provider order entry systems (CPOEs), and clinical decision support systems (CDSSs) are among the tools which can make the delivery of care safer, more effective and more efficient.[1] Whatever support these tools may provide, their relevance to healthcare and for healthcare professionals relies on the correct characterization of actual patients in EHRs and more widely in data repositories.

However, concerns about the quality of recorded patient data have been reported.[2] EHRs have become less reliable than ever before, flawed by errors introduced by medical personnel, patients, and machines.[3] Studies suggest HIT introduces unpredicted and unintended adverse consequences (UACs) potentially harmful for patients. Campbell et

al.[4] have identified 9 types of UACs resulting from CPOE implementation, the major of them being the Type 7 category (new kinds of errors) called "e-iatrogenesis" and defined by Weiner et al.[5] as "patient harm caused at least in part by the application of health information technology".

A US poll among health-care leaders revealed that HIT safety is the hazard of greatest concern for 2013.[6] Of those polled, 89% rated HIT patient/data mismatches in EHRs and HIT systems as a 9 or 10 on a scale of 1 to 10 as a hazard of great concern. In another more recent study, the most often safety event reported is that "Data is incomplete, missing or misleading" in EHRs (52%).* Patient data may thus be missing, incorrectly entered, badly displayed, and erroneously transmitted which may have serious safety consequences at the individual level. Apart from "technoskeptics", some authors consider that technology-related errors are due to the inappropriate use of HIT tools. For instance, Sittig and Singh[7] reported that UACs of HIT occur when HIT is unavailable for use, malfunctioning during use, used the wrong way, or when HIT interacts inadequately with another system component. Numerous patient safety initiatives for HIT are thus carried out to handle the issue of technology-induced errors.[8,9] Gaining knowledge on how to avoid misuse and promote correct usage becomes a requirement when implementing HIT tools to further secure their routine use.

Cancer management is subject to variable practices and to varied levels of compliance with oncology practice guidelines.[10,11] In the last decade, multidisciplinary meetings (MDMs), or tumour boards, have become a standard of practice to improve therapeutic decision quality of cancer patients. The underlying hypothesis is that the management of cancer is by nature pluridisciplinary, and that treatment plans collectively established by cancer specialists (surgeons, oncologists, radiologists, pathologists, radiotherapists, etc.) are better than those proposed by one specialist deciding alone. If studies on cancer care generally associate MDMs with improvements of guideline compliance rates,[12,13,14] the efficiency of MDMs has been recently questioned[15,16] and the debate is still open, suggesting there should be opportunities for decision support in MDMs.[17]

Many studies have indeed shown that clinical decision-support systems (CDSSs) might be effective tools to promote clinical practice guideline (CPG) compliance.[18] We have developed a guideline-based CDSS called OncoDoc2[19] to assist decision-making for non-metastatic breast cancer according to local (CancerEst) CPGs. OncoDoc2 has been routinely used for three years in the breast cancer MDM of a single hospital, showing very good results in terms of compliance rate.[20] More recently, a prospective cluster multicentered randomized controlled trial (RCT)† has been carried on to assess the impact of OncoDoc2 on the compliance of MDM decisions with CPGs. Six hospitals were randomized to separate the three hospitals of the "intervention arm" where OncoDoc2 had to be used during MDMs to decide the care plan of breast cancer patients, and the three hospitals of the "control arm" where MDMs proceeded as usual (OncoDoc2 was not used). This paper does neither discuss the methods nor present the results of the RCT, which will be presented in a different paper. The objective here is to focus on the use of OncoDoc2, and only analyze the intervention group according to the study design. In a previous work with the same context,[21] we studied the accuracy of actual data entry. When using OncoDoc2, we distinguished three different situations: "correct navigations" (*N+*), "incorrect navigations" (*N–*), and "missing navigations" (*N0*). The aim of this work is to evaluate how the way OncoDoc2 was used has an impact on the quality of MDM decisions assessed by their compliance with CPGs.

Material and Method

The OncoDoc2 CDSS. OncoDoc2[19] is a computerized guideline-based CDSS. It provides patient-specific recommendations for the management of any primary breast cancer (in situ and invasive) according to CancerEst (local) CPGs. The knowledge base is structured as a decision tree where nodes represent patient criteria, arcs represent values of criteria, and paths represent theoretical patient profiles. The system can be automatically run from patient data extracted from EHRs. It can also be used according to the document-based paradigm of decision-making where the decision tree is interactively browsed by a user. Thus, using OncoDoc2 consists in answering, by a simple click on the right value, a sequence of closed-ended questions, displayed in the interface, to finally access recommended care plans when the hypertextual navigation is completed and a leaf is reached. The user's navigation through the knowledge base, denoted N, instantiates criteria according to patient-specific data, which selects the theoretical clinical profile that best fits the actual patient, $N = \{criteria = value\}^*$. Figure 1 shows a screenshot of OncoDoc2 displaying a

*http://www.healthit.gov/sites/default/files/onc_safer_jan302014_ppt.pdf.

†This study has been supported and funded by Assistance Publique - Hôpitaux de Paris, France (# K 070603).

navigation, characterized by a path of the decision tree ("Recapitulative"), and the corresponding recommendations ("CancerEst Recommended Treatment Plans"). If the MDM decision is one of the proposed recommendations, then the decision is compliant with the CPGs encoded in the system.

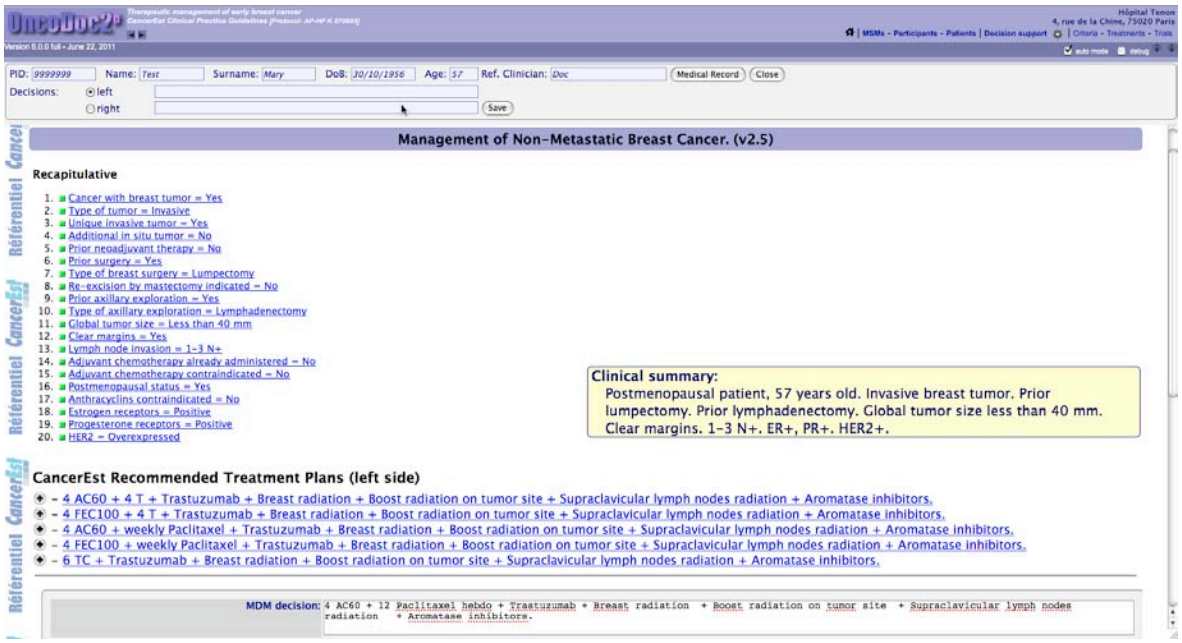


Figure 1. Screenshot of OncoDoc2: display of the patient profile selected by the navigation with clinical summary and corresponding recommendations.

Study background. The MDM is a common organizational feature of cancer care practice.[17] In France, it is one of the measures launched by the first National Cancer Program (2003-2007) and renewed in the second (2009-2013) and third (2014-2019) Programs. Multidisciplinary decision making have become mandatory for all cancer patients.

The six hospitals of the RCT are located in the eastern area of Paris (France). They all organize weekly breast cancer MDMs. In the three hospitals randomized to the intervention arm, MDM physicians had to use OncoDoc2 for all breast cancer patients. Navigations were either performed by one physician among MDM participants or by the MDM secretary. All were informed of the study and of its objectives, and trained to the use of OncoDoc2 before the study started. Navigations were performed while the clinical case was orally presented to MDM physicians. The display of OncoDoc2's interface was shown on a large screen so that all participants could follow the navigation performed and read the system's recommendations. In each hospital H_i, and for each discussed case P_{i_j}, the MDM navigation, N_{ij}^{MDM}, as well as the final therapeutic MDM decision d_{ij} were recorded in OncoDoc2 database.

Data collection. Each week and in each of the six hospitals of the RCT, clinical research assistants (CRAs) collected the data of all the patient cases discussed in MDMs the previous week. None of the hospitals were equipped with a full EHR. Thus, CRAs were given the list of the patients discussed along with their paper-based medical records. CRAs knew which hospitals were in the control arm, resp. the intervention arm, of the RCT, but they didn't know what were the navigations performed by MDMs of the intervention arm, and for each patient case of both arms, they blindly performed their own navigation on the basis of the information they found in paper-based records. CRAs had been previously clinically trained. In both arms of the RCT, MDM decision compliance was computed by the comparison of MDM decisions and the recommendations provided by CRA-performed navigations. For all MDM decisions computed as non-compliant, CRA navigations were reviewed, data were checked back in medical records,

and possibly corrected, by a domain expert. Thus, for each patient case P_{i_j}, the CRA-performed final navigation was considered as the "gold standard" and named "reference navigation" noted $N_{i_j}^{Ref}$. The set of recommendations attached to the reference navigation has been considered as the reference recommendations that apply to the patient and noted $\{R\}_{i_j}^{Ref}$. We stated that a MDM decision d_{i_j} was compliant with CPGs when it belonged to the set of recommendations of the reference navigation, *i.e.* $d_{i_j} \in \{R\}_{i_j}^{Ref}$. In this way, CRA data collection and analysis process was the same in both the intervention and the control arms.

Data analysis. We only considered for the analysis the data from the intervention arm, *i.e.* when OncoDoc2 was used. For each patient case P_{i_j} discussed during a MDM, we compared the MDM navigation $N_{i_j}^{MDM}$ with the reference navigation $N_{i_j}^{Ref}$. Depending on data entry quality when using OncoDoc2, we distinguished three types of decisions:

- Decisions made with "correct navigations" (*N+*): MDM navigations were identical to reference navigations.
- Decisions made with "incorrect navigations" (*N–*): MDM navigations differed from reference navigations. At least one criteria instantiation of the MDM navigation was different from its value in the reference navigation, and considered as a false data entry or an error.
- Decisions made with "missing navigations" (*N0*): Some MDM decisions didn't have a corresponding MDM navigation in OncoDoc2 database. In this case, the system might not have been used and MDMs did not complete the intervention as required by the protocol.

We first studied how comparable were the three samples of decisions using statistical monovariate analyses on patient criteria. We then analyzed the effect of data entry quality on the compliance of MDM decisions with CPGs. The special case of *N–* navigations has been analyzed. In particular, we studied the distribution of the data error rate among the decision criteria used by OncoDoc2. Some navigations were close enough to the reference navigation to include identical care plan recommendations: $\{R\}_{i_j}^{MDM} \cap \{R\}_{i_j}^{Ref} \neq \emptyset$. Such situation was qualified as $Recos^+$. In such cases, the recommendations provided by OncoDoc2 did include some guideline-based care plans actually recommended for the patient although she was not correctly described by the navigation performed by the MDM. On the contrary, when there was no overlap of the care plans proposed by the two navigations, $\{R\}_{i_j}^{MDM} \cap \{R\}_{i_j}^{Ref} = \emptyset$, none of the recommendation provided by OncoDoc2 was appropriate for the patient according to CPGs and this case was noted $Recos^-$. We made a difference between these two scenarios and assessed the effect of *N–* navigations leading to either $Recos^+$ or $Recos^-$ on the compliance of MDM decisions.

Results

MDM decisions were included between June 2009 and April 2010 (11 months). In this work, we only considered decisions made when OncoDoc2 was available for use, *i.e.* in the hospitals of the intervention arm. A total of 557 decisions were collected and considered for analysis. Hospital accrual was respectively of 73, 155, and 329 decisions for the 3 hospitals, H_1, H_2, and H_3.

Navigation types, data accuracy, and effect on decision compliance. We collected 189 decisions *N+* (33.9%), 205 *N–* (36.8%), and 163 *N0* (29.3%). In the subset of 394 (189 + 205) decisions where MDM navigations were available, we observed 52.0% of incorrect uses *N–*. When considering data entry of MDM users of OncoDoc2, 6,025 criteria assignments were performed, 15.3 per decision in average. A total of 252 data entry errors were identified in incorrect navigations *N–*, yielding an overall error rate of 4.2% of all clinical data entry.[21]

The overall compliance rate of the 557 decisions was measured at 87.8% in an intent-to-treat analysis, *i.e.* independently of MSMs fully following the study protocol and using the CDSS, or not, thus taking all *N0* decisions into account. Compliance rates according to the navigation type are reported in Table 1. They were measured at 94.2%, 80.0%, and 90.2% for *N+*, *N–*, and *N0* decisions respectively, and were significantly different ($p < 10^{-4}$, Chi^2, simple

descriptive statistics, no cluster effect taken into account). The compliance rate is the highest when the system is correctly used (*N+*). It is improved when the system is not used (*N0*) as compared to when it is incorrectly used (*N–*).

Table 1. Distribution of compliance rates by navigation type, and hospital accrual.

	N+	*N–*	*N0*		*Total*
	N=189 (33.9%)	*N=205 (36.8%)*	*N=163 (29.3%)*	p	*N=557*
Compliance					
Yes	178 (94.2%)	164 (80.0%)	147 (90.2%)	$< 10^{-4}$	489 (87.8%)
No	11 (5.8%)	41 (20.0%)	16 (9.8%)		68 (12.2%)
Hospital					
H1	32 (16.9%)	26 (12.6%)	15 (9.2%)	$< 10^{-14}$	73 (13.1%)
H2	79 (41.8%)	66 (32.2%)	10 (6.1%)		155 (27.8%)
H3	78 (41.2%)	113 (55.1%)	138 (84.6%)		329 (59.1%)

Table 1 also reports the distribution of navigation types among hospitals. Results show a highly significant difference between hospitals ($p < 10^{-14}$, Chi^2). H_1, H_2, and H_3 have respectively 20.6%, 6.5%, and 41.9% of *N0* decision. H_3 was the main provider of decisions (59.1%), and brought 84.6% of all missing navigations *N0*. Data entry accuracy was 55.2%, 54.5%, and 40.8% for H_1, H_2, and H_3 respectively.

Comparison of decision characteristics by navigation type. Patient profiles were analyzed and compared to identify whether some criteria could be associated with navigation types. A first analysis consisted in comparing the decision criteria used in the reference navigations of the three samples of navigation types. Sixty-three criteria were used in *N+*, 64 in *N–* and 69 in *N0*. Among all these criteria, 60 were common to the three groups but with different frequencies of usage. A few criteria were specific of a navigation type. For instance, "Type of neoadjuvant chemotherapy" has only be used once, and in a *N+* navigation. The same applies to "Suspicion of invasive cancer" only used twice in *N–* navigations. More interestingly and despite they were used no more than twice each, the following criteria were only used in *N0* decisions: "Prior radiotherapy", "Factors of bad prognosis", "Size of the largest invasive tumor larger than 2 cm", "Pleiomorphic lesion". These decision criteria do usually characterize rather complex or unusual clinical situations.

A second analysis consisted in monovariate analyses of variables characterizing the patient, the tumor, or the decision with respect to the navigation type. Table 2 lists the variables usually used to describe cancer patients along with the variables significantly associated with the navigation type. "Bilateral cancer", "Prior breast cancer", and "Contraindication to radiotherapy" were significantly more frequent in *N0* decisions, although "Contraindication to radiotherapy" is correlated with the variable "Prior breast cancer". These characteristics correspond to less frequent conditions and might characterize decisions which are more complex than usual ones. Larger tumors ("Size of the largest nodule", "Cumulative tumor size including spaces") were more frequent in *N–* navigations.

Results for incorrect navigations (*N–*). In this group, there was at least one error in the navigation's data entry. The compliance rate was the lowest, 80.0% on 205 decisions.

These incorrect uses of OncoDoc2 were also analysed according to the fact that the recommendations provided by the system for erroneous patient profiles included ($Recos+$), or not ($Recos-$), care plans recommended in the reference navigation. For the 121 $Recos+$ navigations among 205 (59.0%), the compliance rate was 90.1%, whereas it was 65.5% for the remaining 84 $Recos-$ navigations. Compliance rates in $Recos+$ and $Recos-$ were significantly different ($p < 10^{-4}$, Chi^2).

A total of 252 data entry errors were collected in *N–* navigations. Eighty percent of these navigations had only one error, 18% had exactly 2 errors and 2% had exactly 3 errors. None had more than 3 errors. On the 51 decision criteria that were used in incorrect *N–* navigations, 39 (76.5%) were recorded with at least one error in data entry. Figure 2 shows the list of criteria sorted by their data entry error rate, which ranges from 55% to 1%.

Table 2. Main patient and tumor characteristics by navigation type.

Variables		*N+* *N=189, N (%)*	*N–* *N=205, N (%)*	*N0* *N=163, N (%)*	*P-value*
Age at decision, years [mean, std]		59 ± 14	60 ± 15	59 ± 14	NS
Postmenopausal status					
	Yes	123 (65.1%)	141 (69.1%)	111 (68.1%)	NS
	No	66 (34.9%)	63 (30.9%)	52 (31.9%)	
Cancer laterality					
	Unilateral	183 (96.8%)	200 (97.6%)	146 (89.6%)	$< 10^{-3}$
	Bilateral	6 (3.2%)	5 (2.4%)	17 (10.4%)	
Prior breast cancer					
	Yes	10 (5.3%)	19 (9.3%)	28 (17.2%)	10^{-3}
	No	179 (94.7%)	186 (90.7%)	135 (82.8%)	
Invasive cancer					
	Yes	151 (83.4%)	169 (83.3%)	134 (85.4%)	NS
	No	30 (16.6%)	34 (16.7%)	23 (14.6%)	
Exclusive microinvasive cancer					
	Yes	2 (1.1%)	4 (2.0%)	3 (1.9%)	–
	No	177 (98.9%)	198 (98.0%)	154 (98.1%)	
Multifocality					
	Yes	23 (13.0%)	41 (20.6%)	24 (16.0%)	NS
	No	154 (87.0%)	158 (79.4%)	126 (84.0%)	
Decisional step					
	Pre-surgery	76 (40.2%)	88 (42.9%)	62 (38.0%)	NS
	Post-surgery	113 (59.8%)	117 (57.1%)	101 (62.0%)	
Size of the largest nodule, mm [mean, std]		21.9 ± 17.7	25.7 ± 18.5	22.7 ± 20.0	4.8×10^{-2}
Cumulative tumor size including spaces, mm [mean, std]		22.8 ± 18.8	29.4 ± 20.7	24.7 ± 21.4	1.9×10^{-3}
Contraindication to lumpectomy					
	Yes	4 (2.1%)	10 (4.9%)	14 (8.6%)	–
	No	185 (97.9%)	195 (95.1%)	149 (91.4%)	
Contraindication to axillary lymph node dissection					
	Yes	14 (7.4%)	22 (10.8%)	21 (12.9%)	NS
	No	175 (92.6%)	182 (89.2%)	142 (87.1%)	
Contraindication to radiotherapy					
	Yes	6 (3.2%)	13 (6.3%)	20 (12.3%)	3.4×10^{-3}
	No	183 (96.8%)	192 (93.7%)	143 (87.7%)	
Clear margins					
	Yes	80 (90.9%)	89 (91.8%)	71 (92.2%)	NS
	No	8 (9.1%)	8 (8.2%)	6 (7.8%)	
ER status					
	Positive	57 (80.3%)	73 (86.9%)	68 (90.7%)	NS
	Negative	14 (19.7%)	11 (13.1%)	7 (9.3%)	
PR status					
	Positive	38 (54.3%)	42 (50.0%)	44 (58.7%)	NS
	Negative	32 (45.7%)	42 (50.0%)	31 (41.3%)	
HER2 status					
	Positive	11 (15.3%)	21 (23.9%)	15 (20.8%)	NS
	Negative	61 (84.7%)	67 (76.1%)	57 (79.2%)	
Node invasion					
	Positive	29 (25.7%)	34 (29.1%)	21 (20.8%)	NS
	Negative	84 (74.3%)	83 (70.9%)	80 (79.2%)	

Note: Subtotals may differ from sample size due to missing data. Monovariate analyses were performed with Chi^2 tests for categorical data and Kruskal-Wallis tests for numerical data. Significance level chosen at 5%.

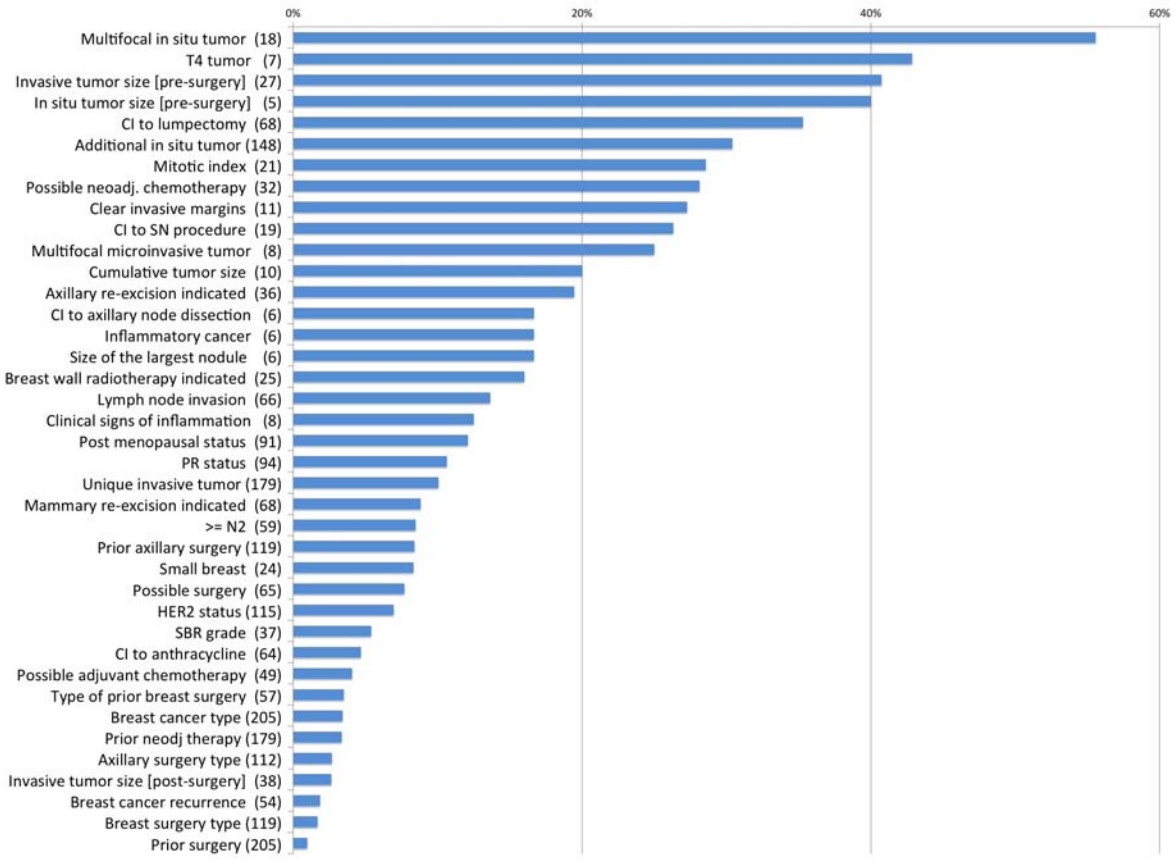


Figure 2. Decision variables involved in data entry errors of *N*– navigations, with their number of assignments and error rate.

Results for missing navigations (*N0*). For 163 decisions, there was no MDM navigation recorded in OncoDoc2 database. The compliance rate in this group, measured at 90.2%, was however higher than for incorrect navigations (*N*–), although lower than for correct navigations (*N*+).

Figure 3 displays the proportion of *N0* decisions by month of the study and by hospital. For H_1, all but one decisions were concentrated in the last 2 months where they represented 100% of MDM decisions, indicating a stop in OncoDoc2 recording for this hospital. In the other two hospitals, *N0* decisions are regularly distributed along the inclusion period, even if the proportions in H_2 and H_3 are different as previously noted.

Discussion and conclusion

Results show that the compliance rate of MDM decisions with CPGs is significantly associated with the navigation type, with compliance rates of 94.2% , 80.0%, and 90.2%, for correct (*N*+), incorrect (*N*–), and non uses (*N0*) of OncoDoc2.

We first studied whether the three samples of decisions associated with navigation types corresponded to different breast cancer patients. Many usual breast cancer variables were not significantly different among the three decision groups. However, there were significantly more patients with bilateral breast cancers or cancer recurrences in *N0* decisions and tumor size was significantly higher in *N*– cases.

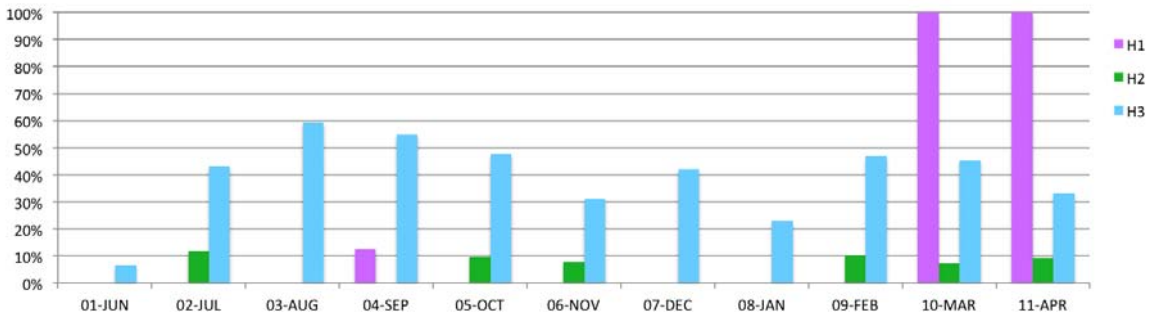


Figure 3. Temporal distributions of the proportion of missing navigations (*N0*) by hospital.

Despite we studied the intervention arm of a RCT, where OncoDoc2 should be used for all breast cancer decisions, there were 29.3% of missing navigations *N0* (163/557). If they correspond to situations where MDM participants actually did not use OncoDoc2 to decide, these cases may also be explained by technical problems (for some reasons, navigations were done but not recorded). Paradoxally, the compliance of MDM decisions when navigations were missing was quite high (90.2%), comparable to the compliance rate of *N+* decisions, and significantly higher than the compliance rate of *N–* decisions. Missing data is a ubiquitous problem in all studies of medical research and inadequate handling of missing data can have a detrimental impact on the data analysis.[22,23] Knowing the mechanism is useful in identifying the most appropriate analysis. However, the missingness mechanism is often unknown and methods for dealing with missing data such as complete case analysis (excluding patients with missing data), mean substitution (replacing missing values of a variable with the average of known values for that variable), and last observation carried forward are commonly used for analysis. Missing data is also often considered to be indicative of participant relapse, and under that assumption, many researchers have relied on setting all missing values to the worst-case scenario for the outcome. This sort of single-imputation method has been criticized for producing biased results. More recently, methods involving multiple imputation approaches have been developed.[24] In our case, H_3 hospital who is also the hospital with the highest rate of patient accrual, had 41.9% (138/329) of missing navigations. These cases may correspond to complex patient cases (bilateral breast cancer or cancer relapse) that MDM participants might have considered to be outside the scope of CPGs, thus not handled by OncoDoc2, and to be discussed on a case by case basis (no need to use the system). Other explanations such as a seasonal effect (more *N0* for H_3 in summer holydays period), or resident schedules, could be given. Concerning H_1 hospital, nearly all *N0* decisions occured in the last two months where the MDM secretary was on sick leave.

Although navigating OncoDoc2 was developed on the principle of a check-list of patient criteria easy to document, and actually necessary for the decision-making process, there was at least one error in 52% of the navigations. Most patient criteria were concerned, whether they were often ("Unique invasive tumor") or rarely ("T4 tumor") used, and quantitative ("Invasive tumor size [pre-surgery]") or qualitative ("Contraindication to lumpectomy"). When using OncoDoc2, the selection of the answer is made by simply clicking on the right value. Since the different values are displayed the some below the others, "juxtaposition errors" may explain some wrong answers. A second hypothesis could be that discordant data are not entry errors but clinical errors. Beyond clinicians' knowledge of CPGs, HIT could reveal the discordance between patient actual information and clinicians' belief of what this information is. Finally, since CRAs used the information recorded in medical records to navigate OncoDoc2, we should also consider that navigation errors correspond to errors of the medical record, corrected on the fly during the MDM navigation, but not revised in the medical record. Another interesting point is the discovery of the compliance rates of 90.1% and 65.5% in $Recos+$, resp. $Recos-$, situations of *N–* navigations. In $Recos-$ situations, MDM physicians made an incorrect navigation and got "wrong", inappropriate, recommendations, but decided to follow them. This can be interpreted as an overtrust in technology (similar to the Type 9 of UACs decribed by Campbell et al.[4], called overdependence on the technology).

This study has some limitations including the small size of the samples of both MDM decisions and hospitals, the

use of a specific CDSS OncoDoc2 including a user-controlled navigation to instantiate patient parameters, and the application to a single pathology (breast cancer management). With respect to *N–* navigations, errors resulted from the manual entry of data, since no full clinical EHR system was available in the hospitals. We don't know whether there would have been less errors if part of the data needed for OncoDoc2 navigations had come from an EHR. This is an important question and a stake for the development of EHRs and surrounding HIT tools. This however reinforces the importance of initial data entry quality into the health care digital environment. Moreover, we considered reference navigations as the gold standard and then "error-free". Thus, only non-compliant decisions were checked to assess the quality of data entry. However, there also could be errors in navigations considered to be compliant. By only evaluating one direction of errors, there is small risk of bias in the results, with a under-estimation of the error rate in compliant MDM decisions. Finally, since H_3 had a large majority of *N0* decisions (84.6%), the compliance difference could also be explained by the hospital.

We observed that the compliance rate of MDM decisions was significantly lower for *N–* navigations (80.0%), than for *N0* and *N+*, with a rate as low as 65.5% in $Recos-$ situations. Surprisingly, the compliance rate of MDM decisions with missing navigations *N0*, where we assumed that the CDSS was not used, was comparable to the one measured for correct navigations *N+*. Moreover, it was higher than the compliance rate published for breast cancer management.[11,10] This raises questions related to the context of non-using the system. Not using the CDSS for certain patient cases, whereas it is used otherwise on a regular basis, seems to be not prejudicial to decision quality. This might stem from a positive contamination effect. Then, from what was observed, it seems that it's better not to use OncoDoc2 than to use it improperly. Though not demonstrated here, similar effects might occur with other CDSSs or HIT tools, which emphasizes the importance of the correct use of HIT tools. This explains why numerous research teams are currently investigating the broad range of human factors in user-CDSS interactions. This raises the issue of usability[25] and enlightens the importance of better system development, testing, implementation, and end user training to promote the good use of HIT tools, like CDSSs, to actually benefit from HIT potential to enhance the quality and safety of health care. In this way, we could avoid the next decade to be a "dangerous decade".[1]

Acknowledgments

Authors thank URC-EST for its logistic support, AP-HP for providing the OncoDoc2 system, clinical research assistants for collecting data, all MDM members who participated to the study, and Prof. S. Uzan for his support in promoting this study.

References

1. Coiera E, Aarts J, Kulikowski C. The dangerous decade. *J Am Med Inform Assoc* 2012;19(1):2–5.
2. Benin A, Fenick A, Herrin J, Vitkauskas G, Chen J, Brandt C. How good are the data? feasible approach to validation of metrics of quality derived from an outpatient electronic health record. *Am J Med Qual* 2011;26(6):441–51.
3. Burnum J. The misinformation era: the fall of the medical record. *Ann Intern Med* 1989;110(6):482–4.
4. Campbell E, Sittig D, Ash J, Guappone K, Dykstra R. Types of unintended consequences related to computerized provider order entry. *J Am Med Inform Assoc* Sep-Oct 2006;13(5):547–56.
5. Weiner J, Kfuri T, Chan K, Fowles J. "e-iatrogenesis": the most critical unintended consequence of cpoe and other hit. *J Am Med Inform Assoc* 2007;14(3):387–8.
6. Denham C, Classen D, Swenson S, Henderson M, Zeltner T, DW B. Safe use of electronic health records and health information technology systems: trust but verify. *J Patient Saf* 2013;9(4):177–89.
7. Sittig D, Singh H. Defining health information technology-related errors: new developments since to err is human. *Arch Intern Med* 2011;171(14):1281–4.
8. Kushniruk A, Bates D, Bainbridge M, Househ M, Borycki E. National efforts to improve health information system safety in Canada, the United States of America and England. *Int J Med Inf* 2013;82(5):149–60.

9. Magrabi F, Aarts J, Nohr C, Baker M, Harrison S, Pelayo S, *et al.* A comparative review of patient safety initiatives for national health information technology. *Int J Med Inf* 2013;82(5):139–48.

10. Lebeau M, Mathoulin-Pélissier S, Bellera C, Tunon-de Lara C, Daban A, Lipinski F, *et al.* Breast cancer care compared with clinical guidelines: an observational study in france. *BMC Public Health* 2011;11:45.

11. Wöckel A, Kurzeder C, Geyer V, Novasphenny I, Wolters R, Wischnewsky M, *et al.* Effects of guideline adherence in primary breast cancer–a 5-year multi-center cohort study of 3976 patients. *Breast* 2010;19(2):120–7.

12. Kesson E, Allardice G, George W, Burns H, Morrison D. Effects of multidisciplinary team working on breast cancer survival: retrospective, comparative, interventional cohort study of 13 722 women. *BMJ* 2012;344:e2718.

13. van Hoeve J, de Munck L, Otter R, de Vries J, Siesling S. Quality improvement by implementing an integrated oncological care pathway for breast cancer patients. *Breast* 2014;23(4):364–70.

14. Brar SS, Hong NL WF. Multidisciplinary cancer care: Does it improve outcomes? *J Surg Oncol.* 2014. [Epub ahead of print].

15. Keating N, Landrum M, Lamont E, Bozeman S, Shulman L, McNeil B. Tumor boards and the quality of cancer care. *J Natl Cancer Inst* 2013;105(2):113–21.

16. El Saghir N, Keating N, Carlson R, Khoury K, Fallowfield L. Tumor boards: optimizing the structure and improving efficiency of multidisciplinary management of patients with cancer worldwide. Am Soc Clin Oncol Educ Book, 2014.

17. Patkar V, Acosta D, Davidson T, Jones A, Fox J, Keshtgar M. Cancer multidisciplinary team meetings: Evidence, challenges, and the role of clinical decision support technology. *Int J Breast Cancer* 2011;2011:831605.

18. Roshanov PS, Fernandes N, Wilczynski JM, Hemens BJ, You JJ, Handler SM, *et al.* Features of effective computerised clinical decision support systems: meta-regression of 162 randomised trials. *BMJ* 2013;346:f657.

19. Séroussi B, Bouaud J, Antoine ÉC. OncoDoc, a successful experiment of computer-supported guideline development and implementation in the treatment of breast cancer. *Artif Intell Med* 2001;22(1):43–64.

20. Séroussi B, Laouénan C, Gligorov J, Uzan S, Mentré F, Bouaud J. Which breast cancer decisions remain non-compliant with guidelines despite the use of computerized decision support? *Br J Cancer* 2013;109(5):1147–56.

21. Séroussi B, Blaszka-Jaulerry B, Zelek L, Lefranc JP, Conforti R, Spano JP, *et al.* Accuracy of clinical data entry when using a computerized decision support system: a case study with OncoDoc2. In: Mantas J, Mazzoleni C, eds, Improved Care Through Health Informatics, (vol180) of *Stud Health Technol Inform*. IOS Press, 2012:472–6.

22. Wood A, White I, Thompson S. Are missing outcome data adequately handled? a review of published randomized controlled trials in major medical journals. *Clin Trials* 2004;1(4):368–76.

23. White IR, Horton NJ, Carpenter J, Pocock SJ. Strategy for intention to treat analysis in randomised trials with missing outcome data. *BMJ* 2 2011;342.

24. Lee K, Simpson J. Introduction to multiple imputation for dealing with missing data. *Respirology* 2013. [Epub ahead of print].

25. Beuscart-Zéphir MC, Elkin P, Pelayo S, Beuscart R. The human factors engineering approach to biomedical informatics projects: state of the art, results, benefits and challenges. In: Geissbuhler A, Haux R, Kulikowski C, eds, *IMIA Yearbook of Medical Informatics 2007*. Schattauer, 2007:109–27.

Impacts of EHR Certification and Meaningful Use Implementation on an Integrated Delivery Network

Watson A. Bowes, III, MD, MS Intermountain Healthcare and Department of Biomedical Informatics, University of Utah, Salt Lake City, UT

Abstract

Three years ago Intermountain Healthcare made the decision to participate in the Medicare and Medicaid Electronic Heath Record (EHR) Incentive Program which required that hospitals and providers use a certified EHR in a meaningful way. At that time, the barriers to enhance our home grown system, and change clinician workflows were numerous and large. This paper describes the time and effort required to enhance our legacy systems in order to pass certification, including filling 47 gaps in (EHR) functionality. We also describe the processes and resources that resulted in successful changes to many clinical workflows required by clinicians to meet meaningful use requirements. In 2011 we set meaningful use targets of 75% of employed physicians and 75% of our hospitals to meet Stage 1 of meaningful use by 2013. By the end of 2013, 87% of 696 employed eligible professionals and 100% of 22 Intermountain hospitals had successfully attested for Stage 1. This paper describes documented and perceived costs to Intermountain including time, effort, resources, postponement of other projects, as well as documented and perceived benefits of attainment of meaningful use.

Introduction

In the U.S. over 89% of 5011 eligible hospitals and over 66% of 527,200 eligible professionals have received an incentive payment from the Medicare and Medicaid EHR Incentive Program, which stems from the Health Information Technology for Economic and Clinical Health Act (HITECH), and American Recovery and Reinvestment Act (ARRA) of 2009.[1] Over 20 Billion dollars of incentives have been distributed for healthcare information technology (HIT) projects to accelerate the adoption of EHRs and other technology, and to have the technology used in a meaningful way[2]. Intermountain Healthcare eligible hospitals (EH) and eligible professionals (EP) were eligible for approximately $35 million in incentives from Medicare and Medicaid for the first year of Stage 1 of the incentive program in 2013. $28M was estimated for our hospitals and $5M for our ambulatory providers.

A paper presented at the AMIA proceedings in 2011 described the decision to move forward with meaningful use and outlined the gaps and challenges that faced Intermountain Healthcare for this endeavor.[3] Because our EHR at Intermountain was self-developed, we had to certify our EHR prior to meeting meaningful use to meet the requirements outlined in the EHR Standards rule[3] and related National Institute of Standards and Testing (NIST) Approved Test Procedures for certification.[4] We identified 20 functionalities that needed no modification, but 47 requirements which required some enhancement or completely new development in order to pass certification. See Table 1.

Table 1. EHR Functional Gaps - 2011

Care Site	Modular System	Functionality Category	Later Removed
Ambulatory	HELP2	Automate measure calculation (ambulatory)	
Ambulatory	HELP2	Electronic prescribing	
Ambulatory	HELP2	Maintain up-do-date problem list	
Ambulatory	HELP2	Patient-specific education resources	
Ambulatory	HELP2	Submission to Public Health Registries	
Ambulatory	HIE	Clinical summaries	
Ambulatory	HIE	Exchange clinical info and summary report	
Ambulatory	Centricity Business	Access control	
Ambulatory	Centricity Business	Audit log	
Ambulatory	Centricity Business	Authentication	
Ambulatory	Centricity Business	Automatic log-off	
Ambulatory	Centricity Business	Emergency access	

Care Site	Modular System	Functionality Category	Later Removed
Ambulatory	Centricity Business	Encryption when exchanging EHI	
Both	HELP2	Access control	
Both	HELP2	Audit log	
Both	HELP2	Authentication	
Both	HELP2	Automatic log-off	
Both	HELP2	Computerized provider order entry	
Both	HELP2	Emergency access	
Both	HELP2	General Encryption	
Both	HELP2	Integrity	
Both	My Health Portal	Audit log	
Both	My Health Portal	Authentication	
Both	My Health Portal	Emergency access	
Both	EDW	Clinical Quality Measures	
Hospital	ECIS	Audit log	x
Hospital	ECIS	Emergency access	x
Hospital	ECIS	Maintain active medication allergy list	x
Hospital	ECIS	Maintain up-do-date problem list	
Hospital	ED System	Audit log	
Hospital	ED System	Calculate BMI	
Hospital	ED System	Emergency access	
Hospital	ED System	Maintain up-do-date problem list	
Hospital	ED System	Smoking Status	
Hospital	ED System	Vital Signs	
Hospital	HELP1	Audit log	x
Hospital	HELP1	Automate measure calculation (Hospital)	x
Hospital	HELP1	Calculate BMI	
Hospital	HELP1	Computerized provider order entry	x
Hospital	HELP1	Emergency access	x
Hospital	HELP1	Smoking Status	x
Hospital	HELP1	Submission to Public Health Registries	
Hospital	HELP2	Patient-specific education resources	
Hospital	HIE	Electronic copy of discharge instructions	
Hospital	HIE	Electronic copy of health information	
Hospital	HIE	Electronic copy of health info (d/c summary)	
Hospital	HIE	Exchange clinical info summary record	

We also reviewed the CMS Incentive Program Rule[5] to understand the Stage 1 requirements necessary to meet meaningful use in the ambulatory and hospital settings. Our workflow analysis identified 10 meaningful use workflows that we projected would require change to clinician workflow to bring all of our hospitals and ambulatory providers above the required thresholds. These workflows were identified as Problem List (hospital and ambulatory), Medication Allergy List (hospital and ambulatory), Computer Provider Order Entry (CPOE List hospital and ambulatory), Smoking Status List (hospital and ambulatory) , Patient Education List (hospital and ambulatory), Electronic Prescribing (ambulatory only), and Timely Access to Health Information (ambulatory only). See Table 2.

Table 2. Meaningful Use Status 2011 – Workflow Challenges

Final Meaningful Use Stage 1 Objectives	Measure Requirement Threshold	Percent of Providers that Meet Measure	Number of Hospitals that Meet Measure out of 22
Use Computerized Provider Order Entry (CPOE)	30%	55%	18
Electronic Prescribing [EP only]	40%	5%	N/A
Problem List	80%	20%	None

Final Meaningful Use Stage 1 Objectives	Measure Requirement Threshold	Percent of Providers that Meet Measure	Number of Hospitals that Meet Measure out of 22
Medication Allergy List	80%	45%	5
Smoking Status	50%	20%	19
Timely Electronic Access to Health Information [EP Only]	10%	20%	N/A
Patient Specific Education	10%	50%	None

Our 2011 analysis demonstrated that we had significant work to do to close the EHR functional gaps and the meaningful use gaps. We divided the work into three major projects; EHR certification, led by our Information Systems (IS) division, hospital meaningful use implementation, and ambulatory meaningful use implementation. Intermountain made reaching meaningful use a board goal for our system, giving the project top 3 status in our list of enterprise projects. This decision by leadership allowed all three project teams to obtain the staffing and clout to mobilize to meet the challenge. Projects prioritized lower were postponed to allow for the meaningful use project to move forward. Likewise, the Intermountain Medical Group of ambulatory physicians and the 22 hospitals prioritized meaningful use highly, making it a board-level goal for their organizations.

Methods

This analysis was performed at Intermountain Healthcare, a not-for-profit integrated health care delivery network which operates 22 hospitals (130,000 admissions per year), employs over 900 physicians working in 170 ambulatory clinics. Intermountain's clinical information systems have been described previously.[7,8] We currently use two home-grown, legacy clinical information systems, HELP in the hospitals and HELP2 primarily in the ambulatory setting. Over 13,500 unique users access HELP to retrieve results and/or document care for over 123,000 patient records per month. Over 13,000 clinicians use the HELP2 EHR each month to access records or document care on over 260,000 unique patients. Providers access different modules for different functionality, including documentation of progress notes, problem lists, medication orders, nursing documentation, etc.

Developing EHR Functionality for Certification

We assembled twelve teams and divided all of the Stage 1 EHR NIST requirements that required enhancement or creation among the teams. These functions are shown in Table 1. Each team consisted of a team lead analyst, programmer, informaticist, terminology analyst, quality assurance analyst, certification analyst and often an interface analyst. These teams reported to a project manager, and certification lead. The project was prioritized at number 1 or 2 in IS during 2012 and 2013

Achieving Meaningful Use

Intermountain formed the Meaningful Use Steering Committee (MUSC) to oversee the goal of attaining meaningful use in our Intermountain Medical Group (IMG) of employed physicians and in our 22 hospitals. This committee included VP level physician and nursing executives, certification lead, project manager, informaticist, MU Initiatives Manager, and regulatory advisor from the IMG and hospital system. The team met weekly and could escalate urgent issues to the Chief Information Officer, Chief Medical Officer, and Chief Nursing Officer for assistance in prioritization for resolution. Each Hospital and IMG regional champion formed Meaningful Use working groups that were responsible for setting and tracking meaningful use goals. These working groups consisted of regional physician and nurse champions, regional nursing consultants, and clinical systems specialists to train on and track meaningful use goals. Meaningful Use metric dashboards with drill-down capability were created for hospitals and IMG regions and clinics and were monitored for progress and/or issues by the MUSC.

Hospital working group tasks included education of the hospital leadership about meaningful use, assembling local teams representing nursing, physicians, information systems, finance, regulatory compliance, and health information management - medical records (HIM). These teams met to review requirements and reinforce or establish workflows to meet the meaningful use targets. Departments with deficiencies in particular areas, such as Problem List or CPOE were encouraged to adopt successful workflows that were in place in other Intermountain locations. Expectations were set by local leadership to stress the importance of compliance with the meaningful use measures.

The IMG ambulatory physicians working groups worked with clinical systems specialists and clinic managers to identify workflow deficiencies, then to modify and improve these workflows. Yearly provider incentives (15% of base pay) were based partially on whether a provider reached meaningful use.

Objective Impacts

All information system team members tracked their time for function development and certification-related tasks in project management software. This time tracking data as well as resource hourly rate was stored in our enterprise data warehouse (EDW) and converted to a data cube for analysis in order to determine time and dollar cost for the EHR certification project. Work-in-progress (WIP) Actual amount in dollars is defined as the WIP times an FTE rate, which is a normalized FTE rate by role. Ambulatory and hospital meaningful use team leads estimated time and effort project tasks by resource group to achieve meaningful use. Hourly rate for meaningful use project team roles was derived from an average of rate by role taken from Intermountain human resources records.

Workflow changes were tracked by following changes in meaningful use automated measure metric values, e.g. percent of a hospital's patients that had a coded problem on their problem list. These metrics, along with clinical quality measure (CQM) measures were generated weekly for all providers and hospitals and stored in our EDW. Total dollar incentive payments received for successful attestation for all providers and hospitals from Medicare and Medicaid were tabulated.

Subjective Impacts

Subjective impacts of the EHR certification and Meaningful Use project, both negative and positive, were elicited from information system, hospital, and physician executive leadership. Clinical leadership was asked about the financial and reputational impact as well as impact on patients, physicians/nurses and ancillary staff. Information systems product managers were asked whether the EHR certification and Meaningful Use project had important impacts such as promotion or demotion of other projects.

Results

Objective Impacts

The total hours and estimated costs to achieve Stage 1 EHR certification for outpatient and inpatient systems are shown in Table 3. Total hours were 221,765 between 2011 and 2013 and estimated costs came to $12.2M.

Table 3. EHR Certification Expenditures

	2011		2012		2013		2011-2013	
IS Systems	WIP Actual HRS	WIP Actual AMT	WIP Actual HRS	WIP Actual AMT	WIP Actual HRS	WIP Actual AMT	Total WIP Actual HRS	Total WIP Actual AMT
CPOE	6,671	$ 389,345	29,904	$ 1,715,264	34,189	$ 2,022,520	70,762	$ 4,127,129
HITECH	2,384	$ 144,385	8,324	$ 481,563	6,091	$ 360,812	16,797	$ 986,759
ANCILLARY	7,988	$ 438,193	26,750	$ 1,487,836	21,950	$ 1,197,713	56,687	$ 3,123,741
HELP1	4,532	$ 230,780	13,759	$ 692,220	12,264	$ 634,661	30,554	$ 1,557,662
HELP2	5,132	$ 264,077	19,147	$ 995,114	22,686	$ 1,173,950	46,965	$ 2,433,141
Totals	26,708	$ 1,466,780	97,885	$ 5,371,997	97,179	$ 5,389,655	221,765	$ 12,228,432

The total hours and estimated costs to meet and attest for Meaningful Use for hospitals are show in Table 4 and for eligible professionals in Table 5. Respectively, these totals were $3.43M and $1.67M totaling $5.1M.

Table 4. Hospital Meaningful Use Expenditures

Intermountain EH		2011			2012			2013		
EP or EH	FTE Type	Number FTEs	average % of annual time on MU	Estimated Cost	Number FTEs	average % of annual time on MU	Estimated Cost	Number FTEs	average % of annual time on MU	Estimated Cost
EH	MU Business Lead	1	30	$ 27,000	1	30	$ 27,000	1	20	$ 18,000
EH	Medical Informaticist	1	60	$ 54,000	1	60	$ 54,000	1	30	$ 27,000
EH	MU Initiatives Manager	1	90	$ 81,000	1.5	90	$ 121,500	2	65	$ 117,000
EH	MU Regional Champions	12	20	$ 216,000	12	20	$ 216,000	20	15	$ 270,000
EH	MU Subject Matter Experts	3	15	$ 40,500	4	15	$ 54,000	5	5	$ 22,500
EH	Clinical Systems Specialists	22	10	$ 198,000	22	40	$ 792,000	22	40	$ 792,000
EH	MU Reports (not in IS or EDW)	2	40	$ 72,000	3	40	$ 108,000	3	30	$ 81,000
EH	Privacy/Security/Regulatory	1	15	$ 13,500	1	15	$ 13,500	1	5	$ 4,500
EH	Finance	1	5	$ 4,500	1	5	$ 4,500	2	2	$ 3,600
		44		$ 706,500	46.5		$ 1,390,500	57		$ 1,335,600
										$3,432,600

Table 5. Eligible Professional Meaningful Use Expenditures

Intermountain EP		2011			2012			2013		
	FTE Type	Number FTEs	average % of annual time on MU	$	Number FTEs	average % of annual time on MU	$	Number FTEs	average % of annual time on MU	$
EP	MU Business Lead	1	50	$ 45,000	1	50	$ 45,000	1	40	$ 36,000
EP	Medical Informaticist	1	80	$ 72,000	1	80	$ 72,000	1	80	$ 72,000
EP	Training Regional Nurse Consultant			$ -	8	20	$ 144,000	8	20	$ 144,000
EP	Clinical Systes Specialists	3			5	50	$ 225,000	5	50	$ 225,000
EP	MU Attestation Lead	1	0	$ -	1	80	$ 72,000	1	80	$ 72,000
EP	MU Attestation Proxies			$ -	9	10	$ 81,000	9	10	$ 81,000
EP	IMG MU Reports			$ -	2	25	$ 45,000	2	50	$ 90,000
EP	MU Subject Matter and Workflow Experts	1			3	10	$ 27,000	4	10	$ 36,000
EP	Privacy/Security/Regulatory	1	20	$ 18,000	2	10	$ 18,000	2	10	$ 18,000
EP	Finance	1	10	$ 9,000	1	10	$ 9,000	1	10	$ 9,000
EP	Other			$ -			$ -			
		9		$ 144,000	33		$ 738,000	34		$ 783,000
										$1,665,000.00

608 of 696 (87%) outpatient eligible professionals (EP) successfully attested for meaningful use in 2012 for Stage 1, year 1. The incentives received professionals are shown in Table 6. Total incentives received for Stage 1, year 1(2012) and projected for year 2(2013) are $17.2M.

Table 6. IMG Ambulatory Provider (EP) Incentives

IMG Provider Grouping	**Number of Providers 2012**	Incentives Received 2012	Number of Providers 2013	Projected Incentives 2013
IMG Providers (Physicians & Mid-Levels)	1090		1342	
IMG Providers who are EPs	696		746	
IMG Providers Attested	629		729	
IMG Providers Attested(processed) successfully	608	$ 9.2M	679	$ 8M

All 22 Intermountain hospitals successfully attested for Stage 1 year 1 in fiscal year 2013. The incentives received and pending for hospitals are shown in Table 7. Received payments totaled $19M. Pending payments total $9.6M.

Table 7 Hospital Stage 1, Year 1 Incentives 2013

Faciltiy Name	Medicare Incentive	Date		Medcaid Incentive	Date
Alta View	$ 975,783	31-Dec		$ 305,730	28-Feb
American Fork	$ 1,058,743	31-Dec		$ 493,712	7-Mar
Bear River	$ 470,792	31-Dec		$ 397,475	7-Mar
Dixie	$ 1,810,081	31-Dec		$ 697,000	PENDING
Garfield	$ 910,616	31-Dec		$ 192,000	PENDING
IMED	$ 2,363,523	31-Dec		$ -	
LDSH	$ 1,029,818	31-Dec		$ 750,000	PENDING
Logan	$ 593,134	31-Dec		$ 623,000	PENDING
MKD	$ 2,424,138	31-Dec		$ 893,000	PENDING
Orem	$ 8,213	31-Dec		$ 20,000	PENDING
Park City	$ 602,896	31-Dec		$ -	
Riverton	$ 397,890	31-Dec		$ 475,000	PENDING
Sevier	$ 958,048	31-Dec		$ 381,000	PENDING
TOSH	$ 865,842	31-Dec		$ -	
UVRMC	$ 1,737,077	31-Dec		$ 1,356,000	PENDING
Valley View	$ 1,056,146	31-Dec		$ 589,000	PENDING
Fillmore (CAH)	$ 57,000	PENDING		$ 360,000	PENDING
Heber (CAH)	$ 138,000	PENDING		$ 417,000	PENDING
Sanpete (CAH)	$ 150,000	PENDING		$ 571,700	28-Feb
Cassia (CAH)	$ -			$ 459,141	31-Dec
Delta (CAH)	$ -			$ 446,000	PENDING
PCMC	$ -			$ 2,094,000	PENDING
received	$ 17,262,742		received	$ 2,227,758	
pending	$ 345,000		pending	$ 9,293,000	

For Stage 1, through 2013 EP (year 1 and year 2) and EH (year 1) received and pending incentives totaled $46.3M. Total costs for EHR certification and meaningful use implementation and attestation were $17.3M.

We identified certain meaningful use clinician workflows we anticipated would be challenging. Meaningful Use could only be achieved by an EP or hospital if all measures were met. These challenging workflows, with before and after results, are shown in Table 8. For instance, only 55% of EPs met the CPOE measure at the beginning of Stage 1. By the end of Stage 1 attestation period, 100% of providers met the CPOE measure.

Table 8 Workflow Challenge Results, Before and After Stage 1 Implementation

Meaningful Use Stage 1 Objectives	Measure Requirement Threshold	Percent of Providers that Met Measure BEFORE	Percent of Providers that Met Measure AFTER	Number of Hospitals that Meet Measure out of 22 BEFORE	Number of Hospitals that Meet Measure out of 22 AFTER
Use Computerized Provider Order Entry (CPOE)	30%	55%	100%	18	22
Electronic Prescribing [EP only]	40%	5%	94%	N/A	N/A
Problem List	80%	20%	84%	None	22
Medication Allergy List	80%	45%	94%	5	22

Meaningful Use Stage 1 Objectives	Measure Requirement Threshold	Percent of Providers that Met Measure BEFORE	Percent of Providers that Met Measure AFTER	Number of Hospitals that Meet Measure out of 22 BEFORE	Number of Hospitals that Meet Measure out of 22 AFTER
Smoking Status	50%	20%	99%	19	22
Timely Electronic Access to Health Info [EP Only]	10%	20%	85%	N/A	N/A
Patient Specific Education	10%	50%	98%	None	22

Subjective Impacts

Information system leaders commented that some important projects were delayed by EHR certification and meaningful use. These included replacement of legacy system functionality in labor and delivery, electronic consent handling, Clinical Health Information Exchange workflow integration, Exchange Provider Directory, enterprise provider master replacement, and replacement/enhancement of inpatient CPOE functionality (vs emergency department CPOE, which was used for meaningful use). IS leadership did mention that some projects were accelerated due to the meaningful use project including CPOE in the Emergency Department, and E-prescribe functionality.

Clinical leadership from the IMG and hospitals both agreed that there was reputational benefit from achieving meaningful use. In addition, they agreed on the benefit of the incentives and potential avoidance of future penalties. IMG leadership felt that patient engagement was likely improved after meaningful use, and was neutral on patient safety and patient satisfaction. Hospital leadership was neutral on the question of patient engagement, patient safety, and patient satisfaction. Both hospital and IMG leadership had feedback that physician and nurse productivity was negatively impacted. The IMG and hospital leadership also felt that clinic and emergency department (ED) physician, ED and clinic nurse, and clinic ancillary staff satisfaction was negatively impacted, due to the extra work necessary to meet meaningful use.

Discussion

This evaluation covers two very large projects at our institution; EHR enhancement/certification and EHR implementation at clinics and hospitals in order to meet meaningful use. This is by no means a comprehensive summary and evaluation of these two projects. However, we do describe with some objective and subjective information, some preliminary findings that inform the cost and benefit proposition for the EHR Incentive Program. The availability of detailed project information such as tasks, time, and resources for both the EHR enhancement/certification effort, and for the organizational work to meet meaningful use both in the outpatient and inpatient setting was very informative to us. We also tracked growth in EHR function usage, such as CPOE, E-prescribe, and Problem List throughout the clinics and hospitals as a result of the meaningful use requirements. Our study had several limitations. We did not track the impact/cost of the workflow changes to clinicians (such as time nor effort spent). Nor did we formally survey all the clinicians about workflow changes and satisfaction.

The decision to certify our EHR and meet meaningful use for Stage 1 use was a success from a strictly financial perspective. However, the true cost and benefit to the organization is not yet completely understood. We can see that the use of the EHR functions that are prescribed by meaningful use has grown. However, this is balanced with preliminary subjective feedback from clinicians that feel that patients may be more engaged, while these same clinicians are feeling less productive and less satisfied. Further analysis on the impact of the EHR incentive program is needed to shed light on this dilemma. Meanwhile, the Stage 2 requirements and workflows are looming.

Conclusion

This paper describes the experience and some preliminary impacts resulting from EHR certification, implementation, and successful attestation for meaningful Stage 1 at Intermountain Healthcare.

References

1. Steinbrook R. Health Care and the American Recovery and Reinvestment Act. N Engl J Med 2009 360: 1057-1060
2. CMS Data Update, Health Information Technology Policy Committee Meeting, March, 2014. Accessed from http://www.healthit.gov/FACAS/sites/faca/files/HITPC_CMS_Update_2014-03-11.pptx
3. Bowes, WA III, Progress and challenge in meeting meaningful use at an integrated delivery network. AMIA Annu Symp Proc. 2011;2011:144-51. Epub 2011 Oct 22.
4. Health Information Technology: Initial Set of Standards, Implementation Specifications, and Certification Criteria for Electronic Health Record Technology; Final Rule, 45 C.F.R. § 170, accessed from http://edocket.access.gpo.gov/2010/pdf/2010-17210.pdf, March, 2011
5. NIST Approved Test Procedures Version 1.0, accessed from http://healthcare.nist.gov/use_testing/finalized_requirements.html
6. Medicare and Medicaid Programs; Electronic Health Record Incentive Program; Final Rule, 42 C.F.R. Parts 412, 413, 422 et al., accessed from http://edocket.access.gpo.gov/2010/pdf/2010-17207.pdf, March, 2011.
7. Gardner RM, Pryor TA, Warner HR. The HELP hospital information system: update 1998. Int J Med Inform. 1999 Jun;54(3):169-82.
8. Clayton PD, Narus SP, Huff SM, et al. Building a comprehensive clinical information system from components. The approach at Intermountain Health Care. Methods Inf Med. 2003;42(1):1-7.

Developing an eBook-Integrated High-Fidelity Mobile App Prototype for Promoting Child Motor Skills and Taxonomically Assessing Children's Emotional Responses Using Face and Sound Topology

William Brown III, DrPH, MA[1,2], Connie Liu, MA[1], Rita Marie John, CPNP, DNP, EdD[3], Phoebe Ford[4]

[1]Department of Biomedical Informatics, Columbia University, New York, NY; [2]HIV Center for Clinical and Behavioral Studies, New York Psychiatric Institute and Columbia University, New York, NY; [3]School of Nursing, Columbia University, New York, NY; [4]School of International and Public Affairs, Columbia University, New York, NY

Abstract

Developing gross and fine motor skills and expressing complex emotion is critical for child development. We introduce "StorySense", an eBook-integrated mobile app prototype that can sense face and sound topologies and identify movement and expression to promote children's motor skills and emotional developmental. Currently, most interactive eBooks on mobile devices only leverage "low-motor" interaction (i.e. tapping or swiping). Our app senses a greater breath of motion (e.g. clapping, snapping, and face tracking), and dynamically alters the storyline according to physical responses in ways that encourage the performance of predetermined motor skills ideal for a child's gross and fine motor development. In addition, our app can capture changes in facial topology, which can later be mapped using the Facial Action Coding System (FACS) for later interpretation of emotion. StorySense expands the human computer interaction vocabulary for mobile devices. Potential clinical applications include child development, physical therapy, and autism.

Introduction

The importance of stories in physical and emotional development, and health practice, cannot be overstated. Stories are a universal concept across cultures that predate the written word. Stories have been told from prehistoric to contemporary to modern times. They also transcend multiple cultural barriers: Language, age, culture, and education. As a result, many stories are shared across cultures and/or are retold with differentiating degrees of variation and purpose. The earliest of storytelling was oral and heavily combined with gestures and expression to enhance the effectiveness of conveyance.[1] As a result, stories can also be a mechanism for physical development and have the ability to elicit emotional responses. Unfortunately, gesture and expression in storytelling has not translated to mobile technology (i.e. eBooks) to the same degree.

Computers and information technology are relatively new to the process of storytelling. Newer still are mobile devices. However, their growing ubiquity and advancing technology potentiate storytelling in new, dynamic, and undiscovered ways. Commonly, mobile devices rely on screen input interface and touch to simulate the standard functions of books. Swiping your finger from one side of the screen to the other performs page turning, and page leaves are animated to resemble real pages. Tapping and pinching are additional features offered by mobile devices. They allow zooming into pages for better views and selection of text to capture content for alternate uses.

Unfortunately, these actions do not meet the range of motion necessary for a child's comprehensive "motor skills" development. They also do not provide sensory and algorithmic identification methods for confirming that the young reader, or the child being read to, has actually performed the task. However, there are many unused mobile device sensors that, if leveraged, can: 1) enhance the way children use motor skills while experiencing eBooks, 2) confirm the execution of specific motor skills, and 3) provide a better understanding of the child's emotional experience during the course of their interactive eBook experience.

We identified several untapped mobile device sensor resources ideal for identifying motor development and assessing emotion during storytelling when using eBooks. We chose to focus on the two sensors that are least used, yet ubiquitously available on most mobile devices. The two sensors are the camera for image capturing and the microphone to capture sound. Advances in camera technologies more accurately simulate the capability of an eye where small and large ranges of motion can be captured, as well as objects detected and identified. Sound detection

is the most common feature of any mobile device, and is able to detect and differentiate greater ranges of sound than ever possible.

Given the knowledge gap related to mobile device assisted motor development and emotion detection, and the opportunities provided by new mobile technologies, we aimed to answer the question "Can an ebook-integrated high-fidelity mobile app prototype be developed for promoting child motor skills and taxonomically assessing children's emotional responses?" The purpose of this work was a proof-of-concept to develop "StorySense", an app that can promote child motor skills and identify emotion, for healthy child development. Our development goal included creation and testing of two sensory functions and identifying necessary programming and topology classification libraries. To evaluate our work we employed a system development life cycle stage-based evaluation model.

Theoretical base

The theoretical basis for this application's use in child development comes from the fundamentals of interactionist theories such as the neuronal group selection theory and the dynamic system theory.[2,3] These theories support that the infant's motor abilities emerge as a result of the child, task, and the environment with individual variability.[4] The plasticity of the brain in children is the neurophysiological basis for promoting motor skills in young children. Research over the past fifty years has confirmed that an infant's brain is built over time and that the development allows for future skills to emerge. Brains are modulated by genetics and experiences that will affect the outcome of the child.[5] StorySense provides interactive experiences similar to those that are known to have a modulation effect on the outcome of a child's gross and fine motor development.

Expression through facial action, gesticulation, and sound

Facial expressions are fundamental to the communication of simple and complex emotion. Movements in the muscles and skin, particularly around the mouth and eyebrows, provide a large visual vocabulary of meaning and emotional terminology.[6] Similarly, gesticulations with hands and the creation of sound (e.g. clapping, rubbing, banging, tapping, snapping) are intrinsically tied to communicating, and can add depth to conveying emotional information while children experience eBooks.[7]

There are two types of facial expressions that contribute to communication, voluntary expression and emotional expression.[8] Voluntary expression follow learned display rules in emotion and are made consciously (e.g. blowing a kiss). Emotional expression, on the other hand, is often displayed unconsciously. This includes facial expressions like distress, disgust, interest, anger, contempt, surprise, and fear. Despite the fact that individuals do not realize they are producing these expressions, this visual autonomic vocabulary is information rich and universally comprehendible.[6] Thus, the eyes and mouth are a fundamental identifier in facial recognition.

Sounds are also used for processes of identification. Sound variation and scale carry varied meaning (e.g. clapping, screaming, and whistling). It is an ideal indication of information that is both consciously and subconsciously conveyed by potential technology users. Both sound and facial movement reveal feelings and thoughts. However, the range and types of feelings and thoughts can be very different and very unique to one mode of information conveyance versus the other. While facial movement can indicate things like attraction, disgust, uncertainty, sound on the other hand can indicate specific emotions through utterances, speech, and various onomatopoeia (e.g. hissing).[7] StorySense both leverages and builds on previous research in: face, motion, and sound recognition; facial action coding and emotion identification; sensing for people-centric applications; and reaction sensing. Below, we discuss work relevant and contributory to this project.

Facial Action Coding System (FACS)

The systematic categorization of facial movements to identify expression of emotion is a historic practice of psychologists. In 1978 Paul Ekman and W.V. Friesen developed a Facial Action Coding System (FACS) by analyzing changes in facial appearance created during various combinations of facial muscular contraction. Their goal was to develop a reliable scoring metric by which human raters can identify facial behaviors. The result was FACS a taxonomic system of human facial movements that can help raters code changes in elements of the face (i.e. eyes, mouth) and their muscular movements.[9]

Today, FACS is the most widely used descriptive measurement tool for facial behaviors, and aids computers to topologically detect faces and their geometry. It also allows for the reduction of subjectivity and the use of high-throughput computational methods. FACS measurement units are Action Units.[9] Consequently, FACS scores do not provide meaning of facial behavior, and can only be used descriptively. To address the need for meaningful interpretation of FACS scores, researchers developed the Facial Action Coding System Affect Interpretation Dictionary (FACSAID). FACSAID links facial expressions with their psychological interpretations and models them to facial behaviors, then stores this information in to a relational database.[10] Thus, the raw FACS scores potentially produced by a StorySense scan can be translated into more psychologically meaningful concepts. By leveraging both the FACS taxonomy and FACSAID as knowledge bases, the final version of StorySense will be able to produce emotional profiles of a reader's facial movement for clinical interpretation and possible therapeutic use.

Sensing for people-centered mobile applications

There have been many applications of sensing technology for people-centered mobile applications. Cameras are one of the most utilized features in mobile devices and have some of the most varied function. Originally the camera's function was solely relegated to pixilated single pictures. Now we find high resolution images that can be instantly manipulated, as well as motion cinematography.[11] Similarly, the microphone has advanced beyond the standard receiver and plays a larger role in capturing sound. It pairs with camera recordings to produce video recordings, takes dictation, and is used to input commands.[12] As previously mentioned accelerometers are at the forefront of much of the motion detection capabilities of mobile devices.[13] What's more, when paired with GPS systems and the gyroscope, the duo maximize the devices ability to orient itself in the universe.[12–14] Though this information can also be helpful in detecting the movement of the user, the combination of accelerometers and GPS are most often an indicator of distance traveled. The amount of distance traveled from one physical location to another is likely to be short and of less relevance during a child's experiencing while reading an eBook. Thus, we did not focus on these to available features.

Reaction sensing

A developing area in HCI is reaction sensing. Reaction sensing goes beyond the normal bounds of input and imputation. Researchers and developers are leveraging sounds and their variations in order to enhance HCI and create a multifaceted user experience.[15] Moreover, advanced reaction sensing leverages the behavioral cues that are autonomic to people.[15,16] In this way the interaction more easily mimics what a person would expect another person to react to. For instance, previous work in sound recognition includes scalable sound sensing for people-centric applications on mobile phones (i.e. SoundSense).[15] Work in face detection includes technologies such as blink detection for real-time eye tracking.[16]

Moreover, advances such as natural language processing (NLP), machine learning, and data mining, can combine with multimodal sensing technologies, and allow mobile devices to anticipate user needs and react organically to people's naturally occurring behavioral cues.[16–18] This has the added benefit of being able to track and log new and existing movements and reactions. As a result, the potential to increase the known vocabulary of reactions to stimuli means we can process, analyze, and understand behavioral cues collectively and longitudinally, making high-throughput analyses a real possibility.

Methods

We built StorySense in Java for Android. We successfully tested and ran our applications on a Nexus 7 tablet and Samsung Galaxy S III mobile phone. We focused on two areas of sensing, sensing sound produced by high motility (clapping) and low motility (snapping) processes, as well as detecting facial changes, with a focus on eye movement. We also systematically identified classic children's stories that were ideal for incorporating movement and promoting motor skills.

Criterion and methods for choosing a story for the prototype

We defined story criteria to optimize our application's user experience, diversify options for interaction, and integrate fluidly with new sensing techniques. Our criteria included: English availability, multiple translations (for universality), internationally stable storyline, high graphic artwork, abstract storyline, availability/expired copy write, and animation potential. Using the Google search engine and leveraging its "Scholars" database, we performed a library literature search of thousands of choices. Based on our criteria we narrowed our results down to three possible stories: Peter Pan, Cinderella, and The Wizard of Oz. All researchers reviewed the final three stories. Ultimately, we chose Peter Pan because, of the three story options, it had the best balance of being well known and a stable storyline across cultures. The stable storyline across cultures was the penultimate criteria because it is an ideal trait for future evaluation.

In addition to identifying an optimal story for the prototype, we also had to edit the features of the story and create additional interactive illustrations and communicate functionality (Figure 1).

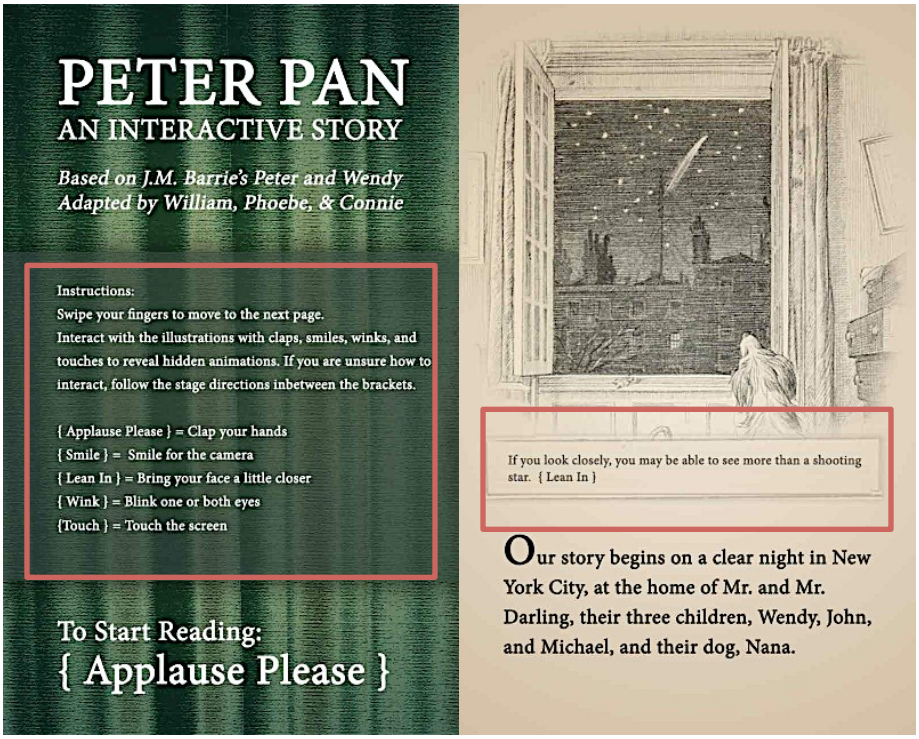


Figure 1. Modified illustrations (outlined in red).

Sensing sound: Clap/snap detection

For the purposes of prototyping, we chose to focus on clap and snap detection. We plan to explore detecting a wider range of sounds in the future. In consideration of the broad range of stories available as eBooks, it would be useful to detect and respond to a wider range of human sounds, such as sneezes, laughs, and snorts. However, sounds such as those we just listed are not produced using high-motility actions. Future noises that could be explored beyond clapping that are produced through high-motility actions include stomping and banging. However, we were unable to find these sound types to be useful in the same story. Originally we focused only on clap detection, however we needed a low-motility sound and verbalization (i.e. talking) for performance juxtaposition and sound identification. Thus, we quickly incorporated snap detection as well. This serves two purposes. In conditions where clapping is the preferred action; we want to make sure that sounds can be distinguished from similar sound patterns. This way the eBook can ensure that the desired physical activity is being performed. Secondly, we want to accommodate users who prefer to hold a mobile device with one hand and use snap detection. Furthermore, a story may elicit these sounds from its audience and perform a pre-designated response or action in return, or just respond to organically produced sounds as they happen naturally.

Sensing the face: Eye detection

We use facial recognition to detect and track user eye movement. Triggered by a user click, our application accesses the mobile device's camera and initiates a series of actions for facial recognition and eye tracking (Figure 2). In the development of the face recognition portion of the application, several decisions were made in regards to how we access a device's camera and how a user would be interacting with our application. First, we found there to be two different ways to capture images using the camera on Android. One is using an intent, which is using an existing camera application to take the photo. Second, is to create a custom camera application to take the photo. The first option would have another application come into the foreground, thereby occluding the story page and disrupting story flow. Thus, we chose the second option.

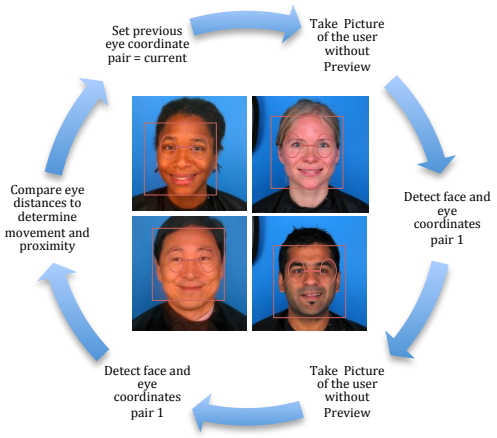


Figure 2. Face and eye detection. (Images from PICS database[19])

However, the second option had the Android required condition where in order to take a picture, there has to be a preview surface, meaning the user has to see what they are taking a picture of and then click to take a picture. The problem here is that we did not want a preview since it would occlude much of the story page as well as disrupt story flow. Ultimately, in order to get around this issue, we created a surface view and resized it to a single pixel on the screen, thus there is a preview, but it is not visible.

Evaluation

Development and evaluation of StorySense followed the systems development life cycle (SDLC), also referred to as the application development life-cycle, we have completed the preliminary analysis, systems analysis, requirements definition, systems design, development phases, integration and testing phase.[20] This evaluation highlights our experience and knowledge gained in the integration and testing phase. The following evaluation results are based on StorySense's current SDLC stage, benchmark lab testing.

Benchmark lab evaluation of system sensors

To evaluate StorySense's ability to detect and distinguish sound topology (i.e. talking, clapping, snapping) we used threshold detection. We looked at amplitude readings from mobile device and laptop microphones for a variety of sounds to determine a threshold. We observed amplitudes for: clapping slow and fast; clapping close and far; clapping slow and fast while talking; snapping; talking; yelling; and tapping the device. We ran these tests with two users, one who snaps loudly and one who snaps quietly. These performance tests were done while watching an amplitude meter to help us understand the basic differences in the amplitude over time. We recorded our results in the system and mapped clapping topologies to event triggers in the story.

We evaluated StorySense's ability to detect eye topology using visual tracking confirmation methods. The series of actions following the user click included capturing the user's photo using the front facing camera without a preview, thereby not disturbing or occluding the story page. The picture taken is then sent to a built-in face detector, and if a face is detected, the eye position is recorded. With the eye screen coordinates in-hand, a Peter Pan character image is drawn on the story page screen where the eyes were detected (Figure 3). That is, once we have one pair of eye coordinates; our camera module is able to take another user photo and extract the eye coordinates from that. If at least two pairs of eye coordinates are available, the application draws a Peter Pan character image at both eye locations and a dashed line between them to show the user's eye movement was tracked. This Peter Pan image allowed us to detect changes in eye topology, and was used in the prototype model for verification of metrics, direction, and actions performed.

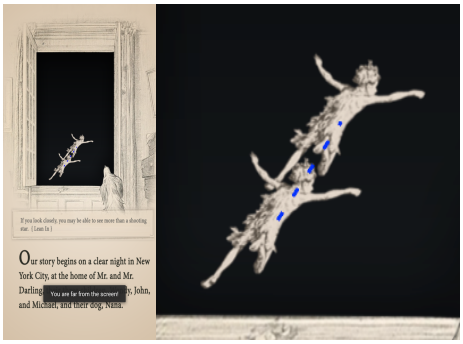


Figure 3. Eye tracking/Peter Pan indicators.

Furthermore, we conducted a performance evaluation of StorySense's face detection software, which consisted of testing if the application is able to detect faces (including recording the time used for detection in milliseconds [converted to seconds], and the resulting confidence factor) for a subset of 40 images from the Psychological Image Collection at Stirling (PICS) University database.[19] The image set from PICS included both ethnic and phenotypic diversity, and consisted of 22 male and 18 female images. The confidence factor used to gauge face detection is a property of the Android Software Development Kit's (SDK) "FaceDetector." This property is also known as the "Face Class", which holds information regarding the identification of a face in a bitmap. The confidence factor returns a value between 0 and 1, and it indicates how certain what has been found is actually a face.

Results

Sound detection evaluation results

The graphs below show that claps and snaps have significantly higher amplitudes than "indoor voice" talking at close range (~ a foot and a half away) (Figure 4). Another difference is that the claps and snap amplitudes go back down to zero very quickly, while talking has more of a zigzagged amplitude corona arch. Both of these features can be combined to yield reasonable results for clap and snap detection. While currently StorySense identifies anything above a threshold as a clap or a snap, it could benefit from further analyzing the sound following the initial trigger to determine if the sound is staying high or going back down quickly. This added feature would dramatically increase the ability of StorySense to detect claps and snaps.

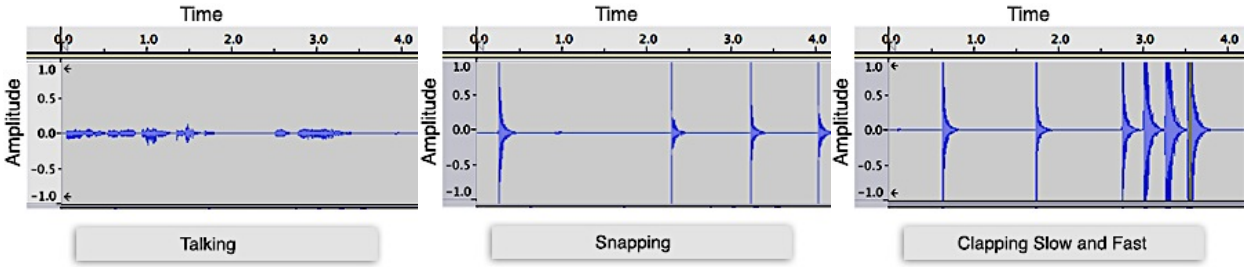


Figure 4. Sound types and amplitudes.

Unfortunately, the threshold we chose to use as our base level made it difficult to distinguish additional sounds such as snaps or short bursts from claps. If someone yells loudly or a nearby dog barks, the sounds will be registered as a clap and the system will respond accordingly. In general, people usually talk with "indoor voices" that stay below the threshold or read silently. More formal user testing could help to define the accuracy of this approach.

Face detection evaluation results

We were able to successfully procure eye coordinates; we also obtained the distance between the eyes. We were then able to use this information to determine additional changes in facial topology as well as the distance of the user from the screen. We could accurately calculate distance of user using eye size. The larger the eye was, the closer the user's face was to the screen. We could also use these variations in distance and eye size to detect eye blinking, proximity, and various movement types. Alternatively, if the user was too close (i.e. distance is greater than mobile device screen size threshold), then a warning successfully triggered and posted to the screen, telling the user to increase their distance from the face of the mobile device. This would remind users not to put their faces too close to the screen; thus, maintaining a safe viewing distance, as well as facial detection accuracy.

According to the Android application program interface (API) documentation, a confidence factor above 0.3 indicates good face detection. Our results from our face detection performance evaluation (Table 1) show that we have a 100% face detection (FD) rate, with average detection time being 2772 milliseconds (or 2.772 seconds) and average confidence factor being 0.522917008.

Table 1. Face detection performance evaluation (Time for face detection (FD) was converted from the original output in milliseconds to seconds).

	Image	FD?	Confidence Factor	Time for detection (Seconds)		Image	FD?	Confidence Factor	Time for detection (Seconds)
1	f4001s.jpg	Y	0.5312345	2.747	21	m4003s.jpg	Y	0.51446885	2.911
2	f4002s.jpg	Y	0.5130435	2.567	22	m4004s.jpg	Y	0.52985317	2.989
3	f4003s.jpg	Y	0.5289631	2.476	23	m4010s.jpg	Y	0.5295981	3.015
4	f4004s.jpg	Y	0.5270577	2.572	24	m4011s.jpg	Y	0.5311005	2.721
5	f4005s.jpg	Y	0.51732284	2.555	25	m4012s.jpg	Y	0.5211092	2.748
6	f4006s.jpg	Y	0.52806115	2.43	26	m4014s.jpg	Y	0.5209863	2.774
7	f4007s.jpg	Y	0.5279383	2.833	27	m4017s.jpg	Y	0.51728237	2.822
8	f4008s.jpg	Y	0.5243423	2.829	28	m4018s.jpg	Y	0.52777547	2.931
9	f4009s.jpg	Y	0.525947	2.888	29	m4021s.jpg	Y	0.5211933	2.963
10	f4010s.jpg	Y	0.5292073	2.837	30	m4024s.jpg	Y	0.5209574	3.002
11	f4016s.jpg	Y	0.512235	3.319	31	m4027s.jpg	Y	0.5180155	2.487
12	f4017s.jpg	Y	0.52486527	2.984	32	m4028s.jpg	Y	0.5112237	2.448
13	f4018s.jpg	Y	0.53505766	3.539	33	m4031s.jpg	Y	0.51334447	2.452
14	f4021s.jpg	Y	0.5200643	2.963	34	m4032s.jpg	Y	0.52210337	2.396
15	f4026s.jpg	Y	0.5255496	3.149	35	m4035s.jpg	Y	0.5270095	2.349
16	f4027s.jpg	Y	0.5266327	2.86	36	m4037s.jpg	Y	0.5241929	2.455
17	f4029s.jpg	Y	0.5280881	2.851	37	m4040s.jpg	Y	0.5187667	2.527
18	f4030s.jpg	Y	0.52181965	2.869	38	m4043s.jpg	Y	0.53127307	2.477
19	m4001s.jpg	Y	0.51076734	2.992	39	m4063s.jpg	Y	0.51823	2.47
20	m4002s.jpg	Y	0.5165685	2.889	40	m4064s.jpg	Y	0.52343065	2.794

Images were taken from the Psychological Image Collection at Stirling (PICS) University database: Category = 2D Face Sets (http://pics.psych.stir.ac.uk/2D_face_sets.htm); Set Name = Utrecht ECVP; Set Description = 131 images, 49 men, 20 women, collected at the European Conference on Visual Perception in Utrecht, 2008. Some more to come, and 3d versions of these images in preparation; Resolution = 900x1200 color.

Once eye movement and distance have been detected and calculated this data can be used to decompose the movement into specific AUs and extrapolated to produce a FACS score. Further, duration, intensity, and asymmetry can improve the accuracy of the score and its related FACSAID interpretation. Although we were currently only tested two uses of eye movement detection, the tracking capability provides possibilities for detecting and identifying other changes in face topology.

Discussion

The stage of our project is proof-of-concept. Our app currently detects discrete interactions, and contains system parameters to detect faces and sound with limited variation for both. Only a sample of the story is being used. System statistics indicate we have reason for concern that there will be device memory space issues when the full story is developed and running; thus, we have focused much of our subsequent work around optimizing the framework to ensure that it handles the loading of pages and animations more efficiently.

Better classifiers to differentiate between sets of sounds and images are highly necessary. Further, development of a larger sound and visual vocabulary and repository are additional goals for the prototype. We may be able to create a

foundational list of desired motor functions from our user survey and Morae 3.3 assessment tool during our first user evaluation, and then compare this list to the existing motions that can be detected by other apps in the Google play store library. Lastly, an immediate goal is to use OpenCV to improve image processing and facial feature detection.

What we have learned in the developmental phase of our app prototype are some of the optimal methods and barriers to creating an app that can be used to detect physical activity of children during their reading of an eBook. We have also identified ways that such functionality can be incorporated directly with the story. In the story of "Peter Pan", the section that tells the reader to clap their hands to bring Tinkerbelle back to life is a clear example of stories moments that directly elicit physical action to produce a story outcome.

Potential clinical applications and target populations

There are no studies using these kinds of application to promote childhood development. It has been shown that early intervention programs that focus on promoting motor skills, preverbal skills, and stimulation of brain development are highly effective. These programs use a variety of techniques that involve multiple senses. The cumulative experience with early intervention programs has confirmed these methods are effective in promoting infant development.[21] StorySense encourages children's motor skills and provides impetus for a response. The functions of StorySense are ideal for increasing child motor skills, and understanding the child's experience even when the child's ability to orally communicate is un-developed.

Our application could potentially facilitate communication skills and social interaction, as well as sustain the child's attention. For children and adolescents with autistic spectrum disorder (ASD) presenting with limitations in conventional forms of verbal and non-verbal communication, this application could provide an alternative form of therapy. Dependent upon the results of our future user evaluation, we could explore the realm of our application as an ASD clinical therapy that could identify limitations and weaknesses in children, as well as strengths and potentials.[22] Furthermore, we could explore application effects on communicative behavior, language development, emotional responsiveness, attention span and behavioral control over a period of time. It may also be possible to put real-time user specific health or physical activity information directly in the eBook story line to help both parents that are reading to their children and children readers to learn how to manage their health.

Contribution to the field(s)

StorySense contributes to the fields of facial recognition, expression coding, and dynamic learning by following the FACS standard of classification protocols and adding to its lexical library. Continued development of StorySense facial recognition algorithms should be able to code nearly any anatomically produced facial expression. During the story telling process, and guided by FACS and FACSAID, StorySense will be able to topologically deconstruct the child's facial movements into specific Action Units, calculate a FACS score, and identify the expression's related emotion. The AUs are independent of variations in human interpretation; thus, higher order decision such as the recognition of basal emotions, can be processed. Subsequently, once a facial expression and its related emotion have been identified, related actions can be pre-programmed into the eBooks story environment. Thus, emotion can also have an impact on the story line in a way that can be designed to address distress, anger, or discomfort, as well as leveraged for therapeutic use.

Future work

Our future work will aim to include geo-location data to trigger a function or change in the story line that could promote walking or encourage other movement (e.g. landscape animations could match the reader's location). We will also improve our application experience by incorporating references to points of interest near the user's current location.

Other future work will incorporate newer technologies and mobile device capabilities. For instance, recent developments such as spritzing technologies (http://www.spritzinc.com/blog/) could provide a way to improve a user's reading skills, speed and focus. Additionally, motion-sensing capabilities of mobile devices could open up a whole new channel of facilitating physical development, conceptually similar to what the Wii Remote motion sensing technology for Wii Sports does.

We were unable to build in a response to other high-motility actions (such as stomping or waving) due to lack of classifiers. Thus, we hope to contribute more classifiers to the established lexicons. Future programming will continue to build off of the work done around reaction sensing.[15,16] Part of our work will also include the development of a facial expression-to-emotion library that is common to children during the process of reading, or is able to distinguish facial movements that are specific to the reading process that may be easily confused with expressing emotion. Additional future work may include train a user specific image set to provide better user recognition with each application use. Thus, future application development will look to implement a more seamless image capture and camera process. More standard functionality such as dictionaries, word pronouncers, bookmarks, and highlighting will also be added.

Future evaluation

Future evaluation will involve usability testing using a user-centered iterative design process, and Morae 3.3 a usability data collection tool to capture audio, video, on-screen activity, and keyboard/mouse/touch input so that we may identify use and error patterns and gain insight into the effectiveness and acceptability of the mobile app's design. NVivo will be used to analyze qualitative data to provide insight into areas for future development.

Limitations

In the development of our application, we had difficulties with sound classification. Loud ambient noises may register as constant clapping. Moreover, as we added animations and more images to test the framework, it became apparent that there were memory related performance issues. Additionally, though face movement can be detected, the app does not yet differentiate between open/close eyes or mouths, which hindered wink and smile detection. Again, we plan on developing classifiers that advances the built-in functions of the mobile device to ameliorate these issues. Also, our sound threshold for testing made it difficult to distinguish additional sounds such as snaps or short bursts from claps. Other loud sounds are at risk of being registered as a clap and the system will respond accordingly. Future iterations of the application might incorporate OpenCV (A library of programming functions mainly aimed at real-time computer vision, http://opencv.org/), open source data sets, and/or more advanced libraries to improve overall face and sound detection.

Using a set of captured 2D images versus actual users is a possible limitation to our evaluation. Moreover, we have thus far only tested StorySense on a Nexus 7 tablet and Samsung SGH-T999 Android 4.1.2 (API 16) (Galaxy S3) 1.9 Mega pixel camera, HD recording @30fps with Zero Shutter Lag, BSI. Newer phone models and other devices that are more widely used (e.g. Apple iPhone) would not be immediately supported. We look to remedy this by creating an iOS version and using emulators as well as acquiring additional mobile devices for testing.

Lastly, the FACS taxonomy is based off of analyzing adult facial movements. There may be unforeseen variations in adult facial recognition. Also, the FACSAID database is based off of FACS codes, which may present a similar problem when classifying child emotions using an emotion dictionary database based on analysis from adult populations.

Conclusion

We accomplished a significant amount of our interactive goals. In the development of this prototype we explored sound and facial expressions, which could potentially contribute to several universal knowledge bases for reaction sensing (i.e. FACS). StorySense leverages unconscious and conscious cues in every day communication. We found that by using mobile sensory technology (i.e. camera and microphone), and integrating common algorithms, it is possible to obtain rich user information. In this way, stories on mobile devices can not only be read by the user, but the mobile device can read the user's conscious or unconscious emotional expression and cues, and adjust the story for a more dynamic and user-centered experience. We also found that sensors used for dynamic and user-centered experiences can be leveraged to promote high-motility action, which would aid in the development of motor skills for young children or children with developmental challenges.

Acknowledgments

Dr. William Brown III is supported by NLM research training fellowship T15 LM007079 and NIMH center grant P30 MH43520. We would also like to thank Dr. Suzanne Bakken for her review on an earlier version of the manuscript.

References

1. VanSledright B, Brophy J. Storytelling, Imagination, and Fanciful Elaboration in Children's Historical Reconstructions. Am Educ Res J. 1992 Dec 21;29(4):837–59.
2. Edelman GM. Neural Darwinism: The theory of neuronal group selection. New York, NY, US: Basic Books; 1987. 371 p.
3. Thelen E. The (re)discovery of motor development: Learning new things from an old field. Dev Psychol. 1989;25(6):946–9.
4. Hadders-Algra M. The Neuronal Group Selection Theory: a framework to explain variation in normal motor development. Dev Med Child Neurol. 2000;42(8):566–72.
5. Adams RC, Tapia C, Murphy NA, Norwood KW, Adams RC, Burke RT, et al. Early Intervention, IDEA Part C Services, and the Medical Home: Collaboration for Best Practice and Best Outcomes. Pediatrics. 2013 Oct 1;132(4):e1073–e1088.
6. Ekman P, Friesen WV. Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues. ISHK; 2003. 200 p.
7. Quek F, McNeill D, Bryll R, Duncan S, Ma X, Kirbas C, et al. Multimodal human discourse: gesture and speech. ACM Trans Comput-Hum Interact. 2002;9(3):171–93.
8. Matsumoto D, Ekman P. Facial expression analysis. Scholarpedia. 2008;3(5):4237.
9. Ekman P, Friesen WV. Facial action coding system: A technique for the measurement of facial movement. Palo Alto: CA: Consulting Psychologists Press; 1978.
10. Merten J. Facial microbehavior and the emotional quality of the therapeutic relationship. Psychother Res. 2005;15(3):325–33.
11. Reynolds F. Camera Phones: A Snapshot of Research and Applications. Pervasive Comput IEEE. 2008;7(2):16–9.
12. Howell J, Schechter S. What You See is What they Get: Protecting users from unwanted use of microphones, camera, and other sensors. In Proceedings of Web 20 Security and Privacy Workshop. 2010.
13. Kwapisz JR, Weiss GM, Moore SA. Activity recognition using cell phone accelerometers. SIGKDD Explor Newsl. 2011 Mar;12(2):74–82.
14. Kim D, Kim J, Choa M, Yoo SK. Real-time Ambulance Location Monitoring using GPS and Maps Open API. Conf Proc IEEE Eng Med Biol Soc. 2008;2008:1561–3.
15. Lu H, Pan W, Lane ND, Choudhury T, Campbell AT. SoundSense: scalable sound sensing for people-centric applications on mobile phones. Proceedings of the 7th international conference on Mobile systems, applications, and services [Internet]. New York, NY, USA: ACM; 2009 [cited 2013 May 15]. p. 165–78. Available from: http://doi.acm.org/10.1145/1555816.1555834
16. Li K. Automatic Content Rating via Reaction Sensing. J ACM. 2013;
17. Morris T, Blenkhorn P, Zaidi F. Blink detection for real-time eye tracking. J Netw Comput Appl. 2002 Apr;25(2):129–43.
18. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. J Am Med Inform Assoc. 2011 Sep 1;18(5):544–51.
19. University of Stirling. Psychological Image Collection at Stirling (PICS) [Internet]. Department of Psychology [Online]; Available from: http://pics.psych.stir.ac.uk/
20. O'Brien JA. Management information systems. 10th ed. New York: McGraw-Hill/Irwin; 2011. 673 p.
21. Dreyer BP. Early Childhood Stimulation in the Developing and Developed World: If Not Now, When? Pediatrics. 2011 May 1;127(5):975–7.
22. Srinivasan SM, Bhat AN. A review of "music and movement" therapies for children with autism: embodied interventions for multisystem development. Front Integr Neurosci [Internet]. 2013 Apr 9 [cited 2014 Mar 13];7. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3620584/

Making Audit Actionable: An Example Algorithm for Blood Pressure Management in Chronic Kidney Disease

Benjamin Brown, MB ChB, MSc[1,2], Richard Williams, BA[1,2], Matthew Sperrin, PhD[1,2], Timothy Frank, MD, FRCGP[1], John Ainsworth, BSc, MSc[1,2], Iain Buchan, MD, FACMI[1,2]
[1]Centre for Health Informatics; [2]Greater Manchester Primary Care Patient Safety Translational Research Centre, University of Manchester, UK

Abstract

Despite widespread use of clinical guidelines, actual care often falls short of ideal standards. Electronic health records (EHR) can be analyzed to provide information on how to improve care, but this is seldom done in sufficient detail to guide specific action. We developed an algorithm to provide practical, actionable information for care quality improvement using blood pressure (BP) management in chronic kidney disease (CKD) as an exemplar. We used UK clinical guidelines and EHR data from 440 patients in Salford (UK) to develop the algorithm. We then applied it to 532,409 individual patient records, identifying 11,097 CKD patients, 3,766 (34%) of which showed room for improvement in their care: either through medication optimization or better BP monitoring. Manual record reviews to evaluate accuracy indicated a positive-predictive value of 90%. Such algorithms could help improve the management of chronic conditions by providing the missing link between clinical audit and decision support.

Introduction

Best practice clinical guidelines are widely used in health systems around the world, however, observed quality of care often falls short of these standards.[1] The gap between ideal and actual care often results in additional morbidity, mortality and preventable hospital admissions, each with human and financial costs.

Chronic kidney disease (CKD) is a prime example. In CKD, both declining kidney function (measured by estimated glomerular filtration rate: eGFR) and increasing urinary albumin:creatinine ratio (ACR) independently increase cardiovascular mortality risk 2-4 fold.[2] Clinical guidelines therefore recommend strict blood pressure (BP) control to reduce this risk.[3–8] Nevertheless, studies invariably demonstrate just 13-66% of CKD patients in the US[9–13] and 35-60% in Europe[14–19] actually achieve controlled BP levels. Not meeting these standards has substantial adverse consequences given the high global prevalence of CKD, estimated at 8-16%.[20]

Data from electronic health records (EHRs) are abundant and ever increasing. They reflect the real-world care that patients receive and are often compared against clinical guidelines for the purposes of audit.[21] EHRs also often capture reasons *why* patients may not have achieved quality standards in the first place (e.g. patient choice or contra-indication to treatment) and practical steps as to *how* they might be achieved in future (e.g. how to optimize current management). This information is rarely exploited by existing informatics tools that perform audit and feedback – both in experimental studies and in routine practice.[11,14,24–28] Feeding such information back to practitioners has been shown in meta-analyses to produce greater improvements in patient outcomes than simple audit because it is more actionable by practitioners.[22,23] Therefore adding this functionality to audit and feedback interventions is likely to lead to greater improvements in patient care by connecting it with consequent quality improvement actions.

Aim of this study

The aim of this study was to develop and estimate the accuracy of an algorithm that searches EHR data to feed back practical, actionable information to clinicians regarding how they could improve care for patients in accordance with clinical guidelines. We used the example of BP management in CKD to test the feasibility of this method.

Importance and relevance of this study

This algorithm could provide the missing link between clinical audit and decision support, thus reducing the latency between clinical quality information and actions to improve care. Focusing on CKD and BP management, the algorithm developed here can be applied directly to EHR data to help reduce adverse cardiovascular outcomes. The approach in general can also be abstracted to other chronic conditions where physiological parameters reflect, at least in part, the quality of care provided to patients (e.g. cholesterol in cardiovascular disease; glycated hemoglobin in diabetes).

Methods

Algorithm development

We used anonymized EHR data from the City of Salford (population 234k) in the UK to develop the algorithm. Since the early 1990s, most UK citizen's primary healthcare records have been stored as machine-readable clinical codes (mainly Read codes v2 and CTV3). Over 300k different entries exist, covering care-processes, diagnoses and medications.[29] Practitioners can enter narrative free-text to supplement certain Read codes, though it is the codes themselves that are used for official purposes in the UK National Health Service (NHS) such as provider payment, public health programmes, health service planning, population health data, clinical performance assessment, and research.[30] Read code collections therefore provide important insights into the UK population's health.

We extracted alphanumeric codes and associated rubrics from a central database of all patients that received care in Salford from 2001-2012. These were then transferred securely to The University of Manchester as text-based vector files for analysis in Microsoft SQL Server 2008. The Salford database has collated EHR data daily from 53 primary care providers and one secondary care provider since 2001. We designed the algorithm to make inferences about how to improve BP management for CKD patients using coded EHR data in the following sequence: 1) identify patients with CKD; 2) identify those from Step 1 in whom it was appropriate to pursue quality standards for BP management; 3) identify those from Step 2 who did not achieve the quality standards; 4) provide information on what action could be taken to enable patients from Step 3 to achieve the quality standards.

In accordance with accepted methodology,[31] we developed and tested the algorithm iteratively. Initial algorithms were written as narrative based on clinical guidelines and authors' (BB and TF) clinical knowledge. The narrative was then translated into machine readable form using existing codes within the EHR dataset and dictionaries of Read codes used for primary care performance related pay.[32] To optimize the performance of the algorithm, we calibrated it with manual clinician review. Two clinicians (BB and TF) manually reviewed the coded data of randomly selected anonymized EHRs of patients receiving care in Salford between 2010 and 2012. For each of the 4 steps in the algorithm sequence, batches of 20 records were reviewed (10 identified positive and 10 identified negative). Based on these findings, the algorithm programming logic was modified as necessary. When additional codes were added to the algorithm, synonyms were identified within the full dataset (2001-2012) and also included. Further reviews of records were undertaken until the algorithm correctly identified all patients in the batch.

Algorithm accuracy study

Participants: We applied the final algorithm to the whole Salford EHR dataset (2001-2012) to make inferences about how to improve the quality of CKD BP management in Salford between 2011 and 2012.

Test methods: To test the performance of the algorithm, we randomly selected patient records it identified that had not been used in the development phase to calculate the positive predictive value (PPV). We chose to use PPV for: 1) clinical importance – in its real-world application, each patient identified by the algorithm will require further action by a clinician, so achieving a high PPV is essential to prevent over-burdening practitioners; and 2) pragmatism – to adequately assess specificity and sensitivity would require the manual review of thousands of records for each stage in the algorithm sequence, for which resources were not available. We defined the PPV as the number of patients identified by the algorithm where the clinicians (BB and TF) deemed they had CKD and that steps could be taken to improve the management of their BP, divided by the total number of records identified by the algorithm. For each patient tested, both BB and TF independently reviewed the anonymized coded data from their EHR and made a clinical judgment in accordance with national UK CKD clinical guidelines.[4] Any disagreements were resolved through discussion.

Statistical methods: We used a worst-case scenario of a predicted 50% PPV to determine the necessary sample size to test the algorithm (though we anticipated the algorithm would perform much better). We calculated that to detect a PPV of 50+/-10% with a two-sided α of 0.05, a minimum sample size of 96 patient records were required. We undertook reviews of 100 records to account for the imperfect nature of the review process. Statistical analysis was performed in R.[33]

Results

Algorithm development

In total, we manually reviewed 440 different patient records over 22 iterative cycles during the development process. Between 2001 and 2012, we found over 165 million individual codes in the Salford data, of which 151

million (92%) were Read code v2 and 14 million (8%) were codes specific to commercial EHR information systems. Most codes (133 million, 81%) originated in primary care and 32 million (19%) originated in secondary care. Often the rubric associated with codes was re-written by clinicians, which added a further layer of complexity when searching for synonyms. We used only the primary care data, as it is largely the responsibility of primary care providers in the UK to manage BP in CKD patients. The narrative description of the final algorithm is in Table 1. The full algorithm and list of codes used is available from www.clinicalcodes.org.

Table 1. Narrative description of the fully developed algorithm sequence.

Step	Criteria

1	**Patients with CKD**
	• Include patients prior to the study period with:
	  o a CKD diagnosis code.
	  o a most recent mean eGFR<60 or ACR≥30 or PCR≥50 over a >90 day period.
	• Exclude patients with a death code at any time.

2	**Patients in whom it is appropriate to achieve quality standards for BP management**
	• Exclude patients with:
	  o a palliative care code or an orthostatic hypotension code at any time.
	  o an 'informed dissent' or 'unsuitability' code for CKD or BP quality indicators within 15 months of the study period.
	  o a 'maximal BP therapy' code within 15 months of the study period, or who were receiving 4 types of anti-hypertensive medication concurrently (or fewer if they were unable to receive 4 types due to allergies and/or contraindications), at the time of their latest BP measurement.

3	**Patients who have not achieved BP management quality standards**
	• Exclude patients with a DM diagnostic code prior to the study period:
	  o *and* microalbuminuria (defined as a microalbuminuria code at any time or most recent ACR≥3.5 [females] or ≥2.5 [males]) *or* proteinuria (defined as a proteinuria code at any time ACR≥70 or PCR≥100), whose latest BP is <130/80 *and* within 9 months of the study period end.[4,34]
	  o but *no* microalbuminuria or proteinuria (as defined above), whose latest BP is <140/80 *and* within 12 months of the study period end.[35]
	• Exclude patients *without* a DM code and *without* proteinuria (as defined above) prior to the study period, whose latest BP is <140/90 *and*:
	  o within 12 months of the study period end if a hypertension or CVD diagnostic code is present prior to the study period.[36–40]
	  o or within 15 months of the study period end otherwise.[4,32]

4	**Patients whose care could have been improved to achieve these quality standards**
	• Count the number of direct and indirect contacts* each patient had with the primary care health system since their last BP measurement or during the period of interest defined in Step 3 (whichever is latest).
	• Count how many patients with a high BP reading whose:
	  o Last BP reading is within 2 months of the study period end.
	  o Medication did not change.
	  o Medication changed.

Key: ACR: albumin:creatinine ratio (mg/mmol); BP: blood pressure (mmHg); CKD: chronic kidney disease; CVD: cardiovascular disease (ischaemic heart disease, peripheral arterial disease, stroke, or transient ischaemic attack); eGFR: estimated glomerular filtration rate (ml/min/1.73m^2); PCR: protein:creatinine ratio (mg/mmol).

* Direct contact: face-to-face or telephone contact; indirect contact: medication prescription or administrative activity where the EHR has been accessed.

As we used UK EHR data, we developed the algorithm based on clinical guidelines and standards issued by the National Institute for Health and Care Excellence (NICE)[41] – the authority for clinical guidelines in the UK. To determine whether a patient had a CKD diagnosis according to biochemical parameters in the guidelines (rather than simply a diagnostic code), we took an average of their most recent readings over a period of more than 3 months to account for any fluctuations, and by discounting readings of zero related to eGFR. When the NICE CKD guidelines did not cater for certain cohorts of patients (e.g. those with diabetes but without microalbuminuria, or those without diabetes), we identified appropriate clinical standards from other related NICE guidelines. Interestingly, we could not find NICE guidance on how often CKD patients without hypertension, diabetes or cardiovascular disease should have their BP measured, so we used the standard of 15 months in the UK Quality and Outcomes Framework.[32] We discounted patients from these quality standards (Step 2) based on our clinical experience and on rules used in the Quality and Outcomes Framework.[32]

We classified anti-hypertensive treatment according to the 6 groups in NICE guidelines: (1) ace-inhibitors and angiotensin-receptor blockers; (2) calcium channel blockers; (3) beta blockers; (4) thiazide-like diuretics; (5) other diuretics; (6) alpha-blockers. When a patient had a high BP reading, we ascertained whether there had been a change in medication by analysing the medication prescribed 63 days either side of the reading to account for bi-monthly prescriptions. We defined whether or not a patient's treatment could be optimised based on the number of medication classes they were prescribed, taking into account their allergies and other contraindications: beta-blockers should not be prescribed to asthmatics; beta blockers and calcium-channel blockers should not be prescribed in moderate-severe aortic stenosis (unless surgically repaired); thiazide-like diuretics should not be prescribed where the latest eGFR<30 ml/min/1.73m^2; diuretics should not be prescribed if the most recent serum sodium<130 mmol or potassium<2.5 mmol; all medications should be avoided in pregnancy except for labetalol or methyldopa; ace-inhibitors and angiotensin-receptor blockers should be prescribed if the most recent serum sodium <130 mmol or renovascular disease is present (unless surgically fixed). We defined maximal anti-hypertensive therapy as being prescribed four concurrent classes of medication,[4] or fewer if patients were unable to receive additional medication due to allergies or the above contraindications.

To identify whether patients had on-going contact with the health system, representing opportunities to improve their care, we classified contacts on any given date as follows: the presence of a code denoting direct contact with a patient (e.g. 'home visit', 'medication review with patient'), examination, symptom, history, diagnosis, or telephone contact counted as a 'direct contact'; everything else was classified as an 'indirect contact' including 'failed encounters', issuing of medication, or administrative activities such as reviewing letters from the hospital.

We categorized patients with high BP measurements needing repeat measurement as those who experienced a change in anti-hypertensive medication following the high reading. We also categorized those who had their high reading within 2 months of the final date of the data extract as requiring a repeat measurement, as this was considered a clinically reasonable period to re-check BP after advising lifestyle changes and therefore did not necessarily indicate sub-optimal management.

Algorithm accuracy study

Participants: We applied the final algorithm to 532,409 individual patient records from the whole Salford dataset to assess the quality of BP management in CKD patients between 2011 and 2012. Please see Figure 1 for a detailed breakdown of results. A total of 11,097 patients had CKD: 8,606 (72%) determined by diagnostic code and 2,491 (21%) determined by biochemical parameters. Of these, 1,435 were excluded from the quality standards: 1,226 (85%) due to the presence of a relevant exclusion code and 229 (15%) due to inferences that they were already prescribed maximal anti-hypertensive medication. Following their removal, over a half the remaining patients (4,940 out of 9,662; 51%) did not achieve quality standards for BP management: 2,072 did not receive a BP measurement and 2,868 had a BP measurement above target. Due to word count limitations, clinical and demographic characteristics of the study population are not presented here but are available on request from the authors.

Test results: Of the 2,072 without a BP measurement, 595 (29%) had ongoing direct contact with primary care, which suggests that their BP could be measured at their next visit. This could be used to create a reminder system for providers prior to the patient's next consultation. A further 64 (3%) had ongoing indirect contacts with primary care, suggesting they could be invited by letter, telephone or text message to arrange a BP measurement. This could be used to automatically send letters or text messages to these patients, or present providers with a list of the patients to telephone. Over two-thirds of this cohort (1,413; 68%) had no contact with primary care. This group is more difficult to draw conclusions about. They may require a different approach than simply inviting by letter, telephone or text message; this may be best guided by local provider knowledge. Of the 2,868 with a BP measurement above

target, 1329 (46%) patients did not receive any change in anti-hypertensive medication and may therefore benefit from optimization of their treatment. This could be used to create an alert system for providers to change these patients' medications; they could be contacted via the appropriate methods described above to implement these changes. A further 365 (13%) experienced changes to their medication regime (which may have included patient non-adherence) but did not have a repeat BP measurement to check the consequences of the change. Of these patients, 305 (84%) and 56 (15%) of patients had ongoing direct and indirect contact with primary care respectively, which represented missed opportunities. This information could again be used as a basis for provider reminders, lists or automatic methods of communication to patients to facilitate BP monitoring. Finally, 1,174 (41%) had their BP measured in the last 2 months of the study period, who also require repeat measurements but may not necessarily represent sub-optimal care. This information could be used as described above when the 2-month period has expired, but in the meantime should not be used as an indicator of suboptimal care.

Estimates: 3,766 patients were considered 'positive' by the algorithm (i.e. did not receive care in accordance with best practice but whose care could be improved upon). On manually reviewing the coded information in 100 randomly selected full patient records from 35 different primary care providers within this cohort (not used in the development of the algorithm), clinicians agreed that 90 were truly 'positive'. No cases needed further discussion. The 10 false positives were deemed to have been incorrectly classified as CKD patients by the algorithm.

Discussion

Summary of findings

This study developed and estimated the accuracy of an algorithm that can search EHR data to feed back practical, actionable information to clinicians regarding how they could improve care for patients in accordance with clinical guidelines. We used BP management in CKD patients as an exemplar and achieved a PPV of 90% (95% CI 82%-95%). Following the application of our algorithm we found 4,490 (51%) CKD patients did not achieve BP quality standards, and 3,766[†] (76%) of these patients had obvious room for improvement in their care – either by optimizing medication or taking opportunities to monitor BP. This information has greater utility than current informatics tools that provide audit and feedback using EHR data, because it provides actionable detail for practitioners to follow rather than simply reporting proportions of patients achieving standards of care[11,14,24–28]. The digital artefacts of this kind of retrospective, continuous analysis could provide the missing link between clinical audit and decision support – reducing the latency between clinical quality information and practitioners' improvement actions/tactics for patients with long-term conditions.

Strengths and weaknesses of this study

A major strength of this study is the thorough process through which the algorithm rules were developed. We based the rules on widely accepted clinical guidance, and used our clinical experience to ensure that the algorithm assessed care quality according to both robust research evidence and the clinical service context. These were further refined to ensure their accuracy by manually reviewing 440 different patient records. In addition, we used the totality of patient records to gain a representative picture of how care is delivered across the local population.

A weakness of this study is our reliance on the coded data within the EHRs to validate the algorithm. We were unable to gain access to the full patient EHR, and it is possible that free-text within the records not reflected in the coded data would detail reasons why some of the patients identified by the algorithm did not achieve the BP quality standards. However, many official processes in UK primary care such as health service planning and pay-for-performance rely on coded data, so any important information not entered as codes into EHRs is likely to be minimal.

We did not calculate the negative predictive value (NPV) of the algorithm due to restrictions on the number of patient records we could reasonably review manually. The prevalence of CKD patients that could have their BP management improved is relatively low in the general population and would therefore require large sample sizes to achieve an acceptably narrow confidence interval. However, as this algorithm is intended as a screening tool to lead to further investigation of patient records, it is arguable that a high PPV is the most important statistic to avoid over-burdening clinicians with false positive results.

[†] Taking into account the PPV of 90% this is 3,389.

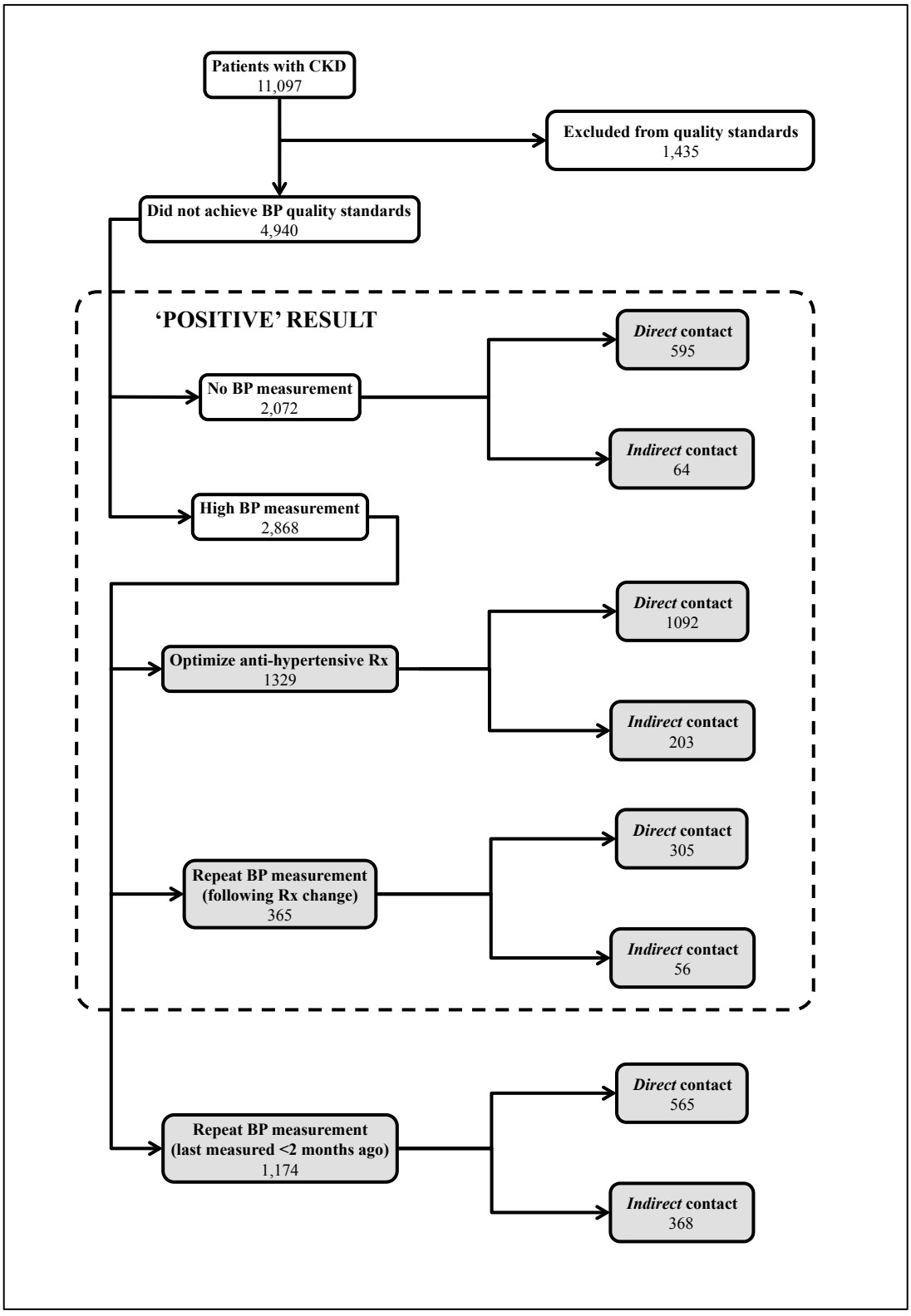


Figure 1. Flow of patients identified by the algorithm.

Key: BP: blood pressure (mmHg); CKD: chronic kidney disease; Rx: medication prescription; shaded areas represent actionable care improvement information for practitioners.

Comparison with existing literature

The algorithm we developed addresses a need identified in the literature that audit and feedback should provide information to clinicians that is actionable, in order to drive effective improvements in patient care.[22,23] Few informatics systems currently use EHR data in an automated fashion to provide audit and feedback in this way,[11,14,24–28] making this algorithm a relatively unique quality improvement tool.

Our finding that 49% of UK CKD patients received NICE quality standards regarding BP management is in broad agreement with other studies. De Lusignan et al. found this proportion to be 48-50%,[14] whereas Fraser et al. found it to be 58%.[16] This agreement suggests our algorithm has external validity.

The PPV of our algorithm is higher than others that use EHR data reported in the literature. Algorithms developed by Singh et al. to detect diagnostic errors in primary care achieved a PPV of between 5-21%;[42] Murphy et al. developed algorithms to detect delayed cancer diagnoses and reported PPVs of 58-70%;[43] and Brenner et al. reported a PPV of 15% in their algorithm to identify adverse drug events.[44] These discrepancies are likely explained by differing prevalence and ease of detection from EHR data of the events under examination. Furthermore, because we were unable to view the free-text within the patient records, the true PPV of our algorithm may be lower than 90%.

Implications for clinical practice and research

The algorithm developed in this study can form the basis of an audit and feedback tool to improve BP care for patients with CKD. Clinical performance systems that provide practitioners with actionable information on how to improve care, such as this, are more effective than those that simply provide proportions of patients meeting quality standards.[22,23] Thus the algorithm could enable practitioners to reduce cardiovascular morbidity and mortality – and as nearly all UK primary care providers use EHRs that capture Read codes, it could be rapidly deployed.

By relating care quality analyses to a practitioner's individual patients in this way, such a system may also improve clinical coding. This effect may reach beyond the specific codes concerned to a more general increase in coding accuracy as the gap between care quality information and improvement-actions is closed.

This study has tested the feasibility of a method that draws clinical process and outcome measurement closer together, which is likely to improve BP control and CKD outcomes. Our approach could be abstracted to other chronic conditions that use both processes and outcomes, which include (but are not limited to) other longitudinal physiological parameters (e.g. cholesterol, glycated hemoglobin).[45] The method could also be applied to other healthcare settings that employ different coding languages in their EHRs, such as hospitals and different countries.

Future work

Our future work will initially focus on improving the performance of the algorithm (particularly with regard to the diagnosis of CKD) and assessing the accuracy of the actionable information it provides. We aim to compare this with manual reviews of full patient records that include free-text. We also hope to improve the functionality of the algorithm to identify total daily doses of medication to provide more specific advice on how treatment could be optimized, and to assess medication adherence. Ultimately our vision is to use this algorithm to form the basis of an audit and feedback informatics tool by integrating it into software previously developed by our group.[46] We anticipate that the tool could be linked to patients' EHRs and provide clear and succinct messages to providers on how to improve care, taking into account comorbidities, allergies and other relevant patient characteristics. We believe that local clinical ownership of quality improvement projects is fundamental to success, but that clinicians may not have the expertise to develop and implement the informatics tools necessary. We will therefore work with local clinicians to understand the needs of their populations and ensure there is clinical engagement with our work. We are also mindful that implementing such tools into practice must align with existing clinical workflows to be effective and efficient. Therefore our approach to development and implementation of the audit and feedback tool will be iterative, and employ mixed methods techniques informed by relevant theoretical frameworks.[47] The effectiveness of the tool will eventually be tested in a randomized clinical trial.

Conclusion

We have developed an algorithm that uses EHR data to provide practical, actionable information to clinicians on how to improve care – a 'rear view mirror' for long-term conditions management. We used BP management in CKD patients as an exemplar and achieved a useful PPV of 90%. The utility of this information might be improved by mining the related EHR narrative, and by instrumenting clinical coding behavior. In future work we will develop

this approach further in variety of conditions and settings, measuring its effects on clinical behavior and patient outcomes.

References

1. Institute of Medicine. Crossing the Quality Chasm: A New Health System for the 21st Century. Washington D.C.: National Academy Press; 2001.

2. Matsushita K, van der Velde M, Astor BC, Woodward M, Levey AS, de Jong PE, et al. Association of estimated glomerular filtration rate and albuminuria with all-cause and cardiovascular mortality in general population cohorts: a collaborative meta-analysis. Lancet. 2010;375(9731):2073–81.

3. Kidney Disease: Improving Global Outcomes (KDIGO) Blood Pressure Work Group. KDIGO Clinical Practice Guideline for the Management of Blood Pressure in Chronic Kidney Disease. Kidney inter, Suppl. 2012;2:337–414.

4. NICE. Chronic kidney disease: Early identification and management of chronic kidney disease in adults in primary and secondary care [CG73]. London; 2008.

5. Heerspink HJL, Ninomiya T, Huxley R, Vlado Perkovic. Cardiovascular effects of blood pressure lowering in patients with chronic kidney disease. Westmead; 2013 p. 1–15.

6. SIGN. Diagnosis and management of chronic kidney disease: A national clinical guideline. Edinburgh; 2008.

7. National Kidney Foundation. K/DOQI Clinical Practice Guidelines for Chronic Kidney Disease: Evaluation, Classification and Stratification. Am J Kidney Dis. 2002;39:S1–S266.

8. Joint Specialty Committee on Renal Medicine. Chronic kidney disease in adults--UK guidelines for identification, management and referral. Nephrology, dialysis, transplantation : official publication of the European Dialysis and Transplant Association - European Renal Association. London; 2006.

9. Tonelli M, Muntner P, Lloyd A, Manns BJ, Klarenbach S, Pannu N, et al. Risk of coronary events in people with chronic kidney disease compared with those with diabetes: a population-level cohort study. Lancet. 2012;6736(12).

10. Bonds DE, Hogan PE, Bertoni AG, Chen H, Clinch CR, Hiott AE, et al. A multifaceted intervention to improve blood pressure control: The Guideline Adherence for Heart Health (GLAD) study. Am Heart J. 2009;157(2):278–84.

11. Greenberg JO, Vakharia N, Szent-Gyorgyi LE, Desai SP, Turchin A., Forman J, et al. Meaningful measurement: developing a measurement system to improve blood pressure control in patients with chronic kidney disease. J Am Med Informatics Assoc. 2013;1–5.

12. Tuot DS, Plantinga LC, Hsu C, Powe NR. Is awareness of chronic kidney disease associated with evidence-based guideline-concordant outcomes? Am J Nephrol. 2012;35(2):191–7.

13. Sarafidis P a, Li S, Chen S-C, Collins AJ, Brown WW, Klag MJ, et al. Hypertension awareness, treatment, and control in chronic kidney disease. Am J Med. 2008;121(4):332–40.

14. Lusignan S De, de Lusignana S, Gallagher H, Jones S, Chan T, van Vlymen J, et al. Audit-based education lowers systolic blood pressure in chronic kidney disease: the Quality Improvement in CKD (QICKD) trial results. Kidney Int. 2013;84(3):609–20.

15. Karunaratne K, Stevens P, Irving J, Hobbs H, Kilbride H, Kingston R, et al. The impact of pay for performance on the control of blood pressure in people with chronic kidney disease stage 3-5. Nephrol Dial Transplant. 2013;28(8):2107–16.

16. Fraser SDS, Roderick PJ, McIntyre NJ, Harris S, McIntyre CW, Fluck RJ, et al. Suboptimal blood pressure control in chronic kidney disease stage 3: baseline data from a cohort study in primary care. BMC Fam Pract. 2013;14:88.

17. Ravera M, Noberasco G, Weiss U, Re M, Gallina AM, Filippi A, et al. CKD awareness and blood pressure control in the primary care hypertensive population. Am J Kidney Dis. 2011;57(1):71–7.

18. Leonardis D, Mallamaci F, Enia G, Postorino M, Tripepi G, Zoccali C. The MAURO study: baseline characteristics and compliance with guidelines targets. J Nephrol. 2012;25(6):1081–90.

19. Van Zuilen D, Blankestijn PJ, van Buren M, Ten Dam MJ, Kaasjager KH, Ligtenberg G, et al. Hospital specific factors affect quality of blood pressure treatment in chronic kidney disease. Neth J Med. 2011;69(5):229–36.

20. Jha V, Garcia-Garcia G, Iseki K, Li Z, Naicker S, Plattner B, et al. Chronic kidney disease: global dimension and perspectives. Lancet. 2013;382(9888):260–72.

21. Brown B, Williams R, Ainsworth J, Buchan I. Missed Opportunities Mapping: Computable Healthcare Quality Improvement. Stud Health Technol Inform. 2013;192:387–91.

22. Ivers N, Jamtvedt G, Flottorp S, Jm Y, Sd F, Ma OB, et al. Audit and feedback: effects on professional practice and healthcare outcomes. Cochrane Database Syst Rev. 2012;(6).

23. Hysong SJ. Meta-analysis: audit and feedback features impact effectiveness on care quality. Med Care. 2009;47(3):356–63.

24. Smith K a, Hayward R a. Performance measurement in chronic kidney disease. J Am Soc Nephrol. 2011;22(2):225–34.

25. Gillam SJ, Siriwardena AN, Steel N. Pay-for-performance in the United Kingdom: impact of the quality and outcomes framework: a systematic review. Ann Fam Med. 2012;10(5):461–8.

26. Persell SD, Thompson JA, Baker DW. Improving Hypertension Quality Measurement Using Electronic Health Records. Med Care. 2009;47(4):388–94.

27. Krein SL, Hofer TP, Kerr E, Hayward R. Whom should we profile? Examining diabetes care practice variation among primary care providers, provider groups, and health care facilities. Health Serv Res. 2002;37(5):1159–80.

28. Kerr E a, Krein SL, Vijan S, Hofer TP, Hayward R a. Avoiding pitfalls in chronic disease quality measurement: a case for the next generation of technical quality measures. Am J Manag Care. 2001;7(11):1033–43.

29. NHS CfH. UK Terminology Centre: Read Codes. NHS Connecting for Health; 2012. Available from: http://www.connectingforhealth.nhs.uk/systemsandservices/data/uktc/readcodes/

30. Gnani S, Majeed A. A user's guide to data collected in primary care in England. Cambridge; 2006.

31. Hripcsak G, Bakken S, Stetson PD, Patel VL. Mining complex clinical data for patient safety research: a framework for event discovery. J Biomed Inform. 2003;36(1-2):120–30.

32. HSCIC. Quality and Outcomes Framework. Primary Care: Support and Guidance; 2013. Available from: http://www.hscic.gov.uk/qof

33. R Core Team. R: A language and environment for statistical computing. Vienna; 2014.

34. NICE. Chronic kidney disease quality standard: Quality statement 5: Blood pressure control Quality. London; 2011.

35. NICE. Type 2 diabetes: The management of type 2 diabetes [CG87]. London; 2009.

36. NICE. Hypertension: Clinical management of primary hypertension in adults [CG127]. London; 2011.

37. NICE. Management of stable angina [CG126]. London; 2011.

38. NICE. MI – secondary prevention: Secondary prevention in primary and secondary care for patients following a myocardial infarction [CG172]. London; 2013.

39. NICE. Stroke: Diagnosis and initial management of acute stroke and transient ischaemic attack (TIA) [CG68]. London; 2008.

40. NICE. Lower limb peripheral arterial disease: diagnosis and management [CG147]. London; 2012.

41. NICE. National Institute for Health and Care Excellence. NICE; 2013. Available from: http://www.nice.org.uk/

42. Singh H, Giardina TD, Forjuoh SN, Reis MD, Kosmach S, Khan MM, et al. Electronic health record-based surveillance of diagnostic errors in primary care. BMJ Qual Saf [Internet]. 2012;21(2):93–100.

43. Murphy DR, Laxmisan A, Reis BA, Thomas EJ, Esquivel A, Forjuoh SN, et al. Electronic health record-based triggers to detect potential delays in cancer diagnosis. BMJ Qual Saf. 2013;0:1–9.

44. Brenner S, Detz A, López A, Horton C, Sarkar U. Signal and noise: applying a laboratory trigger tool to identify adverse drug events among primary care patients. BMJ Qual Saf. 2012;21(8):670–5.

45. Mainz J. Defining and classifying clinical indicators for quality improvement. Int J Qual Heal Care. 2003;15(6):523–30.

46. Ainsworth J, Buchan I. COCPIT: A Tool for Integrated Care Pathway Variance Analysis. Stud Health Technol Inform. 2012;180:995–9.

47. Murray E, Treweek S, Pope C, MacFarlane A, Ballini L, Dowrick C, et al. Normalisation process theory: a framework for developing, evaluating and implementing complex interventions. BMC Med. 2010;8(1):63.

The Challenges of Creating a Gold Standard for De-identification Research

Allen C. Browne, MS, Mehmet Kayaalp, MD, PhD,
Zeyno A. Dodd, PhD, Pamela Sagan, RN, Clement J. McDonald, MD
Lister Hill National Center for Biomedical Communications,
U.S. National Library of Medicine, National Institutes of Health, Bethesda, MD

Abstract

We created a Gold Standard corpus comprised over 20,000 records of annotated narrative clinical reports for use in the training and evaluation of NLM Scrubber, a de-identification software system for medical records. Our experience with designing the corpus demonstrated the conceptual complexity of the task.

Introduction

At NLM we have developed a de-identification software system for medical records called NLM Scrubber intended to automatically de-identify clinical reports in compliance with the Privacy Rule of Health Insurance Portability and Accountability Act (HIPAA).[1] In the course of developing the scrubber, we needed to create a manually annotated corpus of medical records to serve as a "Gold Standard" for testing and evaluation. Annotation is needed to demark identifiers that should be found and scrubbed by the de-identification system and to provide enough information to facilitate both evaluation and further development. The nature of the markup in this corpus determines the kind of evaluation that can be undertaken. Ultimately, our corpus consisted of a set of 21,849 tagged clinical narrative reports. In each report, the identifiers meeting the HIPPA requirements for PII (personally identifying information) had to be hand-tagged. We had to make three types of decisions in the process of developing the tag-set. First, the HIPAA rules had to be interpreted and applied to the actual items found in the records. Second, we needed to identify the items themselves as well as their boundaries and third the internal structure of the items needed to be considered. This paper discusses our approach used in this process, outlines our conclusions and discusses alternatives.

Applying the HIPPA Rule

The HIPPA Privacy Rule requires that clinical documents be stripped of personally identifying information before they can be released to researchers and others. HIPAA Privacy Rule describes 18 identifiers that should be scrubbed in the de-identification of medical records.

Some of these descriptions seem quite specific; For example, "[all] elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older." But even in this seemingly well-defined rule, there is room for interpretation. Do partial dates without mention of the year count as dates e.g. "July" or "July 23rd"; do special days like "Christmas" or "New Year's" count as dates?

The Boundaries and Structure of identifiers

Identifiers do not always appear in discrete packages with clear borders. Are titles like "Mrs.", "Dr.", "Col." or "Adm." part of the name? What about name suffixes like "Jr." or "III" or titles like "MD", "Ph.D." or "Esq."? Some of them, such as "Mrs.", do not have much identifying information value; whereas, others such as "Adm." May have because of their occurrence in the population. Identifiers can also be conjoined in ways that obscure their structure. "The 5th, 6th and 18th of June 1965" seems to contain three full dates but it also contains lexical material that is not really part of any date, "and" in this case. The question/process of delimiting items is further complicated by the internal structure of those items.

Table 1. Per HIPAA Privacy Rule, the following identifiers must be deleted from PHI to fully de-identify health information. (*) As of 2010, there were 18 sets of zip codes with distinct initial three digits whose corresponding population sizes were less than or equal to 20,000.[2]

1. Names
2. All geographic subdivisions smaller than a state, except the first two digits of the zip code of the postal address. The third digit of the zip code can also be left intact, only if the size of the population in the area of the censored two digits is greater than 20,000 according to the most recent census data.(*)
3. All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older.
4. Telephone numbers.
5. Fax numbers.
6. Electronic mail addresses.
7. Social security numbers.
8. Medical record numbers.
9. Health plan beneficiary numbers.
10. Account numbers.
11. Certificate/license numbers.
12. Vehicle identifiers and serial numbers, including license plate numbers.
13. Device identifiers and serial numbers.
14. Web universal resource locators (URLs).
15. Internet Protocol (IP) address numbers.
16. Biometric identifiers, including fingerprints and voiceprints.
17. Full-face photographic images and any comparable images.
18. Any other unique identifying number, characteristic, or code, except the ones that may be generated by the covered entity for re-identification.

Are single letter initials inside a full name significant, e.g. "Q" in "John Q. Public"? In other words, if the scrubber only misses a middle initial, would the middle initial reveal the identity of the person? If not, should we tag or not tag the middle initials?

Methods

As described in Kayaalp, et. al. 2014,[2] we selected a random sample of patient records from the NIH Clinical Center using a method that prevented duplicate records. The selection involved randomly choosing 7,571 patients, collecting all of their records and removing duplicate records. A linguist and a registered nurse on our research team used VTT (Visual Text Tagger), a freely available text tagging system developed at NLM[3] to annotate PII in each record.[3,4] VTT uses a stand-off method to annotate texts so that both the original text and its formatting are preserved.[5] VTT facilitates tagging by allowing the human tagger to select an area of text by smearing the cursor over it and then choose a tag listed in from a drop down menu. It also provides a visual display of the tagged document representing each tag in a distinct visual format. VTT stores these annotated documents in a pure-ASCII machine readable format. In Figure 1 below, a mock-up version of the sort of records that make up our corpus is shown as displayed in VTT.

At the beginning of the process the human annotators tagged overlapping sets of records and came to a consensus on the results. The annotators conferred on specific questions as they worked through different sets of records. Then different sets of records were assigned to each annotator. Organized by patient, all the records of a particular patient would be completed by the same annotator in succession.

We tagged personal names, dates, addresses, ages and alphanumeric identifiers.

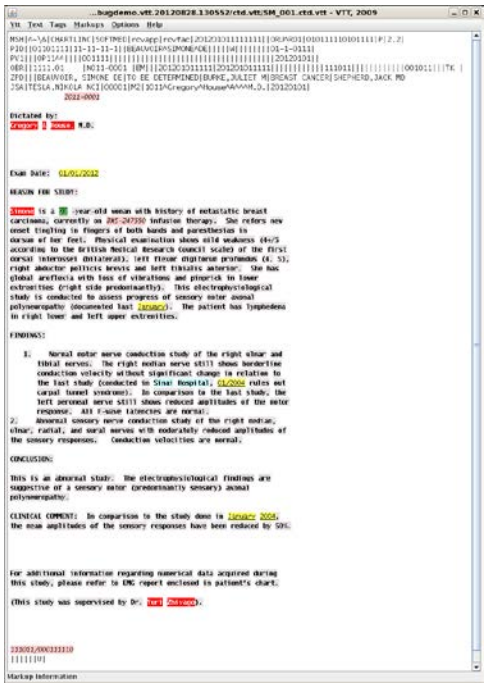


Figure 1

Names

HIPAA specifies that names must be redacted when they refer to patients and their relatives, employers and household members. Names of providers are not considered personally identifying, but almost all de-identifying programs scrub all personal names (care providers as well as patients), and it is probably the safest course because it can be difficult to distinguish between the two. We assigned the tag 'Name' to all personal names in our corpus. VTT allows tags to have subtags further refining the meaning of the tag. This distinction between identifiers related to patients and those not related to patients carries through a number of different tags. A 'Patient' subtag added to the 'Name' tag was used to indicate the names of patients as opposed to the names of hospital personnel and other people mentioned in the records. Names of relatives and members of the patient's household also received the 'Patient' subtag. Names were tagged as whole; that is the entire string "John Q. Public" would be given a single 'Name' tag. Suffixes like "III" or "Jr." were considered part of the personal name and tagged accordingly, but titles like "Mr.", "Dr.", or "Col.", were not unless their occurrences were rare. Initials seen in a longer name string were included in the name. Single first-names and last-names standing alone, e.g. 'John' by itself and "Public" in "Mr. Public", also received the 'Name' tag. Initials standing alone in the text, e.g. 'JQP' or 'JP', were marked 'PNinit' and separated into patient and non-patient using the 'Patient' subtag. While OCR guidelines states that initials are considered names,[6] we separated them because we felt initials were different and perhaps much less identifying than spelled out names. The vast majority of initials that we encountered denoted providers or transcriptionists who used them to sign the record. Names that were neither providers nor patients, for example names appearing as citations to the literature like "Greulich and Pyle", influential authors of articles regarding bone age, or the name of the current president when evaluating a patient's orientation were not tagged at all. These decisions introduced a difficulty in evaluation in those names like 'Greulich' and were counted as false positives when the system labeled them as names.

Addresses

Whole address strings such as "905 Maple Street, Apartment 2, Littletown, Minnesota, 55021" were tagged 'Address'. Cities standing alone like "Baltimore" were tagged as addresses but states and countries, for example "Maryland" and "Argentina", were not in keeping with guidelines of the Privacy

Rule which say that units smaller than a state should be redacted, although Baltimore has a population of well over 20,000, the size limit for Zip-Codes. D.C. was considered a state for this purpose. State-like subdivisions of countries other than the U.S., like "Alberta", were treated as states. Specific locations within the hospital received the 'Location' tag. For example, "OP9" or "3-Northeast" or "Day Hospital" were considered locations.

In a later iteration, we divided the Address tag into 8 tags identifying the street address, unit number, city, state, country, ZIP-code, county, and kept the old address tag as a catch-all. This revision reflected the realization that not all errors in redacting addresses were alike in seriousness. By tagging Address strings whole we made it difficult at the time of evaluation to recognize that redacting "Maryland" and not "Baltimore" from the string "Baltimore, Maryland" would result a violation of HIPAA but redacting "Baltimore" and not "Maryland" would not.

Street address includes strings like "904 South Madison Street" as well as building names and numbers "The Dakota" and "Building 9". The unit number captures apartment and suite numbers. States include U.S. states and their equivalents in other countries, e.g. "British Columbia". ZIP-codes have not been observed to stand alone in our documents. They are individually tagged as part of a larger address. To clarify, a lengthy address like "907 Madison Street, suite #5, Silver Spring, Maryland, 20190, U.S.A" represents 6 entities. A street address "907 Madison Street", a unit number, "Suite #5", a city, "Silver Spring", a ZIP-code "20910", a state "Maryland" and a country "U.S.A". Only the states and the countries need not be redacted under HIPAA.

This breakdown of larger address tags allows us to improve our evaluation by facilitating a more granular understanding of how identifying partially redacted addresses might be. It opens up a better method of counting errors than either a binary question of whether the address was (completely) redacted or a simple count of how many tokens or what percent of the address was redacted. The best evaluation might count the number of identifying parts that are redacted or better yet we should consider an evaluation of how identifying the unredacted parts are. For example "907", "Suite #5" or even "Madison" alone without the rest of the address could hardly be considered identifying.

Alphanumeric Identifiers

Alphanumeric identifiers were defined as strings of letters and/or numbers used as identifiers, excluding those identifiers that are part of a personal name, address, date, and age. We divided them into three different types: communication identifiers, protocol numbers and other Alphanumeric Identifiers, receiving 'Comm', 'Prot', or 'Alphanum' tags respectively. HIPAA calls for all of these identifiers to be redacted. A 'Comm' tag was assigned to Communication identifiers included numbers such as telephone numbers, email addresses, URLs and the like. Protocol numbers were common in our corpus and have a fairly typical form. The remaining numbers comprised the Other Alphanumeric identifiers and came from a range of types, including sample numbers, blood unit numbers, radiologic ids and lab test numbers. Communication number may or may not pertain to a patient so 'Comm' tags can take the patient subtag depending on whether they pertain to the patient. The telephone number of a referring physician would be marked 'Comm' but without a 'Patient' subtag. Protocol numbers and other alphanumeric identifiers were all considered patient related and subtagged 'Patient'.

Ages

We used three tags 'Age-PII' and 'Age-NPII' and 'Age-fract' to mark ages found in the corpus. Age-PII identified ages 90 years and over, since HIPAA specifically requires that ages over 89 be redacted. Age-NPII was used to mark ages in years, less than 90 which are not PII. Ages less than a year e.g. "3 days", "2 ½", "fifth week of life" belong to a special case, because they were not singled out by HIPPA as PII but they certainly could be much more identifying than an age in whole years. In the case of 'Age-PII' and 'Age-NPII' only the numeric part of the age was tagged, e.g. in "patient is 56 years old", we only tagged "56". In the case of age ranges like "3-5 years" only "3" and "5" were tagged. Ages given as decades, e.g. "in his 60's", were not tagged because they represented so large an age span that they were not considered identifying. In the case of fractional ages, both the number and the unit of measure were included in the tag, for example "thirteen months" and 'three weeks' would be tagged 'Age-fract'. By

keeping the unit of measure we allowed the de-identification system to round these ages to a year or redact them completely. All three age tags were used to mark only the ages of patients, their relatives or household members. No provider ages were seen in our corpus. Gestational ages and bone ages were not tagged 'Age'.

Dates

Date strings were initially tagged whole in our first iteration. "Wednesday, June 14, 1996" was tagged as a date as were free standing date parts. Months like "February", days like "Thursday" and years like '1998' when standing alone in text were tagged 'Date'. Decades like "The 60's" and plural days, "Fridays" were not tagged because such long or repetitive dates were not considered identifying. By default all dates were considered to be relevant to patients. In date ranges like "June 3 – July 15" both the beginning and end points are tagged as dates, separately. An exception appears when part of the range is unable to stand on its own, e.g. "2005-6". In this case the whole string was tagged. We considered special days like "Christmas" or "Mother's Day" to be dates and tagged them as such since they are equivalent to a date like 'December 25". Although HIPAA requires redaction of all elements of Dates except the year, strings like "September 23rd" standing alone don't indicate a specific date unless they are in the context of a year. But since HIPAA allows years not to be redacted we should assume that there might be a year in context depending on how the de-identification run is configured. Days of the month standing alone are another matter since "the third" or "the third of the month" does not specify a particular date and would only do so in conjunction with a month name and a year. Since month names must be redacted under HIPAA, weekday names would not be particularly identifying.

Similar to our treatment of Address, in a later iteration of our tagging effort we broke Dates into their parts again facilitating a more granular evaluation. By tagging months, days and years separately in a long date string like "February 27th, 1991" we can not only take account of the fact that the substring "1991" need not be redacted at all, but we can also consider that "27th" without the month does little to identify that actual date even in the presence of the year.

This more granular approach to tagging also facilitates treatment of conjoined and otherwise obscured items. For example the date string "the 5th, 6th, and 18th of May" presents several difficulties for evaluation. The tokens "the", "and" and "of", though parts of the date string, are not really parts of the three dates represented here and might be ignored during evaluation. By tagging "May" as a month and "6th" as a day we will be able to recognize that "May" alone is identifying in a way that "6th" alone is not, even in the context of a year.

Revision of other tags

Although we have not yet moved to a more granular treatment of other identifiers, it would clearly help our evaluation to do so. In the course of evaluation we decided not to count single initials that are part of a name like the "H" in "William H. Macy" as a name token. That is, we did not consider it a false negative if it was left unredacted when the rest of the name was redacted. Although this situation seldom arose imprecise because the scrubber was generally able to recognize middle initials from their position between two name tokens. That decision points to a future re-tagging of the corpus to reflect the parts of full names. Similarly, telephone numbers might be sub-divided into the area code prefix and number. Area codes and prefixes have a geographical association and might be considered more identifying than the 4 digit number itself. HIPAA already contemplates the internal structure of ZIP-codes specifying that the initial three digits of the ZIP code need not be redacted if "The geographic unit formed by combining all the ZIP codes with the same three initial digits contains more than 20,000 or fewer people." Something similar might be applied to telephone area codes or prefixes.

Results

We annotated a total of 21,849 records, representing 7,571 patients. Of those records, 3093 were used for evaluation in our de-identification study[2] and 1,140 were used for training. In addition to the two iterations of tagging, errors found in the course of evaluation were fed back into the gold standard after review by the taggers.

Discussion

The creation of our gold standard represented a number of challenges as described above including the lack of clear definitions of redactable items. One of the main lessons to come out of the effort was the realization that a finer grained analysis of strings representing PII facilitates a better understanding of evaluation results and points to a better method of evaluation. Counting whole strings as either properly redacted or not does not take into account which parts of the string might be left unredacted. Using token counts in the calculation of sensitivity and specificity also has inherent drawbacks, especially when a singly revealed token is a part of a multi-token identifier such as “September 11, 2001.”The potential of a particular token to identify the patient is less clear than the potential of properly tagged parts of the whole string.

Another consideration not explored above is the possibility of tagging more than the PII in a file. We found our evaluation hindered by a lack of knowledge about which tokens could be removed without loss of clinically pertinent information. Examination of actual redacted records shows that precision based on the number of false positive tokens overestimates the loss of information in actual records. Natural language including the sublanguage of clinical records is sufficiently redundant so that the loss of tokens often does not result in a significant loss of readability. Some sections of the record will inevitably contain information not relevant to a particular medical task and loss of that information would not damage the usefulness of the record. We are exploring a methodology to identify and subsequently tag clinical information so as to rationalize future evaluation of precision in our de-identification effort. The challenge in this task would be similar; that is, how should we categorize clinical information and label it with finer granularity so that we can fairly measure the loss of clinical information?

Funding

This work was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

Competing Interests

The second author receives royalties from University of Pittsburgh for his contribution to a de-identification project. NLM’s Ethics Office reviewed and approved his appointment.

References

1. Kayaalp M, Browne AC, Callaghan FM, Dodd ZA, Divita G, Ozturk S, et al. The pattern of name tokens in narrative clinical text and a comparison of five systems for redacting them. J Am Med Inform Assn 2013.
2. Kayaalp M, Browne AC, Dodd ZA, Sagan P, McDonald CJ. De-identification of Address, Date, and Alphanumeric Identifiers in Narrative Clinical Reports. AMIA Fall Symposium, 2014.
3. Lu, Chris J; Divita Guy; Browne, Allen C. “Development of Visual Tagging Tool”. AMIA 2010 Annual Symposium, Washington, DC, November 13-17, 2010, p. 1156
4. National Library of Medicine. Visual tagging tool, 2010. URL: http://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/vtt/current/web/index.html. Accessed in 8/20/2013
5. Kayaalp, M. Separation of Data, Interpreters and Likelihood. Report number: LHNCB-TR-2007-001, Affiliation: Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, 2007.
6. Office of Civil Rights. Guidance regarding methods for de-identification of protected health information in accordance with health insurance portability and accountability act (HIPAA) privacy rule, 2012.

Adoption of Clinical Data Exchange in Community Settings: A Comparison of Two Approaches

Thomas R. Campion, Jr., Ph.D.[1,2,3,4], Joshua R. Vest, Ph.D., M.P.H.[1,3,4], Lisa M. Kern, M.D., M.P.H.[1,3,4,5], Rainu Kaushal, M.D., M.P.H.[1,2,3,4,5,6] and the HITEC investigators

[1]Department of Healthcare Policy and Research, Weill Cornell Medical College, New York NY; [2]Department of Pediatrics, Weill Cornell Medical College, New York NY; [3]Center for Healthcare Informatics and Policy, Weill Cornell Medical College, New York NY; [4]Health Information Technology Evaluation Collaborative, New York NY; [5]Department of Medicine, Weill Cornell Medical College, New York NY; [6]Komansky Center for Children's Health at NewYork-Presbyterian Hospital, New York NY

Abstract

Adoption of electronic clinical data exchange (CDE) across disparate healthcare organizations remains low in community settings despite demonstrated benefits. To expand CDE in communities, New York State funded sixteen community-based organizations to implement point-to-point directed exchange (n=8) and multi-site query-based health information exchange (HIE) (n=8). We conducted a cross-sectional study to compare adoption of directed exchange versus query-based HIE. From 2008 to 2011, 66% (n=1,747) of providers targeted for directed exchange and 21% (n=5,427) of providers targeted for query-based HIE adopted CDE. Funding per provider adoptee was almost two times greater for directed exchange (median (interquartile range): $25,535 ($17,391-$42,240)) than query-based HIE ($14,649 ($9,897-$28,078)), although the difference was not statistically significant. Because its infrastructure can cover larger populations using similar levels of public funding, query-based HIE may scale more broadly than directed exchange. To our knowledge, this is among the first studies to compare directed exchange versus query-based HIE.

Introduction

Electronic clinical data exchange (CDE) across disparate healthcare organizations has the potential to improve care delivery, strengthen public health efforts, and reduce costs (1,2). For large academic medical centers, sharing clinical data with other institutions has been associated with decreased laboratory test ordering (3) and emergency department charges (4). Additionally, regional efforts promoting CDE usage in emergency departments have been associated with reduced costs (5,6). Despite the demonstrated benefits of CDE, widespread adoption remains low (7), particularly in smaller community settings (8).

To increase adoption of CDE, multiple organizational and technological options exist. Non-profit health information organizations (5), accountable care organizations (9), market-oriented health information technology vendors (10), incentives-driven independent healthcare practitioners (11), and health-minded consumers are pursuing different clinical goals, financial motivations, and administrative structures for CDE. Technological approaches for CDE include point-to-point directed exchange of clinical data such as laboratory results and patient referrals, query-based health information exchange (HIE) for aggregating patient data from multiple sites, and consumer-mediated exchange using personal health records (12). The multiple forms of exchange address different but complementary clinical information needs (11).

Despite the diversity of available CDE approaches, studies comparing organizational and technological configurations for CDE are limited, and optimal approaches to CDE are unknown. In New York State, two types of community-based organizations received funding to implement CDE as part of an effort to create a statewide health information network; clinically affiliated groups of providers implemented directed exchange, and regional organizations implemented query-based HIE. The objective of this study was to compare adoption of and barriers to the two CDE approaches.

Background

New York State has 19 million residents and 91,918 healthcare providers, of which 75% (n=68,898) are physicians (medical doctors (MDs) and doctors of osteopathic medicine (DOs)) (75%) and 25% (n=23,020) are physician extenders (nurse practitioners (NPs) and physician assistants (PAs)) (13). Defined policies for CDE exist in New York State. Directed exchange of patient data does not require affirmative patient consent, as point-to-point data

transmission mimics existing fax and telephone-based clinical information sharing. However, a patient must opt in to query-based HIE participation by providing affirmative consent to each healthcare organization that wishes to access his or her data. Clinicians also have the ability to "break the glass" and access query-based HIE during emergencies.

HEAL NY

New York State has invested more than $800 million in health information technology through four phases of the Healthcare Efficiency and Affordability Law for New Yorkers (HEAL NY). Launched in 2006, HEAL 1 invested in community CDE capacity, electronic health record (EHR) adoption, and electronic prescribing. In 2008, HEAL 5 aimed to increase adoption of interoperable EHRs in communities as well as expand and connect multiple community-level CDE efforts to form the Statewide Health Information Network of New York (SHIN-NY). To develop and govern the SHIN-NY, New York State established the New York eHealth Collaborative (NYeC), a public-private partnership with oversight from the New York State Department of Health. Underway at the time of this writing are HEAL 10, which addresses EHR-based patient centered medical home implementation, and HEAL 17, which focuses on EHR-based care coordination; these phases also expand and facilitate statewide CDE. The current study focused on activities funded through HEAL 5.

HEAL 5

To expand CDE in communities, HEAL 5 awarded funding to sixteen organizations of two types: *community health information technology adoption collaborations (CHITAs)* comprised of community-based providers affiliated for care purposes but not under the same corporate umbrella and non-profit, non-governmental *regional health information organizations (RHIOs)* that convened and governed multiple community stakeholders in a defined geographic area. Whereas RHIOs were legally established entities, CHITAs were less formal alliances of providers with leadership from an organization such as a community hospital. CHITAs promoted adoption of directed exchange via software license purchase and implementation assistance of health information technology, particularly certified EHRs, that featured point-to-point interfaces for laboratory result delivery, electronic prescribing, quality reporting, and/or clinical transfer forms. RHIOs promoted adoption of query-based HIE via purchase of software licenses and implementation assistance for standalone web portals and/or EHR interfaces connected to federated or centralized repositories with master patient index, record locator, provider directory, user authentication, and patient consent management services.

CHITAs and RHIOs agreed to participate in a statewide collaboration process and adhere to state technology and policy standards. CDE activities of HEAL 5-funded organizations addressed up to three use cases aligned with state clinical and public health priorities that were based on Office of the National Coordinator for Health Information Technology (14) and Centers for Disease Control use cases (15). Organizations that received HEAL 5 funding also agreed to participate in evaluation activities conducted by the Health Information Technology Evaluation Collaborative (HITEC), the multi-institutional academic consortium charged with evaluating HEAL NY. Members of HITEC conducted this study.

Methods

Using data collected in 2010 and 2011, we conducted a cross-sectional study of organizations awarded funding through HEAL 5 to implement CDE in communities in New York State. Measures included adoption of CDE and barriers to implementation. We compared measures between CHITAs that implemented directed exchange and RHIOs that implemented query-based HIE. Understanding adoption of CDE by providers (16,17) and patients (18) as well as barriers to implementation (19) can inform best practices for CDE. The Institutional Review Board of Weill Cornell Medical College approved this study.

Data collection

To measure adoption of CDE, we obtained data in March 2011 from sources available to the research team. From NYeC, we obtained the number of providers (e.g. MD, DO, NP, PA) who adopted query-based HIE and the number of patients who affirmatively consented to query-based HIE as reported by each RHIO per state requirement. From unpublished survey data collected by the research team as part of general HEAL 5 evaluation activities (20), we obtained the number of providers who adopted EHRs or other health information technology with directed exchange as reported by each CHITA. From HEAL 5 grant applications, we obtained the number of providers and patients targeted for adoption by each organization as well as details about technology implementations. From the New

York State Department of Health website (21), we obtained funding amounts for each organization's HEAL 5 activities.

To measure barriers to implementation, we conducted a survey in November 2010. For each use case that an organization implemented, an organization's executive director (or executive director's designee) rated each of thirteen items as having served as a barrier, served as a facilitator, or had no effect. The thirteen items were based on a previous survey instrument developed and administered by members of the research team (22). If a survey respondent rated an item as a barrier for at least one use case, we considered the item as having served as a barrier for the respondent's organization. The survey contained four other sections not used in the current study addressing organization and governance, financial sustainability, statewide collaboration, and federal context.

Data analysis

We compared adoption of CDE and barriers to implementation between CHITAs and RHIOs. We determined descriptive statistics including mean (±standard deviation (SD)) for normally distributed data and median (interquartile range (IQR)) for non-normally distributed data. To examine differences between CHITAs and RHIOs, we used Fisher's exact tests for proportions, t-tests for normally distributed data, and Wilcoxcon rank-sum tests for non-normally distributed data. As in a prior study (23), the threshold for statistical significance was $p < 0.2$. To perform calculations, we used the R statistical software package.

Results

100% (n=16) of organizations awarded HEAL 5 funding in 2008 to implement CDE in communities participated in the survey (Table 1). Multiple types of entities led the directed exchange efforts of CHITAs while health information organizations exclusively implemented query-based HIE. Of CHITAs, seven implemented certified EHRs with interfaces, and one implemented a secure web application for facility-to-facility care transitions.

Table 1. Characteristics of organizations that implemented CDE. *Denotes $p < 0.2$

	Total (n=16)	**CHITAs (Directed) (n=8)**	**RHIOs (Query-Based HIE) (n=8)**
Leadership entity			
Health information organization, n (%)	9 (56)	1 (13)	8 (100)
Community hospital, n (%)	3 (19)	3 (38)	NA
Federally qualified health center, n (%)	2 (13)	2 (25)	NA
Long-term care facility, n (%)	1 (6)	1 (13)	NA
Municipal health department, n (%)	1 (6)	1 (13)	NA
Total funding in USD millions			
Total, n (%)	105.1	30.4 (29)	74.7 (71)
Organization, median (IQR)	(3.5-9.4)	2.9 (1.7-5.5)*	9.6 (7.3-12.3)*
Providers targeted for adoption			
Total, n (%)	28,094	2,631 (9)	25,463 (91)
Organization, median (IQR)	738.5 (327.5-2,775)	255 (71-450)*	2,850 (1,000-3,516)*
Patients targeted for adoption			
Total, n (%)	13,419,219	1,054,708 (8)	12,364,511 (92)
Organization, median (IQR)	385,000 (103,628-1,290,640)	95,752 (52,176-207,500)*	1,381,281 (986,500-2,275,712)*
Funding per target for adoption in USD			
Provider, median (IQR)	9,224 (4,309-15,798)	12,292 (10,682-25,836)*	3,914 (2,592-5,917)*
Patient, median (IQR)	13 (8-27)	27 (22-37)*	7 (5-9)*

Compared to RHIOs, CHITAs received 59% less total funding (p=0.003) and targeted 90% fewer providers (p=0.029) and 91% fewer patients (p=0.007) for adoption. However, CHITAs received more than twice as much funding per provider targeted for adoption (p=0.003) and almost three times as much funding per patient targeted for adoption (p=0.002) than RHIOs.

Adoption of CDE

Of all providers targeted by organizations for CDE adoption, 24% (n=7,174) adopted CDE. This included 66% (n=1,747) of providers targeted by CHITAs for directed exchange and 21% (n=5,427) of providers targeted by RHIOs for query-based HIE. As shown in Figure 1, the percentage of providers who adopted directed exchange

through CHITAs (mean (±SD): 70 (±42)) was twice as great as adopted query-based HIE through RHIOs (35 (±26)) (p=0.067). Of note, one CHITA achieved 140% CDE adoption after 25 more providers than the 62 targeted adopted directed exchange. Funding per provider adoptee was almost two times greater for CHITAs (median (IQR): $25,535 ($17,391-$42,240)) than RHIOs ($14,649 ($9,897-$28,078)), although the difference was not statistically significant (p=0.234).

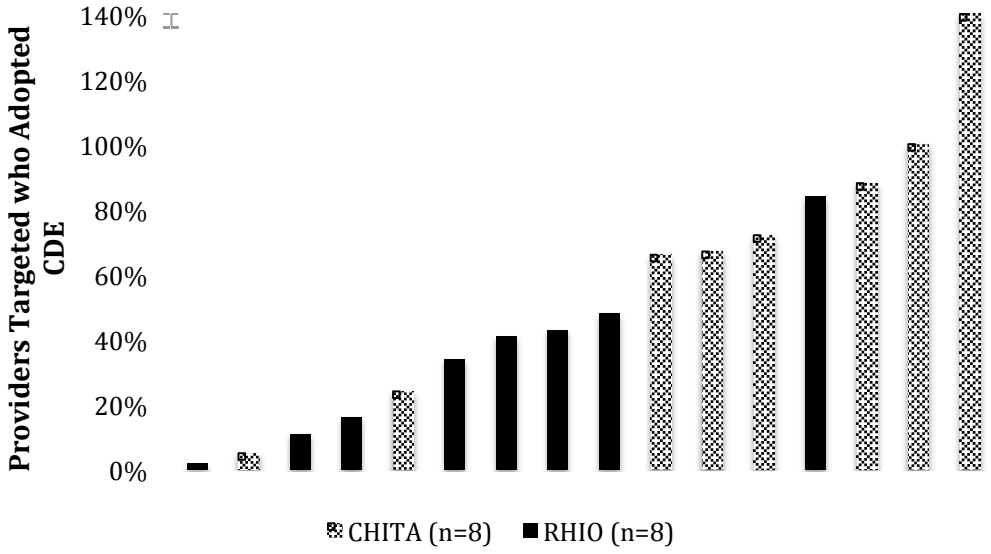


Figure 1. Percentage of providers who adopted directed exchange through CHITAs (n=8) and query-based HIE through RHIOs (n=8).

In total, RHIOs targeted more than 12 million patients to opt-in for query-based HIE adoption per New York State requirement. Of the 1,985,841 patients that RHIOs approached regarding query-based HIE participation during the study period, 1,787,257 (90%) provided affirmative consent, representing 14% of all patients targeted by RHIOs. On average, each RHIO obtained affirmative consent for query-based HIE participation from 16 (±11) percent of patients targeted for adoption (Figure 2).

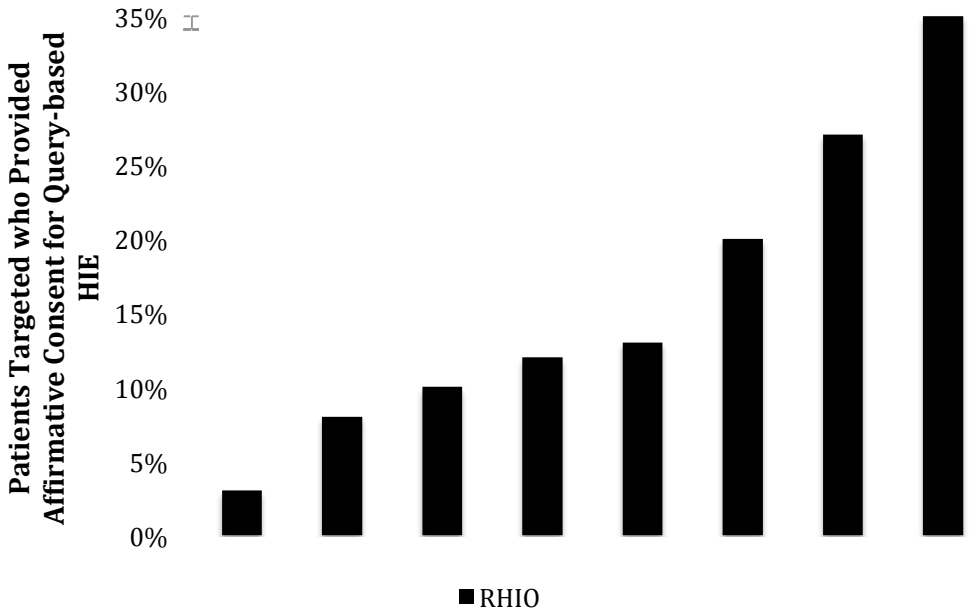


Figure 2. Percentage of patients who adopted query-based HIE through RHIOs (n=8).

Barriers to implementation

Organizations most frequently identified technology maturity (88%, n=14) and vendor participation (63%, n=10) as barriers to implementation (Table 2). Although CHITAs and RHIOs identified barriers similarly in most cases, a greater proportion of RHIOs (88%, n=7) than CHITAs (38%, n=3) identified vendor participation as a barrier to implementation (p=0.118). Vendors for RHIOs included Axolotl, dbMotion, Lawson, InterSystems, MedAllies, and MedPlus while vendors for CHITAs included Allscripts, eClinicalWorks, General Electric, MDLand, Medent, Meditech, NextGen, and Orion Health.

Table 2. Barriers to implementation of CDE. * Denotes p < 0.2

	Total (n=16)	**CHITAs (Directed) (n=8)**	**RHIOs (Query-Based HIE) (n=8)**
Technology maturity, n (%)	14 (88)	7 (88)	7 (88)
Vendor participation	10 (63)	3 (38)*	7 (88)*
Privacy and security policies	9 (56)	4 (50)	5 (63)
Regulatory requirements	9 (56)	3 (38)	6 (75)
Technical support	7 (44)	4 (50)	3 (38)
Organizational structure and culture	7 (44)	3 (38)	4 (50)
Workflow integration	7 (44)	2 (25)	5 (63)
Existing standards	7 (44)	2 (25)	5 (63)
Provider attitudes	4 (25)	2 (25)	2 (25)
Financial resources	4 (25)	2 (25)	2 (25)
Certification rules	4 (25)	2 (25)	2 (25)
Health plan participation	1 (6)	0 (0)	1 (13)

Discussion

This study is among the first to compare adoption of and barriers to directed exchange and query-based health information exchange. Through two types of community organizations, more than 7% of providers and about 9% of patients in New York State adopted CDE between 2008 and 2011. Although a greater percentage of providers adopted directed exchange per CHITA than query-based HIE per RHIO, more than three times as many providers adopted query-based HIE from RHIOs than directed exchange from CHITAs and with New York State funding per provider that was about 50% lower, although not statistically different. Barriers to implementation most frequently identified by organizations were technology maturity and vendor participation, with RHIOs more frequently identifying vendor participation as a barrier than CHITAs.

Federal and state investments have aimed to correct market failure in CDE adoption, but public funding for directed exchange may be to the detriment of existing activities and the public good provided by RHIOs (10). Although financial sustainability of RHIOs is unknown (7) and policymakers expect directed exchange to require less public funding than query-based HIE (10), findings from this study suggest that public funding per provider for query-based HIE may be similar to or lower than public funding per provider for directed exchange. Current federal policy encourages directed exchange that relies on technologies such as provider directories, patient matching, and secure data transport (12), which RHIOs typically already provide for query-based HIE. However, current federal policy does not incentivize use of RHIOs and instead promotes states and commercial entities to build and deliver such services. RHIOs may enable the benefits of CDE to scale more broadly than directed exchange because query-based HIE infrastructure can cover larger populations of providers and patients with similar levels of public funding and more complete data.

As members of Congress question current federal health information technology adoption programs (24) and multiple professional societies request delays to meaningful use deadlines (25), policymakers may find value in revisiting the role of query-based HIE and RHIOs. Existing technology infrastructure for query-based HIE provides a platform for quality improvement, population health management, clinical research, and other rich uses (26). For at least one RHIO, query-based HIE has facilitated care transitions in support of stage two of meaningful use (27), a service other RHIOs can offer to align with provider incentives. Additionally, RHIOs provide a forum for stakeholders to develop trust and govern exchange efforts. For example, health information organizations serving rural areas in the Rocky Mountains and Great Plains recently aligned to support the use of query-based HIE for meeting regional and inter-state patient data needs (28). Query-based HIE and RHIOs add technological and organizational value for CDE absent from current approaches to directed exchange.

Critical attitudes toward vendor-based CDE technology observed in this investigation are consistent with previous reports. In a recent national survey, a quarter of RHIOs expressed concern about "poor customer service, configuration and implementation issues, and cost" of vendor systems for CDE (29). Additionally, small practices and small hospitals have lagged in EHR and CDE adoption (30,31), indicating a need for improved vendor technologies and services in community settings (30,32). Accelerated adoption of CDE, and query-based HIE in particular, will likely require improved relationships between vendors and community organizations.

This study has limitations. First, our sample size was small and statistical comparisons were of borderline significance. Second, self-reported data from community organizations may be subject to respondent bias. Unlike our previous investigations that analyzed HIE system log files (27,33), the current study does not illustrate actual sharing of patient data. Third, unmeasured barriers and facilitators may have affected CDE implementation. Specifically, funding from non-HEAL NY sources may have contributed to adoption of CDE. Additionally, respondents may have interpreted the meaning of barriers differently due to potentially ambiguous definitions. Regardless, results of this study suggest a need to test assumptions about the cost of CDE approaches, including directed exchange (12). Finally, findings may not generalize beyond New York State, as privacy, technology, financial, and other factors may vary.

Conclusion

To our knowledge, this study is the first to quantify differences in adoption of and barriers to directed exchange and query-based HIE. Using similar levels of public funding, query-based HIE may enable broader adoption of clinical data exchange than directed exchange.

Acknowledgements

This study was conducted with funding from the New York State Department of Health (NYS contract number C023699). The study was conducted as part of the Health Information Technology Evaluation Collaborative (HITEC), the multidisciplinary academic consortium charged with evaluating the effects of New York State's investment in health information technology. The authors would like to thank Renny V. Thomas, M.P.H. and Elizabeth R. Pfoh, M.P.H. for data collection assistance; Michael Silver, M.S. and Alison M. Edwards M.Stat. for statistical assistance; and Jessica S. Ancker, M.P.H., Ph.D., Stephen B. Johnson, Ph.D, and Joshua E. Richardson, Ph.D., M.L.I.S. for conceptual feedback. The authors have no financial interests to disclose.

References

1. Vest JR, Gamm LD. Health information exchange: persistent challenges and new strategies. J Am Med Inform Assoc. 17(3):288–94.
2. Walker J, Pan E, Johnston D, Adler-Milstein J, Bates DW, Middleton B. The value of health care information exchange and interoperability. Health Aff (Millwood). Suppl Web W5–10–W5–18.
3. Hebel E, Middleton B, Shubina M, Turchin A. Bridging the chasm: effect of health information exchange on volume of laboratory testing. Arch Intern Med. 2012 Mar 26;172(6):517–9.
4. Overhage JM, Dexter PR, Perkins SM, Cordell WH, McGoff J, McGrath R, et al. A randomized, controlled trial of clinical information shared from another institution. Ann Emerg Med. 2002 Jan;39(1):14–23.
5. Frisse ME, Johnson KB, Nian H, Davison CL, Gadd CS, Unertl KM, et al. The financial impact of health information exchange on emergency department care. J Am Med Inform Assoc. 19(3):328–33.
6. Tzeel A, Lawnicki V, Pemble K. The business case for payer support of a community–based health information exchange: a Humana pilot evaluating its effectiveness in cost control for plan. Am Heal Drug Benefits. 2011;4(4):207–16.
7. Adler-Milstein J, Bates DW, Jha AK. Operational Health Information Exchanges Show Substantial Growth, But Long-Term Funding Remains A Concern. Health Aff (Millwood). 2013 Jul 9;
8. Adler-Milstein J, DesRoches CM, Jha AK. Health information exchange among US hospitals. Am J Manag Care. 2011 Nov;17(11):761–8.
9. Devore S, Champion RW. Driving population health through accountable care organizations. Heal Aff. 2011/01/07 ed. 2011;30(1):41–50.
10. Lenert L, Sundwall D, Lenert ME. Shifts in the architecture of the Nationwide Health Information Network. J Am Med Inform Assoc. 19(4):498–502.
11. Kuperman GJ. Health-information exchange: why are we doing it, and what are we doing? J Am Med Inform Assoc. 18(5):678–82.

12. Williams C, Mostashari F, Mertz K, Hogin E, Atwal P. From the Office of the National Coordinator: the strategy for advancing the exchange of health information. Health Aff (Millwood). 2012 Mar;31(3):527–36.
13. State Health Facts | The Henry J. Kaiser Family Foundation [Internet]. [cited 2013 Jul 28]. Available from: http://kff.org/statedata/?state=NY
14. HITSP - AHIC Use Cases [Internet]. [cited 2013 Jul 30]. Available from: http://hitsp.wikispaces.com/AHIC+Use+Cases
15. CDC - IIS - Immunization Information Systems Homepage - Registry - Vaccines.
16. Goroll AH, Simon SR, Tripathi M, Ascenzo C, Bates DW. Community-wide implementation of health information technology: the Massachusetts eHealth Collaborative experience. J Am Med Inform Assoc. 16(1):132–9.
17. Grossman JM, Bodenheimer TS, McKenzie K. Hospital-physician portals: the role of competition in driving clinical data exchange. Health Aff (Millwood). 25(6):1629–36.
18. Wald JS. Variations in patient portal adoption in four primary care practices. AMIA Annu Symp Proc. 2010 Jan;2010:837–41.
19. Ross SE, Schilling LM, Fernald DH, Davidson AJ, West DR. Health information exchange in small-to-medium sized family medicine practices: motivators, barriers, and potential facilitators of adoption. Int J Med Inform. 2010 Feb;79(2):123–9.
20. Abramson E, Maniccia D, Edwards A, Moore J, Kaushal R. Electronic health record adoption and use among ambulatory care physicians in New York State. 2011;
21. HEAL NY Phase 5 - Advancing Interoperability and Community-wide EHR Adoption in New York State [Internet]. [cited 2013 Jul 30]. Available from: http://www.health.ny.gov/technology/projects/
22. Kern LM, Barron Y, Abramson EL, Patel V, Kaushal R. HEAL NY: Promoting interoperable health information technology in New York State. Health Aff (Millwood). 28(2):493–504.
23. Kern LM, Wilcox AB, Shapiro J, Yoon-Flannery K, Abramson E, Barron Y, et al. Community-based health information technology alliances: potential predictors of early sustainability. Am J Manag Care. 2011 Apr;17(4):290–5.
24. Thune J, Alexander L, Robert P, Burr R, Coburn T, Enzi M. REBOOT: Re-examining the Strategies Needed to Successfully Adopt Health IT [Internet]. 2013. Available from: http://www.thune.senate.gov/public/index.cfm/files/serve?File_id=0cf0490e-76af-4934-b534-83f5613c7370
25. MGMA Requests Delay of Meaningful Use Stage 2 Penalties - iHealthBeat [Internet]. [cited 2013 Aug 31]. Available from: http://www.ihealthbeat.org/articles/2013/8/22/mgma-requests-delay-of-meaningful-use-stage-2-penalties
26. Vest JR, Campion TR, Kaushal R. Challenges, alternatives, and paths to sustainability for health information exchange efforts. J Med Syst. 2013 Dec;37(6):9987.
27. Campion TR, Vest JR, Ancker JS, Kaushal R. Patient encounters and care transitions in one community supported by automated query-based health information exchange. AMIA Annu Symp Proc. 2013 Jan;2013:175–84.
28. Sixteen Health Information Organizations Join Forces As Founding Members Of The Mid-States Consortium Of Health Information Organizations [Internet]. [cited 2014 Feb 22]. Available from: http://business.itbusinessnet.com/article/Sixteen-Health-Information-Organizations-Join-Forces-As-Founding-Members-Of-The-Mid-States-Consortium-Of-Health-Information-Organizations-3070317
29. New Report Presents Mixed Picture of Vendors Helping Doctors Exchange Health Data [Internet]. [cited 2013 Jul 30]. Available from: http://www.ehidc.org/about-us/press/press-releases/727-new-report-presents-mixed-picture-of-vendors-helping-doctors-exchange-health-data-.html
30. Desroches CM, Worzala C, Bates S. Some hospitals are falling behind in meeting “meaningful use” criteria and could be vulnerable to penalties in 2015. Health Aff (Millwood). 2013 Aug 1;32(8):1355–60.
31. Hsiao C-J, Jha AK, King J, Patel V, Furukawa MF, Mostashari F. Office-based physicians are responding to incentives and assistance by adopting and using electronic health records. Health Aff (Millwood). 2013 Aug 1;32(8):1470–7.
32. Vest JR, Yoon J, Bossak BH. Changes to the electronic health records market in light of health information technology certification and meaningful use. J Am Med Inform Assoc. 20(2):227–32.
33. Campion TR, Edwards AM, Johnson SB, Kaushal R. Health information exchange system usage patterns in three communities: Practice sites, users, patients, and data. Int J Med Inform. 2013 Jun 3;82(9):820–810.

Examining the Use, Contents, and Quality of Free-Text Tobacco Use Documentation in the Electronic Health Record

Elizabeth S. Chen, PhD[1,2], Elizabeth W. Carter, MS[1], Indra Neil Sarkar, PhD, MLIS[1,3], Tamara J. Winden, MBA[4,6], Genevieve B. Melton, MD, MA[4,5]
[1]Center for Clinical & Translational Science, [2]Medicine,
[3]Microbiology & Molecular Genetics, University of Vermont, Burlington, VT;
[4]Institute for Health Informatics, [5]Surgery, University of Minnesota, Minneapolis, MN;
[6]Division of Applied Research, Allina Health, Minneapolis, MN

Abstract

Recent initiatives have emphasized the potential role of Electronic Health Record (EHR) systems for improving tobacco use assessment and cessation. In support of these efforts, the goal of the present study was to examine tobacco use documentation in the EHR with an emphasis on free-text. Three coding schemes were developed and applied to analyze 525 tobacco use entries, including structured fields and a free-text comment field, from the social history module of an EHR system to characterize: (1) potential reasons for using free-text, (2) contents within the free-text, and (3) data quality issues. Free-text was most commonly used due to limitations for describing tobacco use amount (23.2%), frequency (26.9%), and start or quit dates (28.2%) as well as secondhand smoke exposure (17.9%) using a variety of words and phrases. The collective results provide insights for informing system enhancements, user training, natural language processing, and standards for tobacco use documentation.

Introduction

Tobacco use continues to be the leading preventable cause of morbidity and mortality in the United States[1, 2]. Worldwide, direct tobacco use is responsible for more than 5 million deaths each year while exposure to secondhand smoke is responsible for over 600,000[3, 4]. Public health initiatives such as Healthy People 2020 Tobacco Use[5, 6], the Centers for Disease Control and Prevention's National Tobacco Control Program[7], the U.S. Preventive Services Task Force[8, 9], and the World Health Organization's Tobacco Free Initiative[10] involve efforts aimed at ending the tobacco epidemic through targeted prevention and treatment strategies for children, adolescents, and adults.

In the last five years, there has been increasing emphasis on the potential role of Electronic Health Record (EHR) systems for identification and treatment of tobacco use. Currently among the Centers for Medicare & Medicaid Services Meaningful Use Objectives[12] is a core measure focused on the recording of smoking status as structured data in the EHR using a specified set of SNOMED CT codes (e.g., for "Current every day smoker," "Former smoker," and "Never smoker")[13] and clinical quality measures for tobacco use screening and cessation intervention[14]. A recent Institute of Medicine report further highlighted the importance of capturing behavioral determinants of health in the EHR and specified nicotine use and exposure among the domains to consider for Stage 3 Meaningful Use[11]. A Cochrane Review of 11 studies that involved using the EHR to improve documentation or treatment of tobacco use found that there were modest improvements and concluded that additional research is needed to understand the role of EHRs in this context[15]. Among these studies was a demonstration project where workflow modifications included incorporating evidence-based prompts in the Epic EHR at Dean Health Systems for guiding the identification of current tobacco users, determining their willingness to quit, and offering a set of tobacco cessation interventions[16]. Recent studies have also described workflow changes such as incorporating medical assistants in the documentation and referral process[17] as well as decision support functionality such as alerts and pre-defined order sets[18].

Within the EHR, tobacco use, secondhand smoke exposure, and related interventions may be documented in various parts as structured data or free-text (e.g., problem list[19, 20], social history[21], medications[19, 22], clinical notes, or patient instructions). A number of efforts have focused on developing natural language processing (NLP) techniques to extract smoking status[23, 24] and tobacco cessation interventions (e.g., searching for the "5 A's" for tobacco treatment and prevention)[25, 26] from clinical notes such as discharge summaries. In a recent study, supplementing structured fields with information from free-text fields was found to substantially improve smoking status data in the EHR[27].

While previous efforts have focused on enhancing the EHR for smoking status and extracting this information from free-text clinical notes, there has been limited discussion on improving the collection of details about tobacco use

(e.g., amount and frequency) and exploring free-text tobacco use documentation throughout the EHR. To this end, the objective of the present study was to examine the use, contents, and quality of free-text comments for tobacco use in the primarily structured social history module of an EHR system. Potential implications of the findings include informing system enhancements, user training, NLP, and standards for tobacco use documentation that may ultimately contribute to improving tobacco use assessment and cessation interventions using the EHR.

Methods

Setting and Study Design

This study involved the retrospective analysis of information collected in the social history module of the Epic EHR (Epic Systems Corporation, Verona, WI)[28] at Fletcher Allen Health Care, the academic health center affiliated with the University of Vermont. This module can be used for primarily structured documentation of tobacco use, alcohol use, illicit drug use, and a range of other social history-related information, which may subsequently be used to pre-populate the social history section in clinical notes. At the time of this study, each tobacco use entry included a set of structured fields associated with smoking, another set of structured fields associated with smokeless tobacco use, and a free-text field for comments (Table 1). Of the 158,608 patients with information documented using the social history module in 2013, this free-text field was used for 18,221 (11.5%) patients where the average length of the comments was 24±19 characters (minimum = 1 and maximum = 255).

Table 1. Example Tobacco Use Entries.

Field	Example 1	Example 2	Example 3
Smoking status	Current Everyday Smoker	Former Smoker	Passive Smoker
Start date	-	-	-
Quit date	-	2/12/02 0:00	-
Types (*Cigarettes*, *Pipe*, or *Cigars*)	Cigarettes	Cigarettes	-
Packs/day	0.5	1	-
Years	15	11	-
Pack years*	7.5	11	-
Smokeless tobacco	Never Used	Current User	Unknown
Quit date	-	-	-
Types (*Snuff* or *Chew*)	-	Chew	-
Comment	Started smoking again in 2010 after quitting a few years	2 cans weekly	Parents smoke outside

** Calculated based on Packs/day and Years*

Three coding schemes were used to manually analyze tobacco use entries in order to characterize: (1) reasons for using the free-text comment field, (2) contents within this free-text field, and (3) data quality issues. The general approach for developing and applying each of these coding schemes (further described below) involved three phases: (1) generating initial coding schemes based on analysis of 100 tobacco use entries from September 2013 and enhancing the schemes using an iterative, consensus-based process involving individuals with expertise in clinical care and biomedical informatics (ESC, EWC, INS, TJW, and GMM); (2) calculating inter-rater reliability using the kappa statistic to ensure consistency in coding between two reviewers (ESC and EWC) using the final versions of each coding scheme for 50 entries from October 2013; and, (3) performing the main analysis on a random sample of 525 tobacco use entries from November 2013 by one reviewer (ESC) where this sample size was based on a total of 4,056 most recent entries for patients during this time, confidence level of 95%, and estimated precision of 4%.

Analysis of Potential Reasons for Using Free-Text Tobacco Use Comments

The first coding scheme for "reasons for use" was developed for identifying potential explanations for why the free-text comment field was used for each patient. In the initial version of the coding scheme, 16 different reasons were identified, which was expanded to 18 reasons (including one for *Other*) in the final version that were grouped into four major categories: (1) Misplaced or redundant information in free-text, (2) Missing values for available structured fields, (3) Limited capabilities of available structured fields, and (4) Other (Table 2). Comments could be associated with one or more potential explanations. For example, the comment "Occasional cigar" would be coded with two reasons: (1) *Misplaced – use Type field* and (2) *Limited ability to describe frequency*. One reviewer then analyzed the set of 525 entries to determine the most frequent reasons for using the free-text tobacco use comment

field. Inter-rater reliability between two reviewers for the set of 50 entries (almost 10%) was calculated using Cohen's kappa, achieving κ of 0.91 for coding reasons.

Table 2. Coding Scheme for Reasons.

#	Potential Reason	Brief Description	Example Comments
Misplaced or Redundant Information in Free-Text			
1	Misplaced – use Smoking status field	Smoking status field includes 10 values, including "Heavy Tobacco Smoker" and "Passive Smoker"	• hx of heavy tobacco use • exposed to second hand smoke
2	Misplaced – use Packs/day field	Could be entered using Packs/day field	• 50 years 2ppd = 100+ pack-years • 30yr x 0.5 ppd
3	Misplaced – use Years field	Could be entered using Years field	• 50 years 2ppd = 100+ pack-years • 30yr x 0.5 ppd
4	Misplaced – use Type field	Could be entered using options for Types (e.g., cigarettes, pipe, or cigars)	• Occasional cigar • Pipe
5	Misplaced – use Start or Quit date field	Could be entered using Start or Quit date fields	• quit 4/3/2011 • Quit just recently. (8/16/13)
6	Redundant Text	Same/synonymous information in comment also entered into structured fields	• Former smoker • nonsmoker
Missing Values in Available Structured Fields			
7	Missing value for Type field	Comment includes type that is not among available values for Type field	• Switched to an inhaler • electronic cigarette
Limited Capabilities of Available Structured Fields			
8	Limited ability to describe amount	Comment includes amount that cannot be described using available fields (Packs/day and Pack years fields)	• 1-2 cigars a day • 2 cans weekly • a few a day
9	Limited ability to describe frequency	Comment includes frequency that cannot be described using available fields (Packs/day)	• Smoked cigars sporadically • occasional pipe • a few a day
10	Limited ability to describe start or quit date	Comment includes a date that cannot be described using Start date or Quit date fields that require mm/dd/yy	• Quit April 2010 • Quit one year ago • Quit 1971
11	Limited ability to describe start or quit age	Comment includes an age related to starting or quitting	• smoked until age 16 • from age 18-26
12	Limited ability to describe duration or timepoint	Comment includes tobacco use or quit duration or timepoint	• Many years • Quit for 10 yr
13	Limited ability to describe situation	Comment includes information about situation or context of tobacco use	• Social • occasional in college
14	Limited ability to describe cessation attempts	Comment includes information about quit attempts, interventions, etc.	• Would like to quit • Working on quitting
15	Limited ability to describe passive smoke exposure	Comment includes information about passive smoke exposure that cannot be described in available fields	• Parents smoke in the home • No second hand smoke
16	Limited ability to specify multiple values	Comment includes additional status, age/date, etc.	• Quit 04/12/2008, restarted in 07/2008, quit 2/2010
Other			
17	Multiple statements	Comment includes multiple pieces of information	• 1-2 cigarettes/day. Quit on /1/13.
18	Other	Any other reason for use	• Smokes marijuana daily

Analysis of Contents within Free-Text Tobacco Use Comments

The second coding scheme for analyzing the "contents" of free-text was focused on categorizing words and phrases within the free-text comments into separate elements. The initial coding scheme included a combination of 10 elements identified in previous work involving the analysis of tobacco use information in clinical notes from multiple EHR systems as well as public health surveys[29, 30]. These elements included: (1) *Status* – current or past tobacco use, (2) *Temporal* – age or date when patient started or quit using tobacco (may be exact or estimated), (3) *Method* – how tobacco is/was used, (4) *Type* – what type of tobacco is/was used, (5) *Subtype* – additional details about type such as brand or filtered/unfiltered, (6) *Amount* – amount of tobacco the patient uses/used, (7) *Frequency* – how often tobacco is/was used, (8) *Certainty* – conviction of source (e.g., patient) regarding tobacco use, (9)

Experiencer – who uses/used tobacco, and (10) *Location* – where tobacco is/was used. An additional four elements were incorporated in the final version of the coding scheme for: (1) *Negation* – absence of tobacco use or exposure, (2) *Duration* – length of time a patient has used or quit using tobacco (explicitly separated out from the Temporal element), (3) *Situation* – context in which tobacco is/was used, and (4) *Cessation* – details about cessation such as attempts, interventions, or treatments.

Each comment was analyzed according to these 14 elements plus an element for *Other* where there could be multiple words or phrases associated with a particular element. For example, the comment "1-2 cigarettes/day" would be coded as Type = "cigarettes," Amount = "1-2," and Frequency = "/day" while the comment "no second hand smoke exposure" would be coded as Negation = "no," Method = "exposure," and Type = "second hand smoke". A κ of 0.94 was obtained for coding contents and main analysis involved determining more commonly used words and phrases for each element where groupings were created to combine those with similar meaning or pattern. For example, the words "chews," "chewed," and "chewing" were grouped together for the Method element while the pattern "*n* years ago" covered specific number of years such as "30 years ago" as well as estimated numbers such as "about 10 years ago" and "over 40 years ago".

Analysis of Data Quality Issues in Tobacco Use Entries

The third coding scheme for "issues" was designed to highlight potential data quality issues based on review of both the structured fields and free-text comment field. From review of the initial 100 tobacco use entries, seven issues were identified that were expanded to a total of 12 data quality issues where an entry could be associated with one or more issues (Table 3). For example, an entry where the value "Never Smoker" is specified in the structured smoking status field while the comment states "OCCASSIONAL CIGAR" would be coded with an issue of *Inconsistent smoking status*. As another example, an entry where the quit date is specified as "11/11/2011" and comment is "quit 2 years ago" would be coded with an issue of *Different temporal references*. Similar to the analysis of reasons for use, a κ of 0.91 was achieved for coding issues and main analysis involved determining the most frequent data quality issues across the 525 entries.

Table 3. Coding Scheme for Issues.

#	Issue	Brief Description	Example
1	Inconsistent smoking status	Contents of comment inconsistent with Smoking status field	Smoking status = Never Smoker Comment = OCCASIONAL CIGAR
2	Inconsistent packs/day	Contents of comment inconsistent with Packs/day field	Packs/day = 1.5 Comment = 01/01/2012 1 pack/day
3	Inconsistent years	Contents of comment inconsistent with Years field	Years = 15 Comment = 10 year smoking hx
4	Inconsistent pack years	Contents of comment inconsistent with calculated Pack years	Packs/day = 1 Years = 50 Comment = 50 years 2 ppd = 100+ pack-years
5	Inconsistent type	Contents of comment inconsistent with Type fields	Types = (not specified) Comment = cigars on occasion
6	Inconsistent start or quit date	Contents of comment inconsistent with Start or Quit date field	Quit date = 8/12/75 0:00 Comment = quit 1980's
7	Different levels of granularity	Contents of comment at different granularity level	Packs/day = 0.5 Comment = 5-10 cigarettes daily
8	Different temporal references	Comment includes relative time rather than absolute time	Quit date = 2/2/05 0:00 Comment = Quit smoking 8 years ago
9	Acronym or abbreviation	Comment includes acronym or abbreviation	• occ. cigar • 2 pks a week
10	Misspelling	Comment includes misspelling	• No smiking x3 days per pt • 3-4 cigarrettes a day
11	Ambiguous	Comment includes ambiguous information	• 2-3/week (*# of times, cigarettes, or packs?*) • 2005 (*start or quit year?*)
12	Not tobacco use	Contents of comment not related to tobacco use	• Smokes marijuana daily • Does not consume alcohol

Results

Based on analysis of the 525 tobacco use entries, Figure 1 depicts the distribution of potential reasons for using the free-text comment field. This field was most often used due to limited ability to describe amount (23.2%), frequency (26.9%), dates associated with starting or quitting (28.2%), and passive smoke exposure (17.9%). In addition, 26.9% of the comments included information considered redundant to what was captured in the structured fields such as smoking status and type.

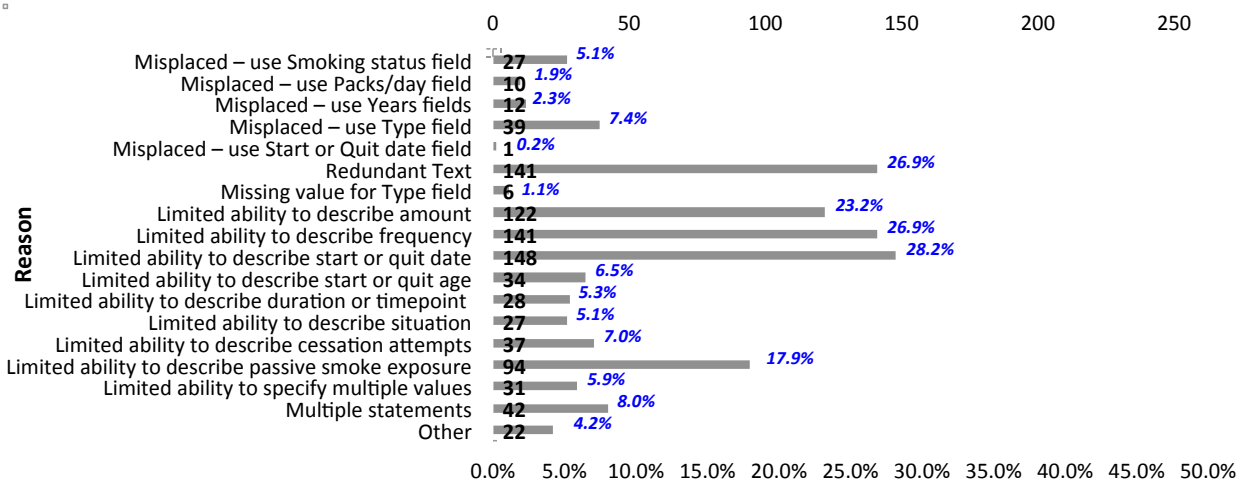


Figure 1. Distribution of Reasons for Use.

With respect to contents, Figure 2 shows the distribution of elements within the free-text comments where words or phrases most frequently described temporal information (38.3%), method (36.6%), type (33.5%), status (31.0%), frequency (29.1%), and amount (28.2%). For each of these elements, Table 4 includes the total number of values (i.e., words or phrases), number of unique values and groups, and the top 3 groups of values along with some examples. For example, of the 220 total values (154 unique values) for the Temporal element that were categorized into 16 groups, 30.0% reflected a specific or estimated year related to use of tobacco or quitting, 27.3% described a specific or estimated number of years ago, and 15.5% provided a specific or estimated age of use or quitting. For the Frequency element, the most frequent words or phrases were related to daily use or use every *n* days where *n* is a specific number or range (50.9%), occasional use (20.0%), and weekly use or use every *n* weeks (13.9%). While occurring less frequently, the majority of phrases categorized as Other were related to decreases in tobacco use (e.g., "cutting back," "down to," and "weaned down") suggesting the need to extend the coding scheme to include an additional element for Change.

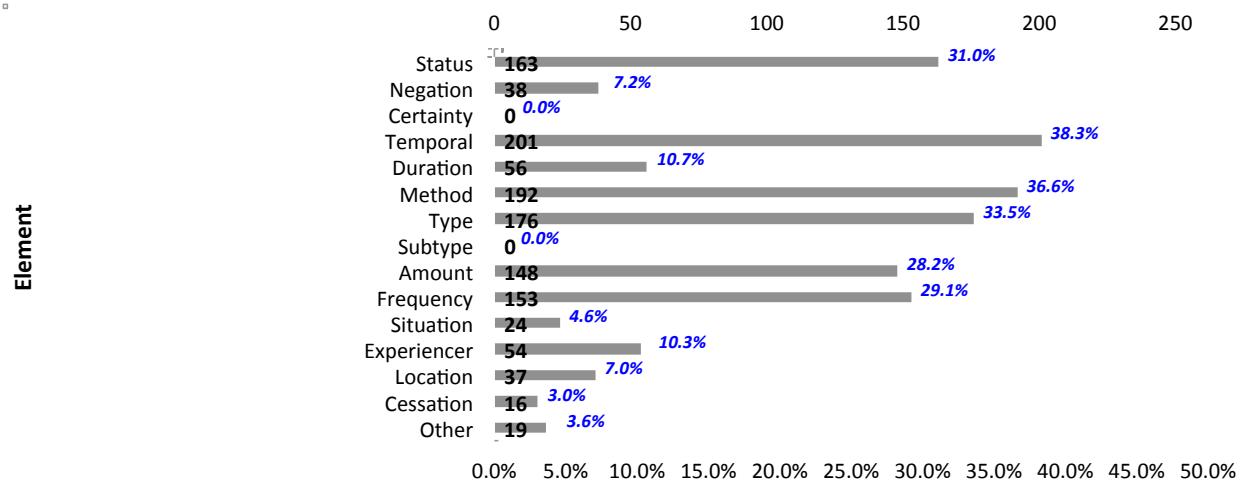


Figure 2. Distribution of Contents.

Table 4. Distribution of Values and Groups of Values for Top 6 Elements.

Element	Total # Values	# Unique Values [# Groups]	Top 3 Groups of Values (*Examples*)	Frequency	
Status	171	24 [8]	• quit (*quit, quitting, former smoker*)	120	(70.2%)
			• smoker (*smoker, smokers*)	18	(10.5%)
			• quit attempt (*trying to quit, process of quitting*)	14	(8.2%)
Temporal	220	154 [16]	• specific or estimated year (*1956, 1970s, early 2000s*)	66	(30.0%)
			• specific or estimated number of years ago (*10 years ago, about 8-10 years ago, over 40 years ago*)	60	(27.3%)
			• specific or estimated age (*age 18, early twenties, teenager*)	24	(15.5%)
Method	195	22 [8]	• exposure (*exposed, exposure*)	110	(56.4%)
			• smoke (*smokes, smoked, smoking*)	61	(31.3%)
			• chew (*chews, chewed, chewing*)	11	(5.6%)
Type	178	36 [13]	• secondhand smoke (*2nd hand, passive smoke, second hand tobacco*)	65	(36.5%)
			• cigarette (*cig., cigs, cigarettes*)	56	(31.5%)
			• cigar (*cigar, cigars*)	25	(14.0%)
Amount	155	85 [17]	• *n* (*3, 7-10, ~8, about 5*)	82	(52.9%)
			• *n* packs (*1/2 pk, 2-3 packs, less than 1 pack, half a pack*)	35	(22.6%)
			• *n* ppd (*1/2-1 PPD, 2-3 ppd, over 1ppd*) *	15	(9.7%)
Frequency	165	60 [22]	• per day or *n* days (*daily, /day, qd, every couple days, 8-10 x/day*)	84	(50.9%)
			• occasional (*occ., now and then, periodically, rarely*)	33	(20.0%)
			• per week or *n* weeks (*weekly, /week, every 2 weeks*)	23	(13.9%)

** addresses both amount and frequency*

Figure 3 reflects the distribution of potential data quality issues associated with the set of tobacco use entries. The most frequent issue was use of acronyms and abbreviations in the comments (18.1%) such as for cigarette or cigarettes where there were 3 different abbreviated forms ("cig,," "cig.," and "cigs") and 2 types of misspellings ("cigarrettes" and "cigarretts"). Other more frequent issues were related to granularity differences (14.1%) such as specifying only the quit year in the comments as opposed to an exact date as provided in the structured quit date field and use of relative rather than absolute temporal references (7.6%) in the comments such as *n* years ago instead of a specific date. Finally, there were several cases of inconsistent number of packs/day (6.7%) that may be due to changes in tobacco use and inconsistent type (5.0%) where the type of tobacco use was specified in the comment but not in the relevant structured fields.

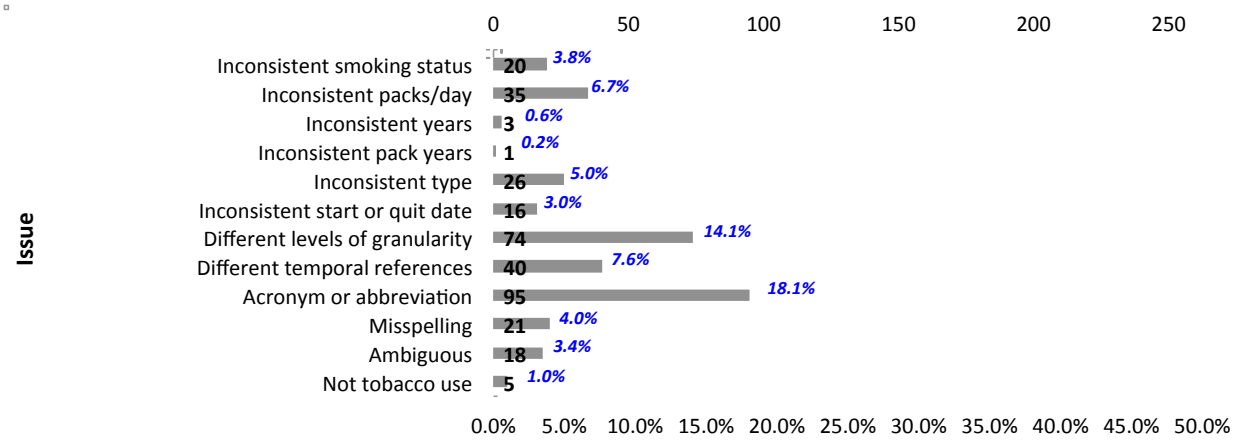


Figure 3. Distribution of Data Quality Issues.

Discussion

The findings of this study provide insights to the current use, contents, and data quality issues associated with free-text tobacco use documentation in the social history module of an EHR at an academic health center. The collective results highlight limitations in capturing details related to tobacco use and secondhand smoke exposure that may be used to inform system enhancements, user training, NLP, and standards for tobacco use documentation. In addition, this work represents a preliminary step towards developing a systematic and semi-automated process for evaluating EHR structure and content. While the study was limited to a single institution with a particular EHR system and focused on a specific module in this system, the overall findings as described below are expected to be generalizable to other institutions that may have the same or different EHR system. It is also anticipated that the three-phased approach for generating, validating, and applying coding schemes to retrospective EHR data could be adapted and applied to accommodate for institutional variations including differences in EHR structure. More broadly, this approach is extensible for performing both quality assessment and content analysis of information in other EHR modules that include free-text and/or structured fields.

The more frequent reasons and contents of the free-text tobacco use comments suggest the need for flexibility in describing amount, frequency, and start or quit dates associated with different types of tobacco or smokeless tobacco (e.g., cigarettes, pipe, cigars, snuff, and chew) that could not be accommodated with existing structured fields (i.e., for packs/day, pack years, and start and quit dates that require that the month, day, and year be specified). Other temporal information within the comments included start age, quit age, and quit duration. Potential system enhancements include incorporating additional structured fields as well as values within existing fields to enable the capture of information such as "occasionally 1-2 cigarettes," "2 cans of chewing tobacco weekly," "quit 1980," "quit in her 20's," "quit x 21 years" in addition to "1 ppd for 5 years" and "quit 2/2/12". The five reasons related to misplaced information (i.e., free-text used instead of available structured fields) were less common; however, they are indicative of gaps in the documentation process that could potentially be addressed through improved user training (e.g., reminders of existing EHR functionality for structured data entry and guidance for when/how to use free-text comments).

While occurring less frequently, there were several entries that included multiple statements or values reflecting changes in status (e.g., patient quit, restarted, and then quit again with associated dates) or amount (e.g., from 0.5 packs to 1 cigarette per day). Such changes could potentially be reflected or re-created by accessing the audit trail for the social history module; however, there may be value in having a more readily-accessible, comprehensive, and flexible "tobacco use history" (or broader "nicotine use history") that could be guided by the findings of this study in addition to existing standards for the representation of tobacco use (Table 5). These standards include those from HL7[31] (e.g., "social history observation," "smoking status observation," and "tobacco use observation" in implementation guides associated with the HL7 Clinical Document Architecture[32-35]) and openEHR[36] (e.g., archetypes for "Tobacco Use" and "Tobacco Use Summary"[37]) that collectively specify the collection of elements and associated values for status, method of use, substance (or type), amount, frequency, start date or age, and quit date or age.

Table 5. Example Tobacco Use History.

Date	Status	Start Date or Age	Quit Date or Age	Duration	Type	Amount	Frequency
2009-07-20	Former	Age: teenager	Date: 30 years ago	Use: 20 years	cigarettes	few	daily
2012-04-16	Current	Date: 2010-08-15			cigarettes	1 pack	weekly
2012-04-16	Current	Date: 2011			cigar	2-3	monthly
2013-10-06	Former	Date: 2011-04		Quit: 6 months	cigar	1	occasionally
2013-10-06	Current	Date: 2010-08-15			cigarettes	0.5 pack	weekly

In addition to describing tobacco use, other uses and contents of the free-text comments were related to secondhand smoke exposure and tobacco cessation (including attempts and interventions). While the list of available values for the structured smoking status field includes one for passive smoking, this is limited to patients who have never smoked and therefore could not be applied to those who were former smokers or who are also current smokers. The occurrence of comments describing exposure or no exposure to secondhand smoke as well as details about experiencer (who smokes – e.g., "parents," "father," or "mother") and location (where smokes – e.g., "outside," "home," or "car") could be used to inform the development of a set of structured fields focused on secondhand smoke exposure.

For tobacco cessation, analysis was performed at a high-level in this study given the breadth of information where further examination is planned to better understand the contents, guide enhancements, and inform how the social history module could promote interventions (e.g., through decision support functionality such as alerts and reminders). As part of this effort, the coding scheme for elements could be extended based on cessation-related elements defined in standards such as the openEHR archetypes for Tobacco Use, Tobacco Use Summary, and Cessation Attempts[37]. Open questions also include determining where and how to document nicotine replacement therapies such as nicotine gum, patches, and inhalers or devices such as electronic cigarettes[38], which were initially coded as missing values for the structured type field when analyzing reasons for use.

Finally, a number of data consistency and other quality issues were noted that could present challenges in using tobacco use information from the social history module for decision support, research, public health, and other primary and secondary uses. In some cases, it was found that information in the free-text comment was inconsistent with the structured fields such as indicating a different status (e.g., never smoker vs. current smoker or former smoker) or number of packs/day where the former could lead to missing or incorrectly identifying patients for tobacco cessation interventions. In other cases, the free-text was found to be the only source of information such as indicating the type of tobacco use (e.g., cigarettes or cigars). For both cases, user training may be one approach for improving and ensuring consistency in documentation prospectively while NLP techniques could be developed to extract details about tobacco use, secondhand smoke exposure, cessation attempts, and interventions both retrospectively and prospectively. Next steps include examining the documentation of tobacco use in other parts of the EHR (e.g., problem list and clinical notes) to further characterize data consistency and quality issues as well as determine how to integrate this information for subsequent uses.

Conclusion

With the increased adoption of EHR systems, there is a need for efforts to explore their potential for improving tobacco use assessment and cessation. This study involved examining the current collection of tobacco use information in the social history module of an EHR with an emphasis on free-text documentation. Based on the preliminary findings, implications for improving the use of information related to tobacco use and secondhand smoke exposure in the EHR include system enhancements, user training, NLP, and standards.

Acknowledgments

Research reported in this manuscript was supported by the National Library of Medicine of the National Institutes of Health under award number R01LM011364. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Centers for Disease Control and Prevention. Smoking & Tobacco Use Fast Facts. [March 2014]; Available from: http://www.cdc.gov/tobacco/data_statistics/fact_sheets/fast_facts/.
2. U.S. Department of Health and Human Services. The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General. 2014 [March 2014]; Available from: http://www.cdc.gov/tobacco/data_statistics/sgr/50th-anniversary/index.htm.
3. World Health Organization. Tobacco Fact Sheet. [March 2014]; Available from: http://www.who.int/mediacentre/factsheets/fs339/en/.
4. World Health Organization. WHO Report on the Global Tobacco Epidemic, 2013. [March 2014]; Available from: http://www.who.int/tobacco/global_report/2013/en/.
5. Centers for Disease Control and Prevention. Smoking & Tobacco Use - Healthy People 2020. [March 2014]; Available from: http://www.cdc.gov/tobacco/basic_information/healthy_people/.
6. Healthy People 2020 Tobacco Use. [March 2014]; Available from: http://www.healthypeople.gov/2020/topicsobjectives2020/overview.aspx?topicid=41.
7. Centers for Disease Control and Prevention. National Tobacco Control Program. [March 2014]; Available from: http://www.cdc.gov/tobacco/tobacco_control_programs/ntcp/.
8. Moyer VA, U. S. Preventive Services Task Force. Primary care interventions to prevent tobacco use in children and adolescents: U.S. Preventive Services Task Force recommendation statement. Ann Intern Med. 2013 Oct 15;159(8):552-7.
9. U. S. Preventive Services Task Force. Counseling and interventions to prevent tobacco use and tobacco-caused disease in adults and pregnant women: U.S. Preventive Services Task Force reaffirmation recommendation statement. Ann Intern Med. 2009 Apr 21;150(8):551-5.

10. World Health Organization. Tobacco Free Initiative. [March 2014]; Available from: http://www.who.int/tobacco/en/.
11. Institute of Medicine. Capturing Social and Behavioral Domains in Electronic Health Records: Phase 1. Washington, DC: The National Academies Press; 2014.
12. Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. N Engl J Med. 2010 Aug 5;363(6):501-4.
13. Meaningful Use Stage 2 - Record Smoking Status. [March 2014]; Available from: http://www.healthit.gov/providers-professionals/achieve-meaningful-use/core-measures-2/record-smoking-status.
14. eCQM Library. [March 2014]; Available from: http://cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/eCQM_Library.html.
15. Boyle R, Solberg L, Fiore M. Use of electronic health records to support smoking cessation. Cochrane Database Syst Rev. 2011(12):CD008743.
16. Lindholm C, Adsit R, Bain P, Reber PM, Brein T, Redmond L, et al. A demonstration project for using the electronic health record to identify and treat tobacco users. WMJ. 2010 Dec;109(6):335-40.
17. Greenwood DA, Parise CA, MacAller TA, Hankins AI, Harms KR, Pratt LS, et al. Utilizing clinical support staff and electronic health records to increase tobacco use documentation and referrals to a state quitline. J Vasc Nurs. 2012 Dec;30(4):107-11.
18. Mathias JS, Didwania AK, Baker DW. Impact of an electronic alert and order set on smoking cessation medication prescription. Nicotine Tob Res. 2012 Jun;14(6):674-81.
19. Zheng K, Hanauer DA, Padman R, Johnson MP, Hussain AA, Ye W, et al. Handling anticipated exceptions in clinical care: investigating clinician use of 'exit strategies' in an electronic health records system. J Am Med Inform Assoc. 2011 Nov-Dec;18(6):883-9.
20. Wang SJ, Bates DW, Chueh HC, Karson AS, Maviglia SM, Greim JA, et al. Automated coded ambulatory problem lists: evaluation of a vocabulary and a data entry tool. Int J Med Inform. 2003 Dec;72(1-3):17-28.
21. Chen ES, Garcia-Webb M. An analysis of free-text alcohol use documentation in the electronic health record: early findings and implications. Appl Clin Inform. 2014;5(2):402-15.
22. Zhou L, Mahoney LM, Shakurova A, Goss F, Chang FY, Bates DW, et al. How many medication orders are entered through free-text in EHRs?--a study on hypoglycemic agents. AMIA Annu Symp Proc. 2012;2012:1079-88.
23. Uzuner O, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. J Am Med Inform Assoc. 2008 Jan-Feb;15(1):14-24.
24. Liu M, Shah A, Jiang M, Peterson NB, Dai Q, Aldrich MC, et al. A study of transportability of an existing smoking status detection module across institutions. AMIA Annu Symp Proc. 2012;2012:577-86.
25. Hazlehurst B, Frost HR, Sittig DF, Stevens VJ. MediClass: A system for detecting and classifying encounter-based clinical events in any electronic medical record. J Am Med Inform Assoc. 2005 Sep-Oct;12(5):517-29.
26. Hazlehurst B, Sittig DF, Stevens VJ, Smith KS, Hollis JF, Vogt TM, et al. Natural language processing in the electronic medical record: assessing clinician adherence to tobacco treatment guidelines. Am J Prev Med. 2005 Dec;29(5):434-9.
27. Wu CY, Chang CK, Robson D, Jackson R, Chen SJ, Hayes RD, et al. Evaluation of smoking status identification using electronic health records and open-text information in a large mental health case register. PLoS One. 2013;8(9):e74262.
28. Epic Systems Corporation. [March 2014]; Available from: http://www.epic.com/.
29. Chen ES, Manaktala S, Sarkar IN, Melton GB. A multi-site content analysis of social history information in clinical notes. AMIA Annu Symp Proc. 2011;2011:227-36.
30. Melton GB, Manaktala S, Sarkar IN, Chen ES. Social and behavioral history information in public health datasets. AMIA Annu Symp Proc. 2012;2012:625-34.
31. Health Level Seven International (HL7). [March 2014]; Available from: http://www.hl7.org/.
32. Dolin RH, Alschuler L, Boyer S, Beebe C, Behlen FM, Biron PV, et al. HL7 Clinical Document Architecture, Release 2. J Am Med Inform Assoc. 2006 Jan-Feb;13(1):30-9.
33. HL7/ASTM Implementation Guide for CDA® R2 -Continuity of Care Document (CCD®) Release 1. [March 2014]; Available from: http://www.hl7.org/implement/standards/product_brief.cfm?product_id=6.
34. HL7 Implementation Guide for CDA® R2: History and Physical (H&P) Notes, Release 1. [March 2014]; Available from: http://www.hl7.org/implement/standards/product_brief.cfm?product_id=19.
35. HL7 Implementation Guide for CDA® Release 2: IHE Health Story Consolidation, Release 1.1 - US Realm. [March 2014]; Available from: http://www.hl7.org/implement/standards/product_brief.cfm?product_id=258.
36. openEHR. [March 2014]; Available from: http://www.openehr.org/.
37. openEHR Clinical Knowledge Manager. [March 2014]; Available from: http://www.openehr.org/ckm/.
38. Fairchild AL, Bayer R, Colgrove J. The renormalization of smoking? E-cigarettes and the tobacco "endgame". N Engl J Med. 2014 Jan 23;370(4):293-5.

Automated Assessment of Medical Students' Clinical Exposures according to AAMC Geriatric Competencies

Yukun Chen, MS[1]; Jesse Wrenn, PhD[1]; Hua Xu, PhD[3,1]; Anderson Spickard III, MD, MS[1,2]; Ralf Habermann, MD[2]; James Powers, MD[2]; Joshua C. Denny, MD, MS[1,2]

Department of [1]Biomedical Informatics and [2]Medicine, Vanderbilt University, Nashville, TN; [3]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX

Abstract

Competence is essential for health care professionals. Current methods to assess competency, however, do not efficiently capture medical students' experience. In this preliminary study, we used machine learning and natural language processing (NLP) to identify geriatric competency exposures from students' clinical notes. The system applied NLP to generate the concepts and related features from notes. We extracted a refined list of concepts associated with corresponding competencies. This system was evaluated through 10-fold cross validation for six geriatric competency domains: "medication management (MedMgmt)", "cognitive and behavioral disorders (CBD)", "falls, balance, gait disorders (Falls)", "self-care capacity (SCC)", "palliative care (PC)", "hospital care for elders (HCE)" – each an American Association of Medical Colleges competency for medical students. The systems could accurately assess MedMgmt, SCC, HCE, and Falls competencies with F-measures of 0.94, 0.86, 0.85, and 0.84, respectively, but did not attain good performance for PC and CBD (0.69 and 0.62 in F-measure, respectively).

1. Introduction

National accreditation bodies including the Accreditation Council for Graduate Medical Education (**ACGME**) and the American Association of Medical Colleges (**AAMC**) have called for competency based curriculum and assessment models for training programs. Many medical schools have responded with education portfolios to capture student experiences in real and simulated environments matched to shared competency goals. Education portfolios, however, have not achieved widespread adoption, partially because current methods require significant manual entry of a limited amount of clinical data. Automatic and valid methods of capturing the richness of students' clinical experiences are needed.

In this project, we developed and validated the machine learning based methods used to identify student experiences in six AAMC geriatrics competency domains: medication management (**MedMgmt**), cognitive and behavioral disorders (**CBD**), falls, balance, gait disorders (**Falls**), self-care capacity (**SCC**), palliative care (**PC**), and hospital care for elders (**HCE**). We tested different feature sets such as bag of words, use of biomedical concepts identified through natural language processing (**NLP**), and a refined list of physician-identified concepts corresponding to a particular competency. This work also highlighted the significant challenges for identifying medical student competency from clinical notes.

2. Background

Competency-based assessment methods combine a variety of modalities to provide a comprehensive evaluation of a learner's knowledge and proficiency. [1,2] Medical schools using competency-based assessments typically rely on education portfolios to track students' progress.[3-5] Portfolio components can include personal reflections, examinations, individual and small group projects, simulation encounter reports such as observed structured clinical examinations (OSCEs), mentoring experiences, and clinical exposures. Handwritten log books or score sheets of clinical data,[6,7] replaced now by portable electronic solutions,[8-11] allow students to enter patient information including demographics, diagnosis, procedures performed, and/or severity of illness. Use of these systems is limited for various reasons, including lack of time. Furthermore, teachers often disagree with students on primary diagnoses. We propose a system that automatically captures all concepts in a student's notes and organizes the data automatically to reflect the student's full experience and proficiency along important clinical outcomes.

The AAMC and the John A. Hartford Foundation developed a minimum set of graduating medical student competencies to ensure competent care of older patients by new interns. [12] With the help of leading geriatric educators and survey responses from educators in a number of clinical domains, the consensus panel established

eight core geriatric competency domains. Each competency domain contains 2-5 competencies, outlining detailed goals for medical students in each domain. For example, the "medical management" domain includes 3 subtopics: 1) age related sensitivity to drug selection and dosing based on patient factors (e.g., renal and hepatic dysfunction); 2) identifying medications (e.g., anticholinergic and analgesics) that should be avoided or used with caution in the elderly; and 3) documenting the patient's complete medication list (including herbal and over-the-counter medications) and recognizing possible side effects. These competency domains represent an agreed-upon framework to guide educational curricula and assessment of medical students.

Competency-based assessment may be amenable to clinical NLP and machine learning (**ML**) methods. Prior work has shown that the logistic regression and concept-based queries can identify student exposure to some common presentations, such as chest pain or fever.[13] Such systems are currently being used at Vanderbilt to track student exposure to 16 common presenting problems through the Learning Portfolio website, which captures all student-authored clinical notes.[14] ML-based methods have not been studied extensively in the medical education domain. ML-based models are trained on available annotated datasets, and then applied to new samples to identify their labels. As a preliminary study, we focused on developing ML-based methods to identify six of the geriatric competency domains from clinical notes written by third and fourth year medical students in Vanderbilt.

3. Methods

The automatic competency detection tool consists of three components. The first is Learning Portfolio, a web-based system that gathers all students' clinical notes from patient encounters in the electronic medical record (**EMR**). [15] Portfolio employs NLP, using KnowledgeMap Concept Indexer (**KMCI**) and SecTag to identify biomedical concepts from these notes mapped to Unified Medical Language System (**UMLS**) with their local context (negation and section information). We then applied ML based competency detectors to the NLP output to determine the relevance of a note with respect to different geriatric competency domains based on identified biomedical concepts. Figure 1 illustrates the components used in this project.

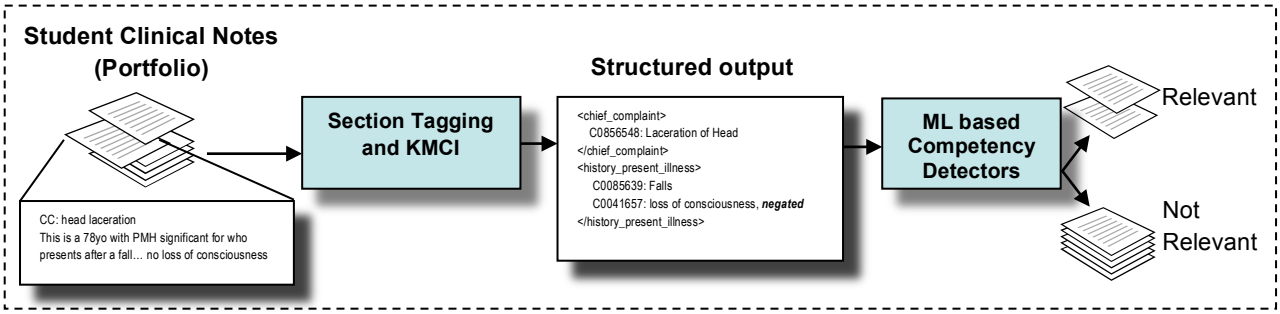


Figure 1. Overview of project: Machine learning based competency detectors assess a clinical note's relevance for each competency based on structured output of the KMCI.

3.1 Section Tagging

We used the **SecTag** section tagger to recognize clinical note sections and their boundaries.[16] The SecTag algorithm uses a concept-oriented, hierarchical section header terminology.[17] In this study, we only used the top sections and some of the first subsections shown in Table 1 as clinician-identified relevant sections. For example, "Family Medical History" under top section "Patient History" is kept as "Family Medical History." Key/first subsections listed in Table 1 replaced their subsections. For example, we have a hierarchical structure of the following sections: Objective Data -> Physical Examination -> Cardiovascular exam. "Objective Data" is the top section, "Physical Examination" is the first or key subsection, and "Cardiovascular exam" is the subsection of the key subsection. Then "Cardiovascular exam" was replaced by "Physical Examination". We also

Table 1. Note sections and subsections for use

- Patient History
  - Chief Complaint
  - History of Present Illness
  - Past Medical History
  - Medications
  - Review of Systems
  - Code Status
  - Health Maintenance
  - Personal and Social History
  - Family Medical History
- Objective Data
  - Laboratory and Radiology Data
  - Physical Examination
- Assessment and Plan
- Problem List
- Reference
- Follow Up
- None

Note: "None" represents the location of words or concepts outside section boundaries.

added a blank section "None" to indicate the content outside of these sections.

3.2 KnowledgeMap Concept Indexer (KMCI)

KMCI is a tool for analyzing unstructured text. For each clinical note, KMCI utilizes UMLS knowledge resources to encode concepts with Concept Unique Identifiers (**CUIs**). Once assigned a CUI, the UMLS provides, for each concept, semantic type information (e.g., "congestive heart failure" is a "disease or syndrome") and relationships to other concepts (e.g., "anterior myocardial infarction" is a type of "heart disease). Locating concepts in the appropriate section of a note can allow an educator or an algorithm to rate a student's mastery of a concept. For example, a search for key concepts related to back pain effectively identifies a note with back pain as a chief complaint, the extent of an appropriate back exam, and the presence or absence of differential diagnoses of back pain found in the student's written assessment.

3.3 Gold Standard of Competency Relevance

Our gold standard note corpus consists of 399 clinical notes from randomly selected inpatients older than 65 years or admitted to the geriatric service and followed by medical students. Geriatric educators rated a student's notes for each admission as containing high, medium, low, or no relevance to each geriatric competency domain. Each reviewer was a board-certified internal medicine physician who was either a geriatrician or has significant experience with geriatrics, including covering the inpatient geriatric service. A total of five reviewers scored each admission. Before scoring, physician reviewers reviewed test sets of admissions and scored them, developing a formal rubric for high, medium, and low/no for each competency domain. Then admissions were scored, disagreements discussed, and consensus achieved. The rubric was developed over the course of several 1-hour meetings of the physicians. The relevance scores of 3, 2, 1, and 0 represent the relevance level from "high", "medium", "low", and "no", respectively. A "high relevance" document contains primary discussion of the components that indicate an experience in that competency domain. In total, 119 admissions were scored by between one and five geriatric educators. For disagreement on scores, we applied the majority vote strategy and finalized the score using the score with the most votes for each student. If scores remain tied, we assign the mean of the tied scores as the final score. We considered students' notes with 'high' or 'medium' relevance (with final score of higher than 1) to have positive relevance and those with 'low' or 'no (with final score of 1 or smaller) to have negative relevance. Table 3 shows the class distribution for each competency domain.

Table 3. Class distribution for six competency domains

Competency Domains	Number of Positive Samples	Number of Negative Samples	Total Samples
Medication Management (MedMgmt)	94 (88%)	13 (12%)	107
Cognitive and Behavioral Disorders (CBD)	46 (43%)	61 (57%)	107
Falls, balance, gait disorders (Falls)	77 (72%)	28 (28%)	107
Self-care capacity (SCC)	78 (74%)	28 (26%)	106
Palliative care (PC)	44 (45%)	54 (55%)	98
Hospital care for elders (HCE)	79 (74%)	28 (26%)	107

3.4 Machine Learning Based Competency Detectors

We applied supervised machine learning techniques to determine a student's experience with geriatric competencies by identifying relevant concepts or other features in the corpus of notes. Experts labeled each case relevant or irrelevant to competency domains. We tested several different feature sets including "Bag of Words" as the baseline feature set, CUIs identified by KMCI, CUIs coupled with Section (SEC), Negation (NEG), and semantic type (STY). We tested other features such as the counts of CUIs in each note as well as the normalized values of CUIs based on term frequency-inverse document frequency (TFIDF). In addition, we added the number of notes in each admission

and the age of patient as two basic features in all experiments. Table 2 shows detailed descriptions of all feature sets in our experiments.

We tried to mimic the way geriatric educators identified the relevance of competency from students' clinical notes. Instead of using all CUIs from the notes, we also tested the feature set with a refined list of CUIs associated with corresponding competency domains, just as geriatric educators use searches for key concepts to assess students' notes. These lists of CUIs were developed by clinicians using web-based tools part of the KnowledgeMap curriculum website,[18] in which complex concept queries are used to track themes in the medical school curriculum. We have previously shown good performance using concept queries of this sort to identify broad themes in the curriculum (e.g., "genetics", "radiology").[18] This reduced the size of features dramatically.

Naïve Bayes (**NB**, the baseline classifier, implemented in CLOP[19]), logistic regression (**LR**), and support vector machines (**SVM**) with linear kernel are three efficient supervised learning tools for the data in high dimensional space. We applied both LR and linear SVM in the package Liblinear. [20] We constructed a classifier for each competency domain using all feature sets including refined lists of CUIs.

Table 2. Description of feature sets used in machine learning experiments

Name of Feature Set	Type of Feature	Description
Note_count	Integer	The number of clinical notes by a medical student for each admission
Age	Integer	Age of patient
Words (baseline)	Binary	Bag of Words features; "1" if word present; "0" if word absent
CUI	Binary	Concept code features; "1" if CUI present; "0" if CUI absent
CUI_NEG	Binary	Concept code with negation: If a CUI is negated in the note, the value of CUI is "0" and the value of CUI_NEG is "1"
CUI_SEC	Binary	Dyad of concept code (CUI) and section: "1" if CUI present in the section; "0" if CUI absent in the section
CUI_count	Integer	The count of each CUI in the notes for one admission
CUI_count_tfidf	Numeric	The TFIDF value of each CUI in the notes for one admission
STY	Binary	Semantic type features (as defined in the UMLS): "1" if semantic type present; "0" if semantic type absent

3.5 Validation

For each competency domain, we ran experiments using 10-fold cross validation for three machine learning algorithms and all feature set candidates. Each fold was stratified so that the distribution of classes in each fold was similar to the original distribution over all samples. These algorithms produced a numeric "relevance score" for the test samples on each fold. We compared the performances of each method with each feature set by computing the average area under receiver operator characteristic curve (AUC) over the cross validation. In addition, using different thresholds for "relevance scores" from the scores of the ML algorithms, we generated different sets of binary predictions and analyzed the precision (i.e., positive predictive values), recall (i.e., sensitivity), and F-measure (the harmonic mean of recall and precision) for all test samples.

4. Results

We evaluated the competencies at the admission level, and a student could write multiple notes for each admission. Totally we had 119 admissions that consisted of 399 clinical notes. There were total 11,249 unique CUIs for all these notes with at least one grade for relevance. Refining this list using just the physician-identified relevant CUIs dramatically reduced the number of features. Given hundreds of refined CUIs for each competency by a geriatric and NLP expert, only 24 CUIs occurred in the notes related to competency domain of medication management, 33 related to CBD, 52 related to Falls, 35 related to PC, 93 related to HCE, and 61 related to SCC. Tables 4 to 9 show the AUC scores for different feature sets and machine learning algorithms for Medication Management, CBD, Falls,

Self-care capacity, Palliative Care, and Hospital care for elders competencies, respectively. The best AUC score for each competency was highlighted in bold.

The SVM classifier generated the models with the best AUC scores of 0.91, 0.76, and 0.69 for competencies MedMgmt, PC, and HCE, respectively. The Naïve Bayes classifier achieved the best models with AUC scores of 0.73, 0.75, and 0.80 for competencies CBD, Falls, and SCC, respectively. For these six best models, we performed a precision-recall analysis based on different thresholds of numeric outputs, or "relevance score", by the classifiers (see Figure 2). The red dot in each graph in Figure 2 represents the precision and recall using the best threshold.

Table 4. AUC results for competency MedMgmt over different feature sets

Feature Sets	Num of Features	Naïve	LR	SVM
Words	27406	0.72	0.73	0.75
CUI	10258	0.84	0.82	0.78
CUI_SEC	25947	0.72	0.77	0.80
CUI_SEC + CUI_NEG	27978	0.67	0.77	0.79
CUI_SEC + CUI_NEG + STY	28237	0.68	0.78	0.81
CUI_count	10576	0.82	0.89	**0.91**
CUI_count_tfidf	10572	0.77	0.68	0.86
Refined CUI_count	24	0.76	0.61	0.74
Refined CUI_count+CUI_SEC+CUI_NEG +STY	318	0.68	0.76	0.84

Table 5. AUC results for competency CBD over different feature sets

Feature Sets	Num of Features	Naïve	LR	SVM
Words	27406	0.61	0.62	0.62
CUI	10258	0.68	0.68	0.69
CUI_SEC	25947	0.65	0.60	0.61
CUI_SEC + CUI_NEG	27978	0.63	0.61	0.62
CUI_SEC + CUI_NEG + STY	28237	0.65	0.60	0.60
CUI_count	10576	0.66	0.68	0.68
CUI_count_tfidf	10572	0.64	0.57	0.66
Refined CUI_count	33	**0.73**	0.68	0.68
Refined CUI_count+CUI_SEC+CUI_NEG +STY	387	0.70	0.66	0.64

Table 6. AUC results for competency Falls over different feature sets

Feature Sets	Num of Features	Naïve	LR	SVM
Words	27406	0.60	0.60	0.65
CUI	10258	0.64	0.59	0.63
CUI_SEC	25947	0.61	0.51	0.59
CUI_SEC + CUI_NEG	27978	0.59	0.50	0.61
CUI_SEC + CUI_NEG + STY	28237	0.60	0.52	0.61
CUI_count	10576	0.62	0.54	0.62

CUI_count_tfidf	10572	0.62	0.66	0.67
Refined CUI_count	52	0.65	0.50	0.57
Refined CUI_count+CUI_SEC+CUI_NEG +STY	469	**0.75**	0.56	0.64

Table 7. AUC results for competency PC over different feature sets

Feature Sets	Num of Features	Naïve	LR	SVM
Words	25904	0.59	0.60	0.62
CUI	9924	0.65	0.75	**0.76**
CUI_SEC	24767	0.63	0.67	0.68
CUI_SEC + CUI_NEG	26707	0.60	0.65	0.66
CUI_SEC + CUI_NEG + STY	26958	0.59	0.64	0.65
CUI_count	10245	0.68	0.68	0.67
CUI_count_tfidf	10241	0.72	0.50	0.59
Refined CUI_count	35	0.58	0.58	0.60
Refined CUI_count+CUI_SEC+CUI_NEG +STY	174	0.60	0.51	0.53

Table 8. AUC results for competency HCE over different feature sets

Feature Sets	Num of Features	Naïve	LR	SVM
Words	27406	0.58	0.62	0.64
CUI	10258	0.57	0.67	0.66
CUI_SEC	25947	0.65	0.66	**0.69**
CUI_SEC + CUI_NEG	27978	0.64	0.67	0.66
CUI_SEC + CUI_NEG + STY	28237	0.64	0.65	0.68
CUI_count	10576	0.62	0.55	0.55
CUI_count_tfidf	10572	0.61	0.65	0.64
Refined CUI_count	93	0.66	0.54	0.51
Refined CUI_count+CUI_SEC+CUI_NEG +STY	670	0.66	0.64	**0.69**

Table 9. AUC results for competency SCC over different feature sets

Feature Sets	Num of Features	Naïve	LR	SVM
Words	27261	0.75	0.71	0.70
CUI	10227	0.78	0.70	0.69
CUI_SEC	25850	**0.80**	0.78	0.77
CUI_SEC + CUI_NEG	27867	0.79	0.75	0.75
CUI_SEC + CUI_NEG + STY	28126	0.78	0.78	0.77
CUI_count	10540	0.71	0.69	0.69
CUI_count_tfidf	10536	0.68	0.54	0.58
Refined CUI_count	61	0.58	0.70	0.68
Refined CUI_count+CUI_SEC+CUI_NEG +STY	290	0.64	0.65	0.64

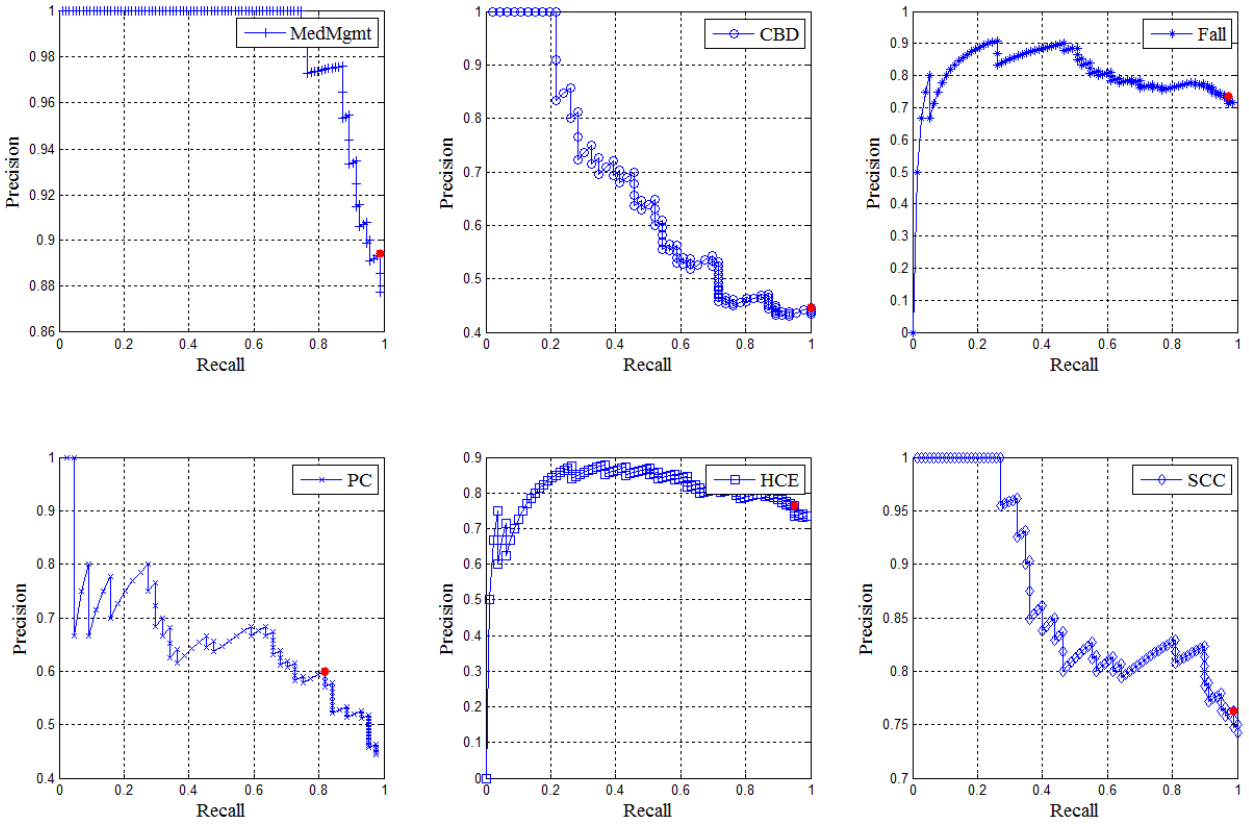


Figure 2. Precision-Recall graphs for the best-performing models for each competency domain. The red dot represents the point with the maximum F-measure (the harmonic mean of recall and precision).

By using the best threshold that can maximize the F-measure (the harmonic mean of recall and precision), we generated the table of precision, recall, and F-measure for each competency in Table 10. ML-model for MedMgmt competency achieved the best result with 0.89 in precision, 0.99 in recall, and 0.94 in F-measure. The next best competencies are SCC, HCE, and Falls with 0.86, 0.85, and 0.84 in F-measure, respectively. PC and CBD are two hard competency identification tasks because the best ML models can only achieve 0.69 and 0.62 in F-measure, respectively.

Table 10. Results of precision, recall, and F-measure for the best model for each competency

	Precision	Recall	Fmeasure
MedMgmt	0.89	0.99	0.94
CBD	0.45	1.00	0.62
Falls	0.74	0.97	0.84
PC	0.60	0.82	0.69
HCE	0.77	0.95	0.85
SCC	0.76	0.99	0.86

5. Discussion

Medical student competency assessment, an ultimate goal of medical education, has typically been largely performed through standardized tests and the subjective assessments of clinical preceptors. This research represents an attempt to implement an objective assessment based on a complete capture of all clinical notes that students write. We chose geriatric competencies due to local interest and 2008 Institute of Medicine report on the importance of geriatrics for the aging US population.[21] We applied NLP and ML methods to 6 of the 8 agreed-upon geriatric competencies as a novel, automated assessment of medical student's competency that are poorly assessed currently, and when they are assessed, it is usually either via a survey or via manual effort by educators. We did not pursue the other two competencies (HCPP: Health care planning and promotion and APD: Atypical presentation of disease) for machine learning. Regarding APD, the physicians failed to resolve an operational definition and consistent rating between them – some Kappas between reviewers for these categories were <0, indicating the difficulty in defining educational competency even despite multiple face-to-face meetings. HCPP was highly unbalanced with respect to the ratio between positive and negative samples, with only 7% of the notes marked as irrelevant to the competency.

This is one of the first applications of NLP and machine learning methods to competency assessment. Based on this preliminary study, identifying the competency domain from the students' clinical notes is a hard problem and requires a variety of features to achieve effective results. The best AUC score of 0.91 in assessing MedMgmt is a desired performance using KMCI CUI counts as feature set. Moreover, methods using NLP methods significantly outperformed "Bag of Words" approaches, whose performance often did not differ significantly from random chance. It is important to note, however, that all of these algorithms had portions of the precision curves in which the precision exceeded 0.8 and 0.9. Such algorithms, implemented with high thresholds, could still significantly enhance upon current methods since they could automatically scan the hundreds to thousands of notes written by clinical students.

We expected that using an expert-refined list of concepts that are highly associated with the competency would improve results, as has performed well in the past.[13,18] In our experiments, the performances on the CBD, Falls, and HCE competencies improved with the refined list of concepts related to the corresponding competency. They are significantly better than the results generated by baseline "Bag of Words". These results told a similar story to the recent I2B2/VA challenge, where most of the research teams used concepts and their related features to improve the text classification performance.[22-26] For the MedMgmt, PC, and SCC competencies, however, the performance decreased with the refined list. This implied that our refined list of concepts for medication management might not be sufficient to cover the entire range of medication management, self-care capacity, and palliative care concepts, or we missed hidden relationships among these concepts that could help the detection. In addition, the best model for each competency domain could generate high recall/precision outcome if we adjust the threshold in figure 2.

Regarding the classifiers we used, Naïve Bayes, Logistic Regression, and linear SVM are all linear classifiers and could run very fast for training. We will try other complex models such as SVM with polynomial or Gaussian kernel[27] and Random Forest[28] to find better models in the future.

Our study has limitation in the following aspects. First, the sample sizes for machine learning are relatively small comparing to other similar NLP tasks. We found the annotation extremely challenging; annotators were spending 30 minutes or more per admission reviewing content for the 8 domains. With the size of training samples less than 150, building a model with high performance when considering multiple documents is extremely hard. Secondly, the quality of annotation result or gold standard in our study is less than perfect. Developing a rubric for relevancy of content was a source of considerable discussion between the physicians. Often, physicians would pick up on subtle hints of disease progression under different situations that would lend them to infer relevance of a given competency – such "hints" of relevance may be difficult for machine learning algorithms to assess. We implemented a majority vote scheme to decide the final label for each admission. There were several cases where the disagreement among raters is high (3 raters voted for high relevance, and another 3 raters voted for low relevance). We have not resolved these cases yet.

In the future, we may try competency detection models at an individual document level instead of a complete admission, as well as a larger annotated set. Machine learning approaches could be more powerful with a refined set of concept related features as well as non-controversial gold standard. Automatic feature section methods could help reduce the dimension of data and extract the most important concept codes with respect to the competency. Expert systems, coupling with machine learning approach, may improve the performance by incorporating the domain knowledge in medical education. Finally, we intend to extend our study to refine the models for all geriatric competency domains by constructing more reliable labels for the training data and test more types of NLP feature sets.

6. Conclusion

In this study, we used machine learning approaches to automate the assessment of geriatric competency for medical students using their clinical portfolios. Use of NLP to generate concept related feature sets as the input of machine learning based competency detectors improved performance. We found use of a physician-generated list of concepts to be our best performing feature set for 3 out of 6 competency assessment tasks. Our model could achieve optimal performance for MedMgmt, SCC, HCE, and Falls with high F-measures, and achieved high precision at some score threshold for all tested competencies.

Acknowledgement

This study was funded under a grant from the Stemmler Foundation of the National Board of Medical Examiners.

Reference

1. Davis MH HR. Competency-based assessment: making it a reality. *Medical teacher.* 2003;25(6):565-568.
2. Whitcomb M. Redirecting the assessment of clinical competence. *Acad Med.* 2007;82(6):527-528.
3. Smith SR, Dollase RH, Boss JA. Assessing students' performances in a competency-based curriculum. *Academic Medicine.* Jan 2003;78(1):97-107.
4. Dannefer EF, Henson LC. The portfolio approach to competency-based assessment at the Cleveland Clinic Lerner College of Medicine. *Academic Medicine.* May 2007;82(5):493-502.
5. Litzelman DK, Cottingham AH. The new formal competency-based curriculum and informal curriculum at Indiana University School of Medicine: Overview and five-year analysis. *Academic Medicine.* Apr 2007;82(4):410-421.
6. Langdorf MI, Montague BJ, Bearie B, Sobel CS. Quantification of procedures and resuscitations in an emergency medicine residency. *J Emerg Med.* Jan-Feb 1998;16(1):121-127.
7. Rattner SL, Louis DZ, Rabinowitz C, et al. Documenting and medical students' comparing clinical experiences. *Jama-J Am Med Assoc.* Sep 5 2001;286(9):1035-1040.
8. Alderson TS, Oswald NT. Clinical experience of medical students in primary care: use of an electronic log in monitoring experience and in guiding education in the Cambridge Community Based Clinical Course. *Med Educ.* Jun 1999;33(6):429-433.
9. Bird SB, Zarum RS, Renzi FP. Emergency medicine resident patient care documentation using a hand-held computerized device. *Acad Emerg Med.* Dec 2001;8(12):1200-1203.
10. Gordon JS, McNew R, Trangenstein P. The development of an online clinical log for advanced practice nursing students: a case study. *Stud Health Technol Inform.* 2007;129(Pt 2):1432-1436.
11. Sumner W, 2nd, Campbell J, Irving SC. Developing an educational reminder system for a handheld encounter log. *Fam Med.* Nov-Dec 2006;38(10):736-741.
12. The Medical Student Competencies in Geriatric Medicine (Accessed October 10, 2007, at http://www.pogoe.org.).
13. Denny JC, Bastarache L, Sastre EA, Spickard A. Tracking medical students' clinical experiences using natural language processing. *J Biomed Inform.* Oct 2009;42(5):781-789.
14. Spickard A, 3rd, Ridinger H, Wrenn J, et al. Automatic scoring of medical students' clinical notes to monitor learning in the workplace. *Med Teach.* Jan 2014;36(1):68-72.
15. Spickard A, 3rd, Gigante J, Stein G, Denny JC. Automatic capture of student notes to augment mentor feedback and student performance on patient write-ups. *J Gen Intern Med.* Jul 2008;23(7):979-984.
16. Denny JC. Evaluation of a novel terminology to categorize clinical document section headers and a related clinical note section tagger. *Nashville, TN: Vanderbilt University.* 2007.
17. Denny JC, Miller RA, Johnson KB, Spickard A, 3rd. Development and evaluation of a clinical note section header terminology. *AMIA Annu Symp Proc.* 2008:156-160.
18. Denny JC, Smithers JD, Armstrong B, Spickard A. BRIEF REPORT: "Where do we teach what?" - Finding broad concepts in the medical school curriculum. *J Gen Intern Med.* Oct 2005;20(10):943-946.
19. http://clopinet.com/CLOP/.

20. Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ. LIBLINEAR: A Library for Large Linear Classification. *J Mach Learn Res.* Aug 2008;9:1871-1874.
21. Retooling for an Aging America: Building the Health Care Workforce. *Report, Institute of Medicine of the National Academies.* April 11, 2008.
22. Torii M, Wagholikar K, Liu HF. Using machine learning for concept extraction on clinical documents from multiple data sources. *J Am Med Inform Assn.* Sep 2011;18(5):580-587.
23. D'Avolio LW, Nguyen TM, Goryachev S, Fiore LD. Automated concept-level information extraction to reduce the need for custom software and rules development. *J Am Med Inform Assn.* Sep 2011;18(5):607-613.
24. Minard AL, Ligozat AL, Ben Abacha A, et al. Hybrid methods for improving information access in clinical documents: concept, assertion, and relation identification. *J Am Med Inform Assn.* Sep 2011;18(5):588-593.
25. de Bruijn B, Cherry C, Kiritchenko S, Martin J, Zhu XD. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *J Am Med Inform Assn.* Sep 2011;18(5):557-562.
26. Jiang M, Chen YK, Liu M, et al. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *J Am Med Inform Assn.* Sep 2011;18(5):601-606.
27. Chang C-CaL, Chih-Jen. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology.* 2011;2(3):27:21--27:27.
28. Breiman L. RANDOM FORESTS. *Machine Learning* 2001;45(1):5-32.

Exploring the use patterns of a mobile health application for alcohol addiction before the initial lapse after detoxification

Ming-Yuan Chih, PhD, MHA, University of Kentucky, Lexington, KY

Abstract

How patients used Addiction-Comprehensive Health Enhancement Support System (A-CHESS)[1], a mobile health intervention, while quitting drinking is worthy exploring. This study is to explore A-CHESS use patterns prior to the initial lapse reported after discharge from inpatient detoxification programs. 142 patients with alcohol addiction from two treatment agencies in the U.S. were included. A comprehensive set of A-CHESS use measures were developed based on a three-level system use framework and three A-CHESS service categories. In latent profile analyses, three A-CHESS system use patterns—inactive, passive, and active users—were found. Compared to the passive users (with the highest chance of the initial lapse), the active users (with the lowest chance of such behavior) participated more in online social activities, used more sessions, viewed more pages, and used A-CHESS longer. However, the chances of the initial lapse between A-CHESS user profiles were not statistically different. Implications of this finding were provided.

Introduction

In 2012, 7% of the U.S. population aged 12 or older were diagnosed with alcohol dependence or abuse.[2] One major challenge in current addiction treatment is to extend care beyond traditional in-patient treatment.[3] The emerging application of mobile communication and sensor technology in health care (called mHealth) has the potential to complement current treatment approaches in alcohol relapse prevention.[4–6] Evidence shows that mHealth is well accepted by the underserved populations with chronic diseases (including addiction).[7,8] Although mHealth is pervasive, highly accepted, and seemingly effective, less is known about how patients actually took advantage of it, and whether different mHealth use patterns may lead to different health outcomes.

Addiction-Comprehensive Health Enhancement Support System (A-CHESS) is a smartphone application developed to support patients with alcohol addiction after they leave residential care. A recent randomized controlled trial showed a significant reduction in the number of risky drinking days for patients who were given the access to A-CHESS.[1] A-CHESS offers patients the ubiquitous access to services designed to promote the three basic needs (i.e., autonomous motivation (AM), coping competence (CC), and social relatedness (RE)) outlined in the Self-Determination Theory (SDT).[7,9,10] For example, the alerts and recommendations in A-CHESS are intended to supplement the impaired cognitive functions of patients by helping them to adopt effective coping strategies when needed. A-CHESS can connect patients to the immediate support network including their addiction counselors, families, and peer patients. Motivational stories and messages from their counselors and peer patients may motivate patients to be on top of their own recovery. The continuous access to services provided in A-CHESS is not available in any existing addiction treatment program, and its use and impact on addiction outcomes is worth exploring.

A recent development of A-CHESS is a predictive feature of alcohol lapses. Using patient-reported recovery status in A-CHESS, such as the levels of urge and alcoholics anonymous (AA) meeting attendance, a Bayesian network model was developed to assess the risks of a lapse within a week of reporting.[11] The model successfully predicts more than 80 percent of the cases.[11] In this model, patients with recent lapse experiences were found to have a higher chance of subsequent lapses. In other words, the model can better identify potential lapsers based on whether they have lapsed recently. However, for those who did not lapse recently (i.e. the initial lapse cases), the prediction could be further explored and improved.

A-CHESS use patterns that occurred before initial alcohol lapses may inform the likelihood of such events and make early interventions possible. Studies have shown that coping activities may reduce the risks of initial lapse.[12] The use of A-CHESS can be considered as a coping behavior. The patterns of A-CHESS use before the initial lapse could potentially offer valuable information to improve initial lapse prediction and prevention. The purpose of this study is to explore the A-CHESS use patterns before the initial lapse reported by the patients who have just completed residential treatment and returned to their own community.

A-CHESS System Use

A-CHESS creates electronic log files containing patients' use of the system. Log file analyses provide rich information about how the technological systems work and how patients may gain benefits from using different components in the systems.[13] In order to go beyond the traditional log file analyses where only the amount of use (e.g., page views) is analyzed, a comprehensive use-measurement framework that contains different levels of system use measures (i.e., system entry, exposure, and engagement) may help to better understand the use patterns of a multifunctional and comprehensive application, like A-CHESS.[14]

In mass media research, system entry is closely related to selection, meaning that people intentionally choose to allocate their time and attention to certain mass media, in this case, A-CHESS. System entry means that users choose to turn on A-CHESS and can be operationalized as a successful login or session.[14] Just as a psychosocial and behavioral intervention received by alcohol addiction patients, A-CHESS system entry means that patients choose to attend (open) the treatment (the applications). Choosing to access the treatments may imply stronger motivation of alcohol addiction patients.[15] Mounting evidence shows that the attendance of treatment sessions, AA meetings, religious activities, or 12-Step group meetings is significantly related to improved abstinence outcomes.[16,17]

System exposure refers to how much information is perceived by the users, often operationalized to be the amount (the number of pages) or the degree (the amount of time) of the exposure. As patients browse through more content in A-CHESS, they may feel more informed, more supported, and more motivated to resist the temptation to lapse. Studies showed that patients who attend more treatment sessions are more likely to achieve better outcomes.[18,19]

System exposure focuses on "what" contents are perceived by patients, while system engagement focuses on the patients' interactions with the system. A-CHESS patients may not only passively receive information but also actively engage in various navigation strategies. They actively learn from and communicate with the other patients or counselors via the social networking tools. Most importantly, A-CHESS users may choose to engage in using the system for a sustained amount of time. Previous studies showed that expressing emotional support and empathy in an online environment was related to a reduction of concerns and improved quality of life.[20] Patients who used the information systems more continuously (i.e. longer over-time use) reported improved health competence, social support and active participation in health care.[21] Similarly, Moos and Moos (2007) found that patients with alcohol addiction who participated in AA meetings for a longer time period were more likely to be remitted at all four follow-ups[22]. The participation of social networks has also been found to be an important protective factor in the recovery journey of patients with addiction.[23]

Although system entry needs to occur before exposure or engagement can be reasonably measured, and exposure must occur for engagement to be meaningful, these three use measurement levels are not in a continuum.[14] Since the tools in A-CHESS offer different functions and contents that are designed to meet the three SDT needs, the combination of three levels of system use measurement with three SDT classified A-CHESS services may offer a more comprehensive view of A-CHESS use patterns (e.g., entry to relatedness services, exposure to coping competence services, and engagement with autonomy motivation services).

Latent Profile Analysis to Study System Use

A holistic approach is needed to study the effects of A-CHESS system use patterns. By considering multiple use measures simultaneously, we can better understand the impact of the system on patient outcomes. Latent class analysis (LCA) has been increasingly used as an exploratory approach to study the underlying patterns of complex, observed response variables in either dichotomized or continuous scales.[24] Compared to traditional clustering algorithm, such as k-mean, LCA is a model-based approach that offers cluster solutions that are supported by rigorous statistical tests and is based on a mixture of the underlying probability distribution, which makes the solution less arbitrary.[25] Latent profile analysis (LPA), a type of LCA in which all indicators are continuous, has been used in medical research to identify relevant and valid groups in skin cancer risks[26], youth risky behavior patterns[27], and alcohol dependence and abuse.[28] In a recent study of the system use patterns of a web-based eHealth intervention for cancer patients, LCA was proved to be a useful technique to identify subgroups of eHealth system users.[29] In the present study, because the use measures are continuous, LPA will be used to explore the underlying A-CHESS use patterns based on a comprehensive set of A-CHESS use measures.

In a recent review of various effective behavior interventions for addiction patients, researchers pointed out that the real challenge is to choose the most appropriate treatments for a given patient.[30] The same challenge was faced in the development of various services offered in A-CHESS. Exploring A-CHESS use patterns before the initial lapse may

offer important information about which patterns of system use matter in A-CHESS. Ultimately, the understanding of effective A-CHESS use patterns may help to design more effective mHealth interventions for addiction patients.

Methods

Participants

170 patients were randomized into the intervention group in a randomized trial.[7] Among them, 142 patients were included in this study because of the inclusion criteria during data cleaning (see section in Data preparation and analysis). The intervention group was given the access to A-CHESS; the control cases did not have access to A-CHESS and so are excluded from this analysis. The patients were recruited from two residential treatment organizations—one in the Midwest and the other in the Northeastern U.S.—from February 2010 to November 2011. They were at least 18 years old, met the criteria for DSM-IV[31] alcohol dependence when they entered treatment, and were able to provide two backup contacts for follow-up. Patients were excluded if they had a history of suicidality, a significant developmental or cognitive impairment that would limit the ability to use A-CHESS, or vision problems.

Procedures

The study was conducted according to the Declaration of Helsinki of 1975 and approved by the Institutional Review Board at the University of Wisconsin-Madison. Patients were recruited at each clinic by project coordinators before they were discharged. Written informed consent was obtained. Before leaving the facilities, patients received training on how to use A-CHESS and the smartphones from their counselors. On the day when they were discharged (i.e. the intervention date), they each received a smartphone (either Palm Pre or HTC Evo) with a pre-installed A-CHESS program. The 3G mobile broadband connection was paid up to 8 months after the intervention dates. The patients agreed to use A-CHESS. However, using A-CHESS is not required to stay on study. The system log data were sent to a secure server at University of Wisconsin-Madison. Details about the trial can be found elsewhere.[1]

Intervention—SDT-Based Services Provided in A-CHESS

A-CHESS services were designed to focus on one or more SDT needs.[10] Detailed information about different A-CHESS services has been reported elsewhere.[7,10,32] The following is a list of services targeting each SDT construct.

Services for promoting autonomous motivation. High-Risk Locator, Notifications, Recovery Motivation, Panic Button, Weekly Check-in, Daily Check-In, Sobriety Date Counter, Our Stories, and Recovery Podcasts.

Services for improving coping competence. Notifications, Panic Button, Discussions, Ask an Expert, News, Easing Distress, Instant Library, Recovery Information, Open Expert, FAQs, Weblinks, Tutorials, Our Stories, and Recovery Podcasts.

Services for building relatedness. Panic Button, Weekly Check-in, Daily Check-In, Discussions, Ask an Expert, Events and Meeting Planner, My Friends, My Messages, My Profile, My Team, and Team Feed.

Measures

Initial lapse and prediction relationship. The substance use item from the Weekly Check-in was used as the indicator of a patient's initial lapse status. The Weekly Check-in is a self-monitoring service in A-CHESS for patients to track their recovery progress.[11] Patients filled out an online questionnaire in the Weekly Check-in every 7 days in A-CHESS on smartphones. The status of their self-reported substance use behavior (i.e., drug and alcohol use) in the last 7 days was measured on a dichotomous scale (yes/no). Therefore, a lapse is defined to have occurred when a patient indicated that he/she used alcohol or took drugs in the last 7 days in their Weekly Check-in reports.[11] Patients may report lapses in multiple Weekly Check-in reports during the 8-month study period. The first reported lapse cases in the Weekly Check-in after patients left the treatment facilities were identified as the initial lapse cases.

The lapse status that patients reported in the Weekly Check-in was about their substance use in the last 7 days. If the time between two consecutive Weekly Check-ins was more than 14 days, the prediction was considered invalid because it was separated by too many days from the prior report.[11] Therefore, only a current Weekly Check-in with a subsequent Weekly Check-in within 14 days was considered an effective data point. If the next Weekly Check-in reported a lapse, it was taken as such; otherwise, it was considered a non-lapse. The use data that happened between the current Weekly Check-in and the previous Weekly Check-in were retrieved and used to develop measures of pre-lapse A-CHESS use (Figure 1).

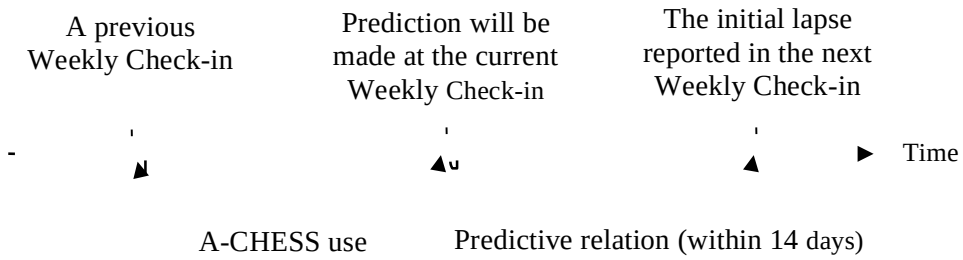


Figure 1. Initial lapse prediction conceptual diagram
Note: A-CHESS: Addiction-Comprehensive Health Enhancement Support System

Use measurements. A-CHESS was programmed to constantly record log files at the level of individual keystrokes or clicks of hyperlinks. The log files—with unique user identifiers and timestamps—can reveal precisely who requested what A-CHESS services and when. The log files can be used to establish activities in which the user was engaged (e.g., posting or reading) as well as the specific services that were requested (e.g., news or discussion group messages). From the log file data, A-CHESS use metrics were developed based on a combination of the SDT-focused A-CHESS service categories and the three levels of system use: entry, exposure, and engagement. SDT-based A-CHESS services were defined above. The operational meanings of the use measures are listed below.

Average daily sessions as system entry. The average number of daily sessions between the previous and current Weekly Check-in was calculated as the measurement for system entry. A new session is recorded when patients switched from using one service, such as discussion group, to another, such as easing distress. Four system entry measures were developed: Total entry (i.e. system entry measure of all A-CHESS services), AM entry (i.e. system entry measure of autonomous motivation services), CC entry (i.e. system entry measure of coping competence services), and RE entry (i.e. system entry measure of relatedness services).

Average daily pages viewed as system exposure. System exposure could be operationalized by pages viewed or time spent. Compared to those using computer-based applications, smartphone users were not restricted to a physical location, and therefore, could be more easily distracted by other tasks, like crossing the roads. This could cause time-based exposure measures to be overestimated. In addition, the contents in A-CHESS are often short and simple, and therefore may not require much time to achieve comprehension. Therefore, the duration of time between pages may not be the best measure in the context of a smartphone application. The amount of content (the number of pages viewed) may be a more useful and realistic measure of system exposure. The average number of daily pages viewed between the previous and current Weekly Check-in was calculated as the measure for system exposure. Four system exposure measures were developed: Total exposure (i.e. system exposure measure of all A-CHESS services), AM exposure (i.e. system exposure measure of autonomous motivation services), CC exposure (i.e. system exposure measure of coping competence services), and RE exposure (i.e. system exposure measure of relatedness services).

Percentage of days using A-CHESS as over-time system engagement. Over-time engagement has been used to describe the commitment and the continuous nature of system engagement in other studies.[14,21] In this study, the measure of over-time system engagement was operationalized by dividing the number of days a patient used A-CHESS by the total number of available days between the patients' previous and the current Weekly Check-in. The scale for this percentage measure is from 0% to 100%. Four over-time system engagement measures were developed: Total over-time engagement (i.e. over-time system engagement measure of all A-CHESS services), AM over-time engagement (i.e. over-time system engagement measure of autonomous motivation services), CC over-time engagement (i.e. over-time system engagement measure of coping competence services), and RE over-time engagement (i.e. over-time system engagement measure of relatedness services).

Average daily posting messages as social system engagement. The average number of daily messages posted or sent by patients between the last and the current Weekly Check-in was used to measure social system engagement. Because this measure describes the activity in online social services without a strong reference to SDT content categories, only one use measure was developed for social system engagement.

Data Preparation and Analysis

The use data retrieval was based on the submission date/time of the selected Weekly Check-in reports. Initial lapse cases were retrieved from the Weekly Check-in database as described in the process flow chart (Figure 2). In August 2011, a new intervention component was added to provide additional, automatic alerts sent out to both patients and counselors about patients' risk of upcoming lapses.[11] To avoid the impact of the new intervention on the predictive relationship studied here, four patients were excluded because all of their Weekly Check-in reports were submitted after this new intervention feature. Of the remaining patients, six did not have effective initial prediction data points: that is, a Weekly Check-in followed by a subsequent Weekly Check-in submitted within 14 days (Figure 1). Among 152 patients with effective prediction cases, 58 reported a lapse in at least one report, but 7 of them were excluded because their initial lapse was reported during their first Weekly Check-in—no prior Weekly Check-in can be used together as a prediction point. Because the 51 initial lapse cases occurred in different weeks of the 8-month study period, a matching process was performed to select compatible non-lapse cases. The process selected a number of non-lapse reports at a particular time point (e.g., 4th week after intervention) to match to the same proportion of lapse cases reported on the same week. Because all the initial lapse cases happened after the 2nd week of the intervention, 3 patients who only have non-lapse reports at the 2nd week were excluded as not matched non-lapse cases. The final dataset contains 51 lapse cases and 91 non-lapse cases.

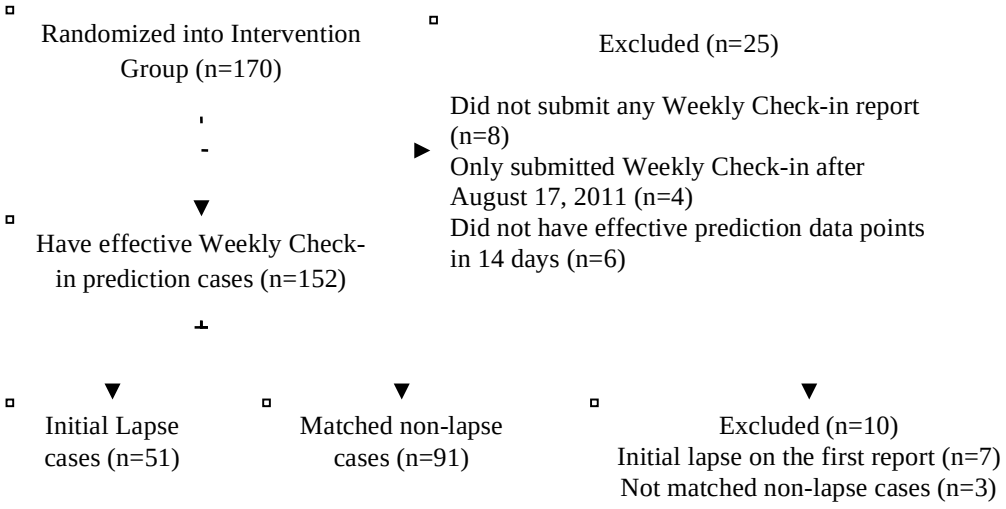


Figure 2. Initial lapse data process flow chart

After selecting these 142 cases (i.e. 51 lapse and 91 non-lapse cases), A-CHESS log data between reports (Figure 1) were retrieved. Usage measures were calculated for each case. Outliers (with a z-score over 3 or below -3) were recoded to the next most extreme values to minimize the impact of outliers.[33] A LPA was conducted to identify the underlying A-CHESS use patterns. For some variables with high proportion of extreme boundary values (e.g. 38% of non-users for coping competence services and 33% of patients had a 100% in total over-time engagement), a censored normal distribution was used in LPA using Mplus v7.11.[34,35]

This LPA identified the underlying use patterns (i.e. latent profiles) based on the comprehensive A-CHESS use measures. The number of latent profiles was increased until the most parsimonious model solution was found. The most parsimonious model was determined by minimizing Bayesian Information Criteria (BIC), statistically significant results in Lo-Mendell-Rubin likelihood ratio test (LMR-LRT), and entropy (over 0.9).[36] During each latent class modeling procedure, a three-step process was adopted to test the predictive relationship between latent profile groups and the distal outcome variable (i.e. the subsequently-reported initial lapse in the following Weekly Check-in). The three-step procedure can avoid the impact of the distal outcome variable on the latent profile groups when they are in the same model.[37] The three-step model, implemented in Mplus v7.11, was used to offer the pair-wise chi-square test of the probabilities of the initial lapse among the latent profiles.[37]

Results

142 patients (51 initial lapse cases or lapsers; and 91 non-lapse cases or non-lapsers), had a mean age of 37.93 years old (from 20 to 64), mostly Caucasians (82%) and mostly male (60.6%). About 40% of patients have received some college education or higher. Only 19% of patients are currently employed. About 60% of patients have abused drugs. Table 1 shows the descriptive statistics of demographic characteristics and A-CHESS use measures. Less than 4% outliers ($z > 3$) were found in all A-CHESS use measures except the Total, CC, and RE over-time engagement measures. The results of t-tests (for continuous variables) and chi-square tests (for categorical variables) showed that lapsers and non-lapses were not statistically different on most demographics and use measures. However, compared to non-lapsers, lapsers are more likely to be female (p=0.005) and currently unemployed (p=0.036).

Table 1. Descriptive statistics of patient characteristics and A-CHESS use

	Lapsers (n=51, 36%)	Non-lapsers (n=91, 64%)	p-value
Demographics			
Age, M(SD)	38.01(9.60)	37.86(9.12)	0.901
Female, N(%)	28(54.9)	28(30.8)	0.005
Race, N(%)			0.895
Caucasian	41(80.4)	73(80.2)	
African American	7(13.7)	11(12.1)	
Other	3(5.9)	7(7.7)	
Education, N(%)			0.248
High school and lower	25(49)	57(62.6)	
Some College	17(33.3)	25(27.5)	
2 and 4 year College degree	8(15.7)	9(9.9)	
Other	1(2)	0(0)	
Drug use, N(%)	33(64.7)	57(62.6)	0.806
Unemployed, N(%)	46(90.2)	69(75.8)	0.036
Use measures			
System entry (Average daily sessions), M(SD)			
Total Entry	4.01(4.41)	4.30(5.02)	0.731
AM entry	0.29(0.30)	0.27(0.30)	0.746
CC entry	0.73(1.36)	0.70(1.46)	0.919
RE entry	1.72(1.62)	1.83(1.88)	0.719
System exposure (Average daily pages viewed), M(SD)			
Total exposure	13.47(10.17)	14.99(12.93)	0.472
AM exposure	1.93(1.45)	1.77(1.26)	0.499
CC exposure	1.96(2.93)	1.53(2.60)	0.370
RE exposure	9.42(7.32)	9.73(8.96)	0.831
Over-time system engagement (% of days using A-CHESS), M(SD)			
Total over-time engagement	75.37(26.26)	72.86(30.92)	0.625
AM over-time engagement	35.46(16.84)	30.75(14.55)	0.083
CC over-time engagement	26.66(29.59)	23.02(28.59)	0.474
RE over-time engagement	68.38(27.46)	64.70(30.64)	0.477
Social system engagement (Average daily posting messages), M(SD)			
	0.32(0.43)	0.27(0.43)	0.451

Abbreviations: A-CHESS: Addiction-Comprehensive Health Enhancement Support System; AM: Autonomous motivation; CC: Coping competence; RE: Relatedness

A latent profile model was developed and estimated. The model contains all use measures outlined in the Measures section. Model fits indices based on different numbers of profile solutions were listed in Table 2. Among these

indices, BIC keeps decreasing when models become more complicated, and all models have good delineation of profiles with high entropy over 0.9. However, by examining LMR test result, the 3-profile solution was selected because the LMR test showed that the four-level model does not significantly improve the model fit (p=0.3849).

Table 2. Fit statistics of latent profile models

Full model				Decision
Number of Profiles	BIC	Entropy	LMR(p)	
2	9286	0.971	0.0452	
3	8943	0.968	0.0244	Selected
4	8760	0.976	0.3849	

Abbreviations: BIC: Bayesian Information Criteria, LMR: Lo-Mendell-Rubin likelihood ratio test.

In Table 3, model estimated profile means were listed to provide an overview of the latent profiles' characteristics. Based on the profile means of the use measures, three profiles can be named as inactive (35.9%), passive (49.3%), and active users (14.8%). The inactive users have the lowest A-CHESS use with lower than one service session per day, and about 40% over-time system engagement. The passive users have modest A-CHESS use, but their use in several over-time system engagement measures (except for CC over-time engagement) approached to the active users. Compared to the passive users, the active users have used more sessions, viewed more pages, and used A-CHESS longer especially in coping competence and relatedness services. The active users also more actively posted messages (almost one message a day) via online discussion groups or personal messages, than the passive users (about one message every three days). The active users have the lowest percentage of initial lapsers (28.6%) while the passive users have the highest chance (40.2%) for an initial lapse. However, the probabilities of the initial lapse between these three groups were not statistically different (p=0.315 to 0.691) at the present study.

Table 3. Model details

Profiles		Inactive Users	Passive Users	Active Users
% of profile membership counts[1]		35.9	49.3	14.8
% of the initial lapse		33.3	40.2	28.6
Model estimated profile means	System Entry (Average daily sessions)			
	Total Entry	0.82	3.98	13.14
	CC Entry	0.21	0.65	3.00
	RE Entry	0.39	1.83	5.09
	AM Entry	0.14	0.32	0.49
	System exposure (Average daily pages viewed)			
	Total exposure	4.22	15.26	36.70
	CC exposure	0.37	1.45	6.49
	RE exposure	3.01	9.70	25.51
	AM exposure	1.26	2.08	2.39
	Over-time system engagement (% of days using A-CHESS)			
	Total over-time engagement	41.90	90.54	95.46
	CC over-time engagement	6.17	24.12	68.22
	RE over-time engagement	33.57	81.71	93.58
	AM over-time engagement	23.87	35.37	43.72
	Social system engagement (Average daily posting messages)			
	Total	0.05	0.30	0.89

Note: 1. The percentages of user counts in profiles were based on the most likely latent profile membership. 2. Abbreviations: A-CHESS: Addiction-Comprehensive Health Enhancement Support System; AM: Autonomous motivation; CC: Coping competence; RE: Relatedness

Discussion

The purpose of this study is to explore the underlying use patterns of a mobile health application, A-CHESS, before the initial lapse after a period of inpatient alcohol detoxification. A-CHESS use measures were developed based on three-level system use framework (i.e., system entry, exposure, and engagement), as well as three SDT service

categories (i.e., autonomous motivation, coping competence, and relatedness services). In LPA, three profiles and their unique use patterns were found (Table 3). The inactive users really did not use A-CHESS much. Although the inactive users turned on A-CHESS on average in 4 days out of a 10-day period, they utilized very few (like one or two pages) coping competence and autonomous motivation services. In this case, they may just read the "daily thought" (i.e. an encouragement message sent out each day) or use the Weekly or Daily Check-in, which A-CHESS showed a reminder message on the phone screen. Compared to the inactive users, the passive users used the "social network (relatedness)" services more frequently. However, this use pattern may also be due to the "daily thoughts" message sent to them daily. The passive users did not post or read as many messages as the active users. Therefore, the inactive users did not use A-CHESS regularly even with the daily alerts. The passive users turned on A-CHESS (probably because of the daily prompts) but were not very engaged in using A-CHESS. The active users truly used A-CHESS regularly and actively. The passive users have the highest probability (40.2%) of the initial lapse in the following week while the active users have the lowest probability (28.6%) of such events. However, the probabilities of the initial lapse between these three user profiles were not different.

The non-statistically different likelihoods of the initial lapse between A-CHESS use profiles do not allow for further inferential conclusion to be drawn. However, the different A-CHESS use patterns and their relative chances of lapses—seemingly matching to the literature—showed the potential to explore further and worth mentioning. Compared to the passive users with the highest likelihood of the initial lapse, the active users who have the lowest likelihood of the initial lapse used more service sessions, viewed much more content about coping competence, and especially actively reached out to others by posting or sending messages. This observation is similar to what was found in the addiction literature where patients who participated more treatment sessions, improved their coping skills, and built stronger support network usually experienced better recovery outcomes.[22,23,38,39] Therefore, the active users' A-CHESS use patterns may be worth promoting, especially for those who consumed fewer coping competence services and less actively participated online social network. The inactive users may be a mix of those who have smooth recovery and did not feel the need of using A-CHESS, and those who really needed helps but just did not want to use A-CHESS for some reasons. Therefore, their chance of the initial lapse (33.3%) fell in between the other two user groups.

This study has several strengths in both system science and addiction recovery research. The analysis of system use behavior in this study was based on the actual log files, which are the objective observation of each activity that patients have with the intervention system, A-CHESS. Therefore, the results in this study have provided a different perspective of user behaviors other than the self-reported system use. Future studies that examine both the log files and the self-reported system use may offer a more complete picture of user behaviors. In addition, a comprehensive set of A-CHESS use measures were developed and analyzed together, which offered rich information about the different A-CHESS use patterns before the initial lapse. The LPA results, although non-conclusive, seemingly match well with the existing addiction literature regarding treatment effects on the alcohol lapse. The findings from this study may benefit to the design of A-CHESS-like technological interventions. The method used to study the system use patterns may serve as an example for researchers who develop and test new mHealth interventions.

Several limitations were related to the data used in this study. First, the data used in this analysis were based on patients' Weekly Check-in. If patients did not fill out Weekly Check-in in a way as described in the Methods section, their data would not be included in the analyses. Patients who did not fill out Weekly Check-in may have some valid and non-ignorable reasons that could be related to their substance use status. For example, some patients who drank might not want to disclose their drinking status in A-CHESS but would tell their counselors at a different setting. This kind of non-random missingness may limit the generalizability of the results. The latent use profiles found in this study may not be applied to all patients but only to those who would use A-CHESS and Weekly Check-in. Besides, the overlapped SDT categories may lead to potentially high correlations between the A-CHESS use measure indicators. A more complex model with a higher number of profiles may be needed in order to fit a complex correlation structure in our data. However, the relatively small sample size in this study precluded the construction of a more complex model and may not offer enough power to detect the group differences. Besides, patients from two treatment agencies may not be representative to the general alcohol addiction population. Therefore, future studies should acquire more participants from more sites in order to confirm and further understand the results of the present study. According to a recent PEW research report, adult smartphone owners in the U.S.A. increased from 30% to 50% in early 2013.[40] As mobile communication technology becomes more common, it is more likely to run population-based trial on A-CHESS or similar systems at a lower cost.

Patients who suffer from alcohol addiction need continuous support in order to recover from their addiction. In this case, ubiquitous mobile technology, like a smartphone, can be useful to deliver such interventions. This study

offered an understanding of the patients' A-CHESS system use patterns, and how these patterns were related to the chances of patients' initial lapse after the residential detoxification programs. Future research in this area to test and extend these findings is needed in order to develop and refine effective mHealth interventions for addiction patients.

Disclosure and Acknowledgement

The author reports on a part of his dissertation and his involvement in a clinical trial testing A-CHESS. The reason that the author is listed as the only author in this paper is because the author has the solo responsibility to the integrity and quality of his dissertation. The author appreciates the support from his mentor, Dr. David H. Gustafson and all the committee members, as well as the dedicated and talented project team behind this trial.

References

1. Gustafson DH, McTavish FM, Chih M-Y, et al. A smartphone application to support recovery from alcoholism: a randomized clinical trial. JAMA psychiatry 2014;71:566–72. doi:10.1001/jamapsychiatry.2013.4642
2. Substance Abuse and Mental Health Services Administration, Results from the 2012 National Survey on Drug Use and Health: Mental Health Findings, NSDUH Series H-47, HHS Publication No. (SMA) 13-4805. Rockville, MD: Substance Abuse and Mental Health Services Administration, 2013..
3. McKay JR. Continuing care research: what we have learned and where we are going. J Subst Abuse Treat. 2009 Mar;36(2):131–45.
4. Gustafson DH, Palesh TE, Picard RW, Plsek PE, Maher L, Capoccia VA. Automating addiction treatment: enhancing the human experience and creating a fix for the future. In: Bushko RG, editor. Future of Intelligent and Extelligent Health Environment. Birmingham, Alabama: IOS Press; 2005. p. 186–206.
5. Marsch L a., Ben-Zeev D. Technology-Based Assessments and Interventions Targeting Psychiatric and Substance Use Disorders: Innovations and Opportunities. J Dual Diagn. 2012 Nov;8(4):259–61.
6. Cohn AM, Hunter-Reel D, Hagman BT, Mitchell J. Promoting behavior change from alcohol use through mobile technology: the future of ecological momentary assessment. Alcohol Clin Exp Res. 2011 Dec;35(12):2209–15.
7. McTavish FM, Chih M-Y, Shah D, Gustafson DH. How patients recovering from alcoholism use a smartphone intervention. J Dual Diagn. 2012 Nov 10;8(4):294–304.
8. Rajput Z a, Mbugua S, Amadi D, Chepng'eno V, Saleem JJ, Anokwa Y, et al. Evaluation of an Android-based mHealth system for population surveillance in developing countries. J Am Med Inform Assoc. 2012 Feb 24;19(4):655-9.
9. Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am Psychol. 2000 Jan;55(1):68–78.
10. Gustafson DH, Shaw BR, Isham A, Baker T, Boyle MG, Levy M. Explicating an evidence-based, theoretically informed, mobile technology-based system to improve outcomes for people in recovery for alcohol dependence. Subst Use Misuse. 2011 Jan;46(1):96–111.
11. Chih M, Patton T, McTavish FM, Isham AJ, Judkins-Fisher CL, Atwood AK, et al. Predictive modeling of addiction lapses in a mobile health application. J Subst Abuse Treat. 2014 Jan 10;46(1):29–35.
12. Witkiewitz K, Masyn KE. Drinking trajectories following an initial lapse. Psychol Addict Behav. 2008 Jun;22(2):157–67.
13. Han JY. Transaction logfile analysis in health communication research: challenges and opportunities. Patient Educ Couns. Elsevier Ireland Ltd; 2011 Mar;82(3):307–12.
14. Shah D, Namkoong K, Moon TJ, Chih M-Y, Han JY. Explicating Use of ICTs in Health Contexts: Entry, Exposure, and Engagement. Association for Education in Journalism and Mass Communication 2011 Annual Conference. St. Louis, MO
15. Cunningham J a, McCambridge J. Is alcohol dependence best viewed as a chronic relapsing disorder? Addiction. 2012 Jan;107(1):6–12.
16. Bonn-Miller M, Zvolensky MJ, Moos RH. 12-Step Self-Help Group Participation as a Predictor of Marijuana Abstinence. Addict Res Theory. 2011;19(1):76–84.

17. Robinson EAR, Krentzman AR, Webb JR, Brower KJ. Six-month changes in spirituality and religiousness in alcoholics predict drinking outcomes at nine months. J Stud Alcohol Drugs. 2011 Jul;72(4):660–8.
18. Lash SJ. Increasing participation in substance abuse aftercare treatment. Am J Drug Alcohol Abuse.1998 Feb;24(1):31–6.
19. Lash SJ, Stephens RS, Burden JL, Grambow SC, DeMarce JM, Jones ME, et al. Contracting, prompting, and reinforcing substance use disorder continuing care: a randomized clinical trial. Psychol Addict Behav. 2007 Sep;21(3):387–97.
20. Namkoong K, Shah D V, Han JY, Kim SC, Yoo W, Fan D, et al. Expression and reception of treatment information in breast cancer support groups: how health self-efficacy moderates effects on emotional well-being. Patient Educ Couns. 2010 Dec;81 Suppl:S41–7.
21. Han JY, Hawkins RP, Shaw BR, Pingree S, McTavish F, Gustafson DH. Unraveling uses and effects of an interactive health communication system. J Broadcast Electron Media. 2009 Feb 27;53(1):112–33.
22. Moos RH, Moos BS. Treated and untreated alcohol-use disorders: Course and predictors of remission and relapse. Eval Rev. 2007 Dec 1;31(6):564–84.
23. Hunter-Reel D, McCrady BS, Hildebrandt T, Epstein EE. Indirect Effect of Social Support for Drinking on Drinking Outcomes: The Role of Motivation. J Stud Alcohol Drugs. 2010;71(6):930–7.
24. Finch WH, Bronk KC. Conducting Confirmatory Latent Class Analysis Using M plus. Struct Equ Model A Multidiscip J. 2011 Jan 13;18(1):132–51.
25. Magidson J, Vermunt JK. Latent class models for clustering: A comparison with K-means. Can J Mark Res. 2002 Jan 1;20(1):36–43.
26. Steffen AD, Glanz K, Wilkens LR. Identifying latent classes of adults at risk for skin cancer based on constitutional risk and sun protection behavior. Cancer Epidemiol biomarkers prev. 2007 Jul;16(7):1422–7.
27. Cole M, Stanton B, Deveaux L, Harris C, Cottrell L, Clemens R, et al. Latent class analysis of risk behaviors among bahamian young adolescents: relatoinship between values prioritizatio and latent class. Soc Behav Personal an Int J. 2007 Jan 1;35(8):1061–76.
28. Muthén B. Should substance use disorders be considered as categorical or dimensional? Addiction. 2006 Sep;101 Suppl 6–16.
29. Børøsund E, Cvancarova M, Ekstedt M, Moore SM, Ruland CM. How user characteristics affect use patterns in web-based illness management support for patients with breast and prostate cancer. J Med Internet Res. 2013 Jan;15(3):e34.
30. Witkiewitz K, Marlatt GA. Behavioral Therapy Across the Spectrum. Alcohol Res Heal. 2011;33(4):313–9.
31. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision (DSM-IV-TR). Arlington, VA: American Psychiatric Association; 2000.
32. Gustafson DH, Boyle MG, Shaw BR, Isham A, McTavish F, Richards S, et al. An e-health solution for people with alcohol problems. Alcohol Res Health. 2011 Jan;33(4):327–37.
33. Osborne JW, Overbay A. The power of outliers (and why researchers should always check for them). Pract Assessment, Res Eval. 2004;9(6):1–12.
34. Tobin J. Estimation of Relationships for Limited Dependent Variables. Econometrica. 1958 Jan;26(1):24.
35. Muthén LK, Muthén BO. Mplus User ’s Guide, ver 7. 7th ed. Los Angeles, CA: Muthén & Muthén; 2012.
36. Nylund KL, Asparouhov T, Muthén BO. Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo Simulation Study. Struct Equ Model A Multidiscip J. 2007 Oct 23;14(4):535–69.
37. Asparouhov T, Muthen B. Auxiliary Variables in Mixture Modeling : A 3-Step Approach Using Mplus, version 6. Mplus Web Notes No 15. 2013;(15).
38. Witkiewitz K, Marlatt GA. Emphasis on Interpersonal Factors in a Dynamic Model of Relapse. Am Psychol. 2005;60(4):341–2.
39. Litt MD, Kadden RM, Cooney NL, Kabela E. Coping skills and treatment outcomes in cognitive-behavioral and interactional group therapy for alcoholism. J Consult Clin Psychol. 2003;71(1):118–28.
40. Fox S, Duggan M. Mobile Health 2012. Pew Internet & American Life Project. 2012.

Online Deviation Detection for Medical Processes

Stefan C. Christov, George S. Avrunin, Lori A. Clarke
School of Computer Science, University of Massachusetts Amherst, MA

Abstract

Human errors are a major concern in many medical processes. To help address this problem, we are investigating an approach for automatically detecting when performers of a medical process deviate from the acceptable ways of performing that process as specified by a detailed process model. Such deviations could represent errors and, thus, detecting and reporting deviations as they occur could help catch errors before harm is done. In this paper, we identify important issues related to the feasibility of the proposed approach and empirically evaluate the approach for two medical procedures, chemotherapy and blood transfusion. For the evaluation, we use the process models to generate sample process executions that we then seed with synthetic errors. The process models describe the coordination of activities of different process performers in normal, as well as in exceptional situations. The evaluation results suggest that the proposed approach could be applied in clinical settings to help catch errors before harm is done.

1 Introduction

Human errors are a major concern in many medical processes. In 1998, an Institute of Medicine (IOM) report estimated that preventable medical errors cause the death of 98,000 people each year in the U.S.[1] More than a decade later, a 2009 U.S. National Research Council report[2] indicated that the problem with errors still persists and that "it is widely recognized that today's health care … suffers substantially as a result of medical errors." A recent study[3] from 2013 reported an even higher estimate of deaths per year, between 210,000 and 400,000, in the U.S. due to medical errors.

To help address this problem, we are investigating an approach for automatically detecting when performers of a medical process deviate from the acceptable ways of performing that process. Such deviations could represent *planning errors*[*] and, thus, detecting and reporting deviations as they occur (i.e., online) could help catch such errors before something bad happens. Here we use the word *process* to refer to the coordination of activities to accomplish a task or a goal, where the activities may be performed by humans, devices, or software systems. We refer to the *execution* of the process as the sequence of steps that are actually performed to accomplish the task or the goal. Our process models capture the process executions typically described in clinical guidelines or protocols, but also include additional detail, such as how guidelines should be executed when exceptional situations arise and customization that might occur for a particular clinical setting.

Detecting deviations could be challenging, however, because of the complexity of medical processes. Medical processes often contain multiple decision points that might require expert judgment, multiple possible exceptional situations that might require special handling, and multiple subprocesses that might be performed concurrently by different medical professionals. This typically results in a very large set of acceptable executions of a medical process. The proposed deviation detection approach uses techniques from software engineering for constructing a compact model describing all acceptable executions of a process. The large number of possible executions, however, combined with the fact that the process execution that human process performers intend to follow is often unknown (and could change as the process is being performed), makes online deviation detection difficult.

The proposed deviation detection approach relies on monitoring an executing process to determine whether the way the process is being performed deviates from the way it should be performed as specified by the process model. In particular, the approach aims to detect the manifestation of errors, i.e., *error phenotypes*[18], in a sequence of performed activities. Recent developments in healthcare and informatics should enable monitoring and recognizing many of the activities performed in real time, that is, as the process is being executed. For example, electronic medical records are being introduced in more hospitals and data entries in such records could be used to infer what activities are being/have been performed. Medical scribes are increasingly being used in medical processes to document the executing process[4] and thus could capture the activities as they are being performed. Computer vision techniques could also be utilized to recognize performed activities in a live video stream from cameras installed in a hospital room.

[*] A *planning error*, as defined by the IOM report[1], is "the use of a wrong plan to achieve an aim". Examples of planning errors are omitting an activity that should have been done (error of omission) or performing an activity that should not have been done (error of commission).

Even if we assume a perfect mechanism for monitoring an executing process, however, there are important issues related to the feasibility of the proposed deviation detection approach that need to be explored before the approach could be deployed in a clinical setting. One of these issues is potential harm due to *delayed deviation detection*. Delayed deviation detection arises when process performers deviate from the acceptable executions of a process, but this deviation cannot be detected until additional process activities are performed. Since the intent of process performers is often unknown to the monitoring system and some decisions at branch points in a process could depend on subjective (and potentially error-prone) human judgment based on experience and domain expertise, a monitoring system might not know which branch(es) should be taken in a given execution of the process. For example, suppose that at a branch point a human intends to perform one branch, but, due to distraction or fatigue, makes a commission error by first performing steps from another branch before performing the steps from the intended branch. In that case, a monitoring system will not be able to detect the commission error until the human performs the first step of the intended branch, or possibly until even later, if the two branches have common steps.

Ideally, we would like to detect deviations before harm is done as a result of these deviations, but the presence of detection delay could prevent achieving this goal in some situations. In this paper, we investigate how frequently delayed deviation detection could arise in medical processes, how often harm could occur because of the delay, and what the causes for detection delay could be. Harm could also occur as a result of a deviation even when there is no deviation detection delay, and we investigate this issue as well. We also investigate the performance of the deviation detection approach in terms of computation time. If the computation time to detect a deviation is too long, harm could be done before process performers are notified of the deviation.

Our initial evaluation of the deviation detection approach has been promising. Delayed deviation detection rarely arose in the medical processes we studied and when it did arise, no harm to the patient could occur as a result of the delay. The running time of the deviation detector is low relative to the human speed of performing process activities, suggesting that it could be used for online deviation detection.

Section 2 discusses related work, the proposed deviation detection approach, and some issues associated with that approach. The experimental methodology for evaluating the approach with respect to these issues is presented in section 3. The results are presented in section 4 and discussed in section 5. Section 6 summarizes the paper and describes future work.

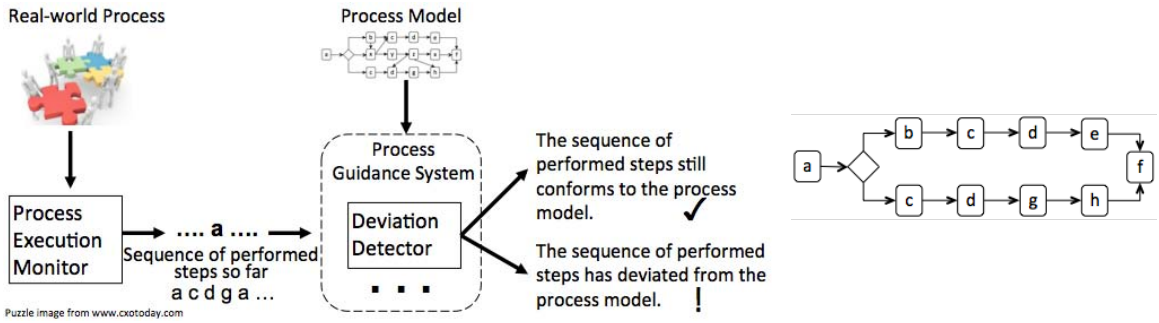
2 Background

2.1 Related Work

There are several approaches that aim to reduce the number of medical errors by encouraging conformance with some specification of the acceptable ways to perform a process (e.g., process aids such as checklists[5,6] and care sets[7]). Such process aids, however, tend to specify only the major steps during normal flow, omitting important details such as exceptional scenarios and concurrent process execution[8,9]. Such process aids often add to the workload of already heavily-burdened medical professionals. The use of checklists, for example, often requires medical professionals to check what needs to be done, to remember what steps they completed, and to determine the appropriate checklist to use in a given context.

To remove some of the burdens that process aids, such as checklists, place on medical professionals, there have been attempts to create systems that automatically check the compliance of an executing process with a specification of that process[11]. To support such systems, various medical guideline modeling languages have been developed[10]. Subsequent work has utilized formally specified medical guidelines to drive careflow management systems. For example, Fitzgerald et al. designed and deployed a careflow management system in a trauma center to guide medical professionals during the first 30 minutes of trauma resuscitation[11]. This system increased compliance with the modeled medical protocol and reduced error rates, but the model did not support complex process behaviors, such as concurrency and exception handling. The framework we are proposing, including the work reported in this paper, is intended to go beyond these limitations.

In prior work, we constructed detailed models of several medical processes[8,9,12] in consultation with medical professionals. We also constructed a framework and a set of tools for analyzing such models to detect various safety problems and evaluate proposed changes[12]. Although building and analyzing such detailed models takes considerable care and time, this work resulted in an improved understanding of the processes and the detection of several errors and vulnerabilities, leading to nearly 70% reduction in the number of errors reaching the patient in one of our studies[13].



(a) Process Monitoring and Guidance Framework.

(b) An Example Control Flow Graph

Figure 1: The Process Monitoring and Guidance Framework and an example of a simple CFG.

We are now beginning to examine the potential for using such models, validated by the analyses, for the monitoring and guidance of executing processes.

2.2 Deviation Detection Approach

The proposed deviation detection approach is a component of a framework we are developing for process monitoring and guidance. This framework also supports other process guidance aspects, such as pro-actively informing process performers of pending activities depending on how the execution of the process develops and providing capabilities for inspecting process execution history.

In this paper, we focus on the deviation detection aspect of the framework. Figure 1(a) shows the relevant components of the framework. As a process is being performed, the *Process Execution Monitor* captures events associated with the performed steps and incrementally creates the *sequence of performed step events*. We use the term *step* to refer to an activity of interest performed as part of a process by humans, software or hardware devices. For this paper, we assume that the sequence of performed step events is accurate, i.e., the sequence contains events for all steps of interest that have been performed, these events are in the order in which the actual steps were performed, and events that are not of interest have been accurately removed[†]. The precise way that the Process Monitor captures events is beyond the scope of this paper.

Every time the *Deviation Detector* receives a step event from the Process Execution Monitor, it checks if the corresponding sequence of steps performed so far is one of the acceptable sequences as specified by the *Process Model*. If it is not, a deviation is detected and process performers are warned that a potential error might have occurred.

Process Model. The process model captures the acceptable ways to perform the corresponding process. This includes the nominal ways, but also how the process should be performed when various exceptional situations arise. The model can also capture different ways to perform the process depending on the level of expertise of process performers.

Eliciting information from domain experts to create such process models and translating the elicited information into an actual model are known to be difficult and time consuming. In our work, we have used several process elicitation methods, such as observations, structured, and unstructured interviews[13]. After creating process models based on the elicited information, we have used several formal analysis techniques to validate these models[12]. Regardless of how extensive the process elicitation and model validation efforts are, however, it is likely that process models will still contain some inaccuracies. Furthermore, process performers might execute a process in innovative and yet acceptable ways that are not captured in the process model. In practice, we expect process models to be continually updated as defects in the models are uncovered and/or as the modeled processes change.

In our experience, to support the complexity of medical processes, the process model needs to be written in a notation with rich and well-defined semantics. Specifically, such a notation should provide support for modeling human choice, exception handling, concurrency, and synchronization. For our work, we have chosen the Little-JIL[14] process modeling language, because it satisfies these requirements and because we have found that the Little-JIL's compact visual notation has facilitated communication with the medical professionals who have been helping us create

[†]To simplify the discussion, hereafter we shall refer to the sequence of step events associated with the sequence of actual performed steps as the sequence of performed steps.

and validate the models of the relatively large and non-trivial processes on which we have been working (two of these processes are briefly described in section 3.1). The proposed deviation detection approach, however, is independent of the process modeling language; it can be used with any language that has well-defined semantics that are also sufficiently rich to capture complex real-world processes. For the purposes of deviation detection, a Little-JIL model is translated into a typical control flow graph (CFG) representation of the possible step execution orderings associated with a process model, so in this paper only a simple control flow graph representation is shown.

Deviation Detector. The Deviation Detector traverses the CFG representation in a breadth-first way to determine whether the sequence of performed steps so far is a legal sequence of steps through the graph. For efficiency, this traversal is done incrementally, maintaining a frontier of possible nodes in the graph that could correspond to the last performed step. When this frontier becomes empty, i.e., when there is no path through the graph that corresponds to the sequence of performed steps, a deviation is detected.

2.3 Example of Applying the Deviation Detection Approach to a Medical Process

Consider the process of transfusing a unit of blood to a patient. This process usually starts after a nurse receives a physician's order to perform a transfusion. The nurse then needs to check if the patient's blood type and screen are known and, if they are not, obtains a blood sample from the patient and sends it to the lab for analysis. Once the type and screen are known, the nurse can request the appropriate blood product from the blood bank and pick it up when it becomes available. To administer a single unit of blood product, the nurse needs to provide some documentation and perform a series of checks to ensure that the right patient receives the right blood product at the right time. During the transfusion, the nurse needs to periodically check for complications (such as a transfusion reaction) and, if any arise, handle them appropriately. Once the transfusion is done, the nurse needs to evaluate the patient, create final documentation, and dispose of the infusion materials.

Most of the activities in the blood transfusion process consist of a fair number of subactivities that can occur in different orders and potentially different number of times. Problems could arise while performing many of these activities, making the blood transfusion process challenging to perform correctly and potentially error-prone[8].

A common error reported in the medical literature, and one that can cause severe harm to patients, is not fully following the procedure for verifying the patient's identity[8] while performing the series of checks before beginning the infusion. A possible instance of this error is omitting to verify that the patient is wearing the correct ID band. Consider the situation where in a busy emergency department patient X is wearing the incorrect ID band—that of patient Y. Perhaps a registration clerk had to place ID bands on several patients and inadvertently switched the ID bands. Suppose that patient Y is the one for whom a blood transfusion was ordered. Since patient X is wearing patient Y's ID band, if the nurse does not verify patient X's ID band prior to infusing the blood, patient X might receive the blood ordered for patient Y. Note that the nurse might still successfully perform other checks, such as the blood product checks, since they are done against the ID band and not against the patient's real identity.

Potential harm as a result of this error might be avoided if the nurse is warned that the process is being performed incorrectly before the infusion is started. The deviation detector could achieve this is by comparing the sequence of steps the nurse has performed against a model of the process. The sequence of steps that omits the step *verify that the patient is wearing the correct ID band* would not be a legal sequence through the process model and the deviation detector could establish that. Informing the nurse about such a deviation might help the nurse recover from the error before harm is done (i.e., infusing blood into the wrong patient).

2.4 Issues

Delayed Deviation Detection and Potential Harm Due to Delay. As discussed earlier, delayed deviation detection is an important issue to consider as it might constitute a significant threat to the usefulness of the proposed deviation detection approach, depending on how often delayed deviation detection arises and the severity of the consequences from the delay. For a concrete illustration of delayed deviation detection, consider the example control flow graph in Figure 1(b), assumed to have been created from a process model. In this CFG, nodes represent process steps and are labeled with letters. There is an edge from one node to another, if the step the first node represents can be immediately followed by the step the second node represents in the modeled process. The diamond node represents a branch point. When no branch conditions are modeled, such as when a medical professional needs to make a decision based on personal judgment and experience, the deviation detector cannot determine which branch should be taken in a particular process execution and must consider all options.

The two legal sequences of steps allowed by the CFG in Figure 1(b) are *abcdef* and *acdghf*. Suppose that during a particular execution of the process, the performers planned to carry out the sequence *abcdef*, but forgot to perform step *b*. In this situation, the deviation occurs when *b* is omitted, but it cannot be detected until *e* is performed, because *acd* is a valid sequence through the model. In this situation, we say that there is a detection delay of 2 (we measure the delay in terms of performed steps between the error—the omission of step *b*—and the step when the deviation is detected—step *e*). If the two branches in Figure 1(b) had a longer sequence of identical substeps (in this case this sequence is *cd*), the deviation detection delay could be even longer. In principle, the delay can be arbitrarily long and serious harm could potentially result from not detecting a deviation at the time when an error is made.

In this work, we explore the following research questions related to delayed deviation detection: How often do deviation detection delays occur in complex medical processes? How long are deviation detection delays? How harmful are deviation detection delays? What causes deviation detection delays and can they be reduced or avoided?

These questions could potentially be analytically explored by statically analyzing a set of process models. We have not been able to develop a reasonable static approach, however, that can handle multiple kinds of planning errors effectively (e.g., single-step omission, single-step commission, single-step substitution, omission of n consecutive steps, …, omission of n not necessarily consecutive steps, …, single subprocess omission, …, multiple subprocess omission, etc.). Thus, as a first step in exploring the above questions, we chose an experimental approach that applies the deviation detector to synthetically generated erroneous executions from realistic models of two medical processes.

Potential Harm When Deviations Are Detected without Delay. It is sometimes possible that deviations are detected immediately as an error is made (there is no detection delay), but harm could potentially still occur as a result of the deviations. For example, in the blood transfusion process discussed above, one of the checks that needs to be performed before infusing the blood is to ensure that the blood product has not expired. Suppose that the process model allows this check to occur immediately before infusing the blood. If the nurse forgets to check that the blood product has not expired, the deviation will be detected when the nurse starts the infusion. Even though there is no detection delay, harm could still occur as the patient might receive expired blood. This is an example of a vulnerability in the process model or in the process itself (assuming the model is accurate) with respect to the proposed deviation detection approach—a deviation cannot be detected until the potentially harmful step is already started.

In this work, we explore the following research questions related to the above issue: How often do situations arise where harm could occur even when deviations are detected without delay? What can be done to reduce or avoid such situations?

Running Time of the Deviation Detector. It is important that the deviation detector does not take too long to compute whether a deviation has occurred after a new step is added to the sequence of performed steps. Otherwise, if a deviation occurs, the warning about that deviation might get issued after harm was already done as a result of the deviation. We explore the following research question related to the performance of the deviation detector: What is the running time of the deviation detector when applied to realistic medical processes?

3 Methods

To perform an initial evaluation of the proposed deviation detection approach with respect to the above issues, we applied it to models of two medical processes—chemotherapy and blood transfusion. We generated sequences of performed steps from these models and then mutated the sequences to represent process executions with simple errors. Our evaluation approach is based on *mutation testing*[16], a common software engineering technique where a computer program is systematically mutated (usually by making a predefined set of simple changes to the source code) to evaluate software testing and analysis approaches.

3.1 Process models

The chemotherapy process model we used was elicited from medical professionals participating in an outpatient chemotherapy ordering and administration process in a regional cancer center in Western Massachusetts. At the time of writing this paper, the chemotherapy process model covers multiple phases of the chemotherapy process, including diagnosing the patient and ordering chemotherapy; thorough review of the treatment plan and medication orders by a medical assistant, a nurse, and a pharmacist; conducting an informational/teaching session with the patient and obtaining informed consent form; preparing chemotherapy drugs, performed by a pharmacist and pharmacy technicians; assessing the patient and administering the drugs, performed by a clinical nurse[13].

The blood transfusion process model is based on a standard blood transfusion checklist from the medical literature[5] and includes additional information about exceptional situations that might arise during the process and their handling. This model is part of a blood transfusion benchmark[15] that we developed with a nursing faculty member working on patient safety[8]. The model covers the process activities starting from receiving a physician order for transfusion to infusing the blood into the patient and performing subsequent follow-through checks. It specifies multiple phases of the process, including verifying that the patient blood type and screen are available (and if they are not, performing the subprocess of obtaining and labeling a blood specimen); ordering and obtaining a blood product from the blood bank; performing various verifications on the patient and on the blood product before starting the transfusion; monitoring the patient during the transfusion and appropriately reacting if a transfusion reaction is suspected.

The chemotherapy and blood transfusion process models are of significant size and complexity. The chemotherapy process model includes 283 steps performed by human process performers and specifies 59 exception handling situations[‡]. The corresponding low-level CFG representation has 2,358 nodes and 701,887 edges. The blood transfusion process model includes 102 steps, including 63 exception handling situations, and the corresponding CFG representation has 97,237 nodes and 242,442,845 edges. Even though the blood transfusion model has fewer steps than the chemotherapy model, the exception handling behavior in the blood transfusion process is more complex, requiring a large number of low-level CFG nodes and edges.

3.2 Sequences with synthetic errors

To evaluate the deviation detection approach, we applied it to step sequences containing typical planning errors. Although there is no standardized taxonomy of human errors[1,17], two error kinds appear in the intersection of most of the proposed planning error taxonomies[1,3,18]—omission and commission errors. An omission error occurs when one or more steps are not performed; a commission error occurs when one or more wrong step(s) are performed.

In our initial evaluation of the deviation detection approach, we focused on several kinds of planning errors that involve a single step and on the error of omission of a single subprocess. This decision was based on the errors available from a study of common errors in the blood transfusion process[8] and from conversations with blood transfusion and chemotherapy domain experts.

Single-Step Errors. From each of the two process models, we generated 50 sequences of steps by performing random walks through the process model. To represent "nominal-path" process executions, we also generated 50 sequences from each of the models but disallowed exceptional situations. Thus, we created 100 legal sequences for each process. Statistics about the lengths of the generated sequences are shown in Figure 2.

Each generated sequences was then mutated to represent a process execution in which the process performers deviated from the acceptable ways to perform the process. The applied mutations were deletion, insertion, and substitution of a single step at almost[§] every position of each sequence. A mutated sequence is called a *mutant*.

Subprocess Errors. Based on common blood transfusion and chemotherapy errors reported in the medical literature and on our interaction with domain experts involved in these processes, we identified 5 blood transfusion and 5 chemotherapy subprocesses deemed most likely to be omitted. The selected blood transfusion subprocesses were: (1) *ensure correct patient is present* (performed by a nurse before notifying the blood bank to prepare the blood to prevent the possibility the blood product to expire because the patient is not available for transfusion or the wrong patient is in the room); (2) *verify patient ID band* and (3) *verify blood product information* (part of the bedside checks performed by the nurse before beginning the infusion); (4) *assess patient* and (5) *evaluate patient clinically* (part of the clinical evaluation prior to the infusion, performed again by the nurse). The selected chemotherapy subprocesses were: (1) *record height and weight* (performed during patient registration by a clerk); (2) *confirm all necessary information is present* (part of the consultation and assessment, performed by an oncologist before creating the treatment plan and chemotherapy orders); (3) *confirm pretesting has been done* and (4) *confirm existence and not staleness of height/weight data* (part of the treatment plan and orders verifications performed by a Practice Registered Nurse (RN)); (5) *obtain patient informed consent* (performed by a Nurse Practitioner or a Clinic Nurse prior to chemotherapy administration).

For each identified subprocess, we generated 50 sequences from the corresponding process model, such that these sequences contain the subprocess. We then mutated each generated sequence by deleting all steps pertaining to the specific subprocess selected to be omitted. The deviation detection approach was applied to each mutant and various statistics were collected.

[‡]An exception handling situation usually involves multiple steps to deal with the exception. A step (e.g., *assess patient*), can be part of different subprocesses and be involved in different nominal and exceptional situations (e.g., patient develops an allergic reaction).

[§]There were a small number of cases, such as deleting the last step from a legal sequence, where the resulting mutated sequence did not correspond to a sequence with an error. Such cases were excluded from our analysis.

	Process definition	Blood transfusion process						
Set-up	Original traces	50 random			50 random, no exceptions			50 random sequences for each of 5 subprocesses
	Avg. trace length (number of steps)	21.52			70.62			70.54
	Min. trace length (number of steps)	4			69			68
	Max. trace length (number of steps)	114			73			73
	Mutation kind	Exp. 1: Deletion at every position (except last)	Exp. 2: Insertion before every position	Exp. 3: Substitution at every position	Exp. 4: Deletion at every position (except last)	Exp. 5: Insertion before every position	Exp. 6: Substitution at every position	Exp. 7: Deletion of all steps belonging to subprocess of interest
Results	Number of mutants	1026	1076	1076	3481	3531	3531	250
	Number of mutants where deviation is detected after the mutation index	0 (0.00%) [27] *	6 (0.56%) [1] ♮	5 (0.47%) [1] ♭	6 (0.17%) [294] *	25 (0.7%) [7] ♮	19 (0.54%) [1] ♭	0 (0.00%) [23] *
	Avg. deviation detection delay (number of steps), for mutants with detection delay	0.00	1.00	1.00	1.00	1.08	1.00	0.00
	Min. deviation detection delay (number of steps), for mutants with detection delay	0	1	1	1	1	1	0
	Max. deviation detection delay (number of steps), for mutants with detection delay	0	1	1	1	2	1	0
	Number of mutants for which delay could be "potentially harmful"	0	0	0	0	0	0	0
	Number of mutants without deviation detection delay for which harm could potentially occur due to the deviation	1 (0.01%)	24 (2.2%)	35 (3.3%)	50 (1.4%)	96 (2.72%)	90 (2.52%)	50 (20.00%)
	Avg. time per mutant (sec.)	5.98	5.88	5.74	13.01	13.12	13.09	11.17
	Avg. time per step (sec.)	0.28	0.27	0.27	0.18	0.19	0.19	0.16

* Number of mutants for which delay was due to a deletion in a shuffle region and deviation was detected right after the shuffle region.
♮ Number of mutants for which the delay was due to insertion of a step from a shuffle region in the model before the step has occurred in the corresponding shuffle region in the original sequence.
♭ Number of mutants for which delay was due to subbing in a step from a shuffle region in the model before the step has occurred in the corresponding shuffle region in the original sequence.
For the mutants in square brackets we used the "minimum interpretation" for the deviation detection delay (see the Discussion section for more detail).

Figure 2: Applying the deviation detection approach to the blood transfusion process model. The results for the chemotherapy process model are available at http://laser.cs.umass.edu/deviation-detection/amia2014.html.

4 Results

Figure 2 shows the results of applying the deviation detection approach to the blood transfusion process. Due to lack of space, we do not include the results table for the chemotherapy process, but make it available at http://laser.cs.umass.edu/deviation-detection/amia2014.html. The results from both processes were similar and are discussed in the next section. A deviation detection *delay* is defined as the number of steps between the *mutation index* (the index in the sequence of steps where the mutation was done for single-step errors or the index of the first mutation for subprocess errors) and the index where that sequence was recognized as not being a sequence from the process model.

5 Discussion

Delayed Deviation Detection and Potential Harm Due to Delay. Delayed deviation detection occurred infrequently—in less than 1% of the mutants from all experiments. We analyzed each mutated sequence with deviation detection delay by tracing that sequence through the corresponding process model to determine what structure(s) in that model caused the delay. The main cause for the delay was branching in the process models due to exception handling or optional steps/subprocesses. For example, suppose a process can be performed by executing step sequence *abcd*, when there are no exceptional situations. If an exceptional situation arises while performing *b*, however, then the sequence of steps *xyz* should be performed to address this situation before continuing with steps *c* and *d*, resulting in sequence *abxyzcd*. If the sequence *abcd* is mutated by inserting step *x* after step *b*, the deviation cannot be detected until step *c* is performed (which is one step after the mutation index) because *abx* is a prefix of valid sequence through the process model.

Another reason for detecting deviations after the mutation index were *shuffled* steps, which are steps that should be done sequentially but are allowed to occur in any order. For example, suppose that a process is performed by doing step *a*, followed by steps *b*, *c*, and *d* in any order, and then step *e*. Thus, if the step sequence *abcde* is mutated by deleting step *b*, a deviation will not get detected until step *e* is performed because *acd* is a prefix of a valid sequence through the process model. This results in a deviation detection delay of 2, using the measure of delay previously described.

Shuffled steps are different from other branch points in a process where there is more than one step to perform next, because any of the shuffled steps are acceptable to be performed next. Since shuffled steps are common in the processes we studied, we decided that deviation detection delays due to shuffled steps should be measured differently to avoid distorting the results. For example, if a mutation deletes one of several shuffled steps in a sequence of steps, we measure the delay as the number of steps between the last shuffled step in the sequence and the step where a deviation is detected. We call this the *minimum interpretation of the delay* and we use similar minimum interpretations of delay

when insertion and substitution mutations involving shuffled steps are performed.

In each of the experiments with the blood transfusion and chemotherapy processes, the average deviation detection delay was small—the largest average was 2.31 steps—and in most cases it was close to 1. In critical processes, such as medical procedures, however, even a delay of 1 could be harmful, if some potentially dangerous step, such as *administer chemotherapy medications*, is performed before detecting the deviation.

To address the question of whether/how often harm might occur due to delayed deviation detection, we identified a set of potentially harmful steps for both processes. These are steps, such that if an error has occurred prior to their performance, performing them could potentially result in immediate harm. Two such steps from the blood transfusion and chemotherapy processes are *begin infusion of blood product* and *administer chemotherapy drug*, respectively.

Having identified the set of potentially harmful steps, we then inspected the mutated sequences described above to determine whether a potentially harmful step occurs between the mutation index and the index of deviation detection. For sequences for which this was the case, we manually analyzed whether the error (the mutation) could affect the potentially harmful step. If harm could occur as a result of the error, we counted the mutated sequence as one for which deviation detection delay could be potentially harmful. There were no cases where the delay could result in harm for the two processes we examined.

Potential Harm When Deviations Are Detected without Delay. We found some mutated sequences for which the deviation is detected at the index of mutation, i.e., there is no deviation detection delay, but harm could potentially still be done (third row from the bottom of the table in Figure 2). In experiments 2, 3, 5, and 6, this was due to mutating a sequence by inserting or substituting in a harmful step. In general, if the error is a commission of a harmful step (e.g., *administer a drug*), a deviation cannot be detected before the harmful step is started because the sequence of performed steps up to the index where the error is made would be a legal sequence through the process model.

In experiments 1, 4, and 7, the cases where potential harm could occur, even when there is no deviation detection delay, were due to omitting a step or a subprocess that a) can immediately precede a harmful step, and b) can affect that harmful step. For example, in the blood transfusion process model, the subprocess of verifying the blood product immediately precedes the step *begin infusion of blood product*. Thus, if steps from this subprocess (such as *ensure that blood product has not expired*) are omitted, or the subprocess is omitted altogether, the deviation cannot be detected until the step *begin infusion of blood product* is started. The 50 cases in experiment 7 of mutants in which harm could occur even when there is no detection delay are due to omitting the subprocess *verify blood product information*.

Such structures represent process vulnerabilities with respect to the proposed deviation detection approach. They also represent vulnerabilities in general, as there is little opportunity for process performers to realize that an error is made before they start a potentially harmful step. A possible strategy to deal with such process vulnerabilities is to introduce a non-harmful step before the potentially harmful one. In fact, such steps are already in place in some medical procedures, such as surgeries where the clinicians performing the surgery are required to stop at certain points of the procedure and confirm that everyone has performed the necessary steps and is aware of the relevant information before proceeding further. The presence of such verification steps would allow the proposed deviation detection approach to detect deviations when such verification steps are performed and before potentially harmful steps are started.

Running time of the deviation detector. The running time results for the blood transfusion process are shown in the last two rows of the table in Figure 2. It took less than 6 seconds to determine whether a sequence of about 21.5 steps, on average, is a deviant; for the mutated sequences where exceptions were disallowed, it took around 13 seconds to determine whether a sequence of about 70.5 steps, on average, is a deviant. This results in less than a third of a second per step in a sequence. The running time of the deviation detector was similarly low for the chemotherapy process. Given that humans usually take more than a third of a second to perform an activity in a medical process, the running time results indicate that the deviation detector could detect deviations in real time before harm is done as a result of a deviation.

Threats to validity. The experimental evaluation was synthetic—we did not monitor a real executing process and did not apply the deviation detection approach to real process executions. The experimental evaluation was based on models of two medical processes. Even though the models were relatively large and complex in terms of covering a large set of process executions (including exceptional executions and concurrency within a single process execution), the results might change if the size or the complexity of the models increases or models of different processes are used.

We mutated the generated sequences by performing single-step deletions, insertions, and substitutions and also deletion of subprocesses. The results might change if mutations that represent different kinds of errors are performed or if multiple errors are considered.

Limitations of proposed deviation detection approach. Even though the preliminary investigation of the proposed deviation detection approach is promising, there are several research challenges that need to be tackled before

the approach can be applied in clinical settings. As discussed in section 2.2, the deviation detection approach relies on receiving an accurate sequence of performed steps of interest from the Process Execution Monitor. Capturing what human process performers do in an accurate and timely manner, however, may be difficult. We expect the use of electronic devices in processes will facilitate process execution monitoring. For example, starting to receive infusion data from an infusion pump could be automatically interpreted as having started the step *begin infusion of blood product*; similarly, when patient height and weight data are entered in an electronic medical record, this could be an indication that the step *measure height and weight* was performed.

While the use of electronic devices in processes increases the opportunities for monitoring process executions, events from electronic devices could be misinterpreted. Furthermore, electronic devices cannot capture certain steps in processes, such as cognitive steps (e.g., a doctor making sure the patient information on two artifacts matches). Human scribes, whose participation in medical processes seems to be increasing[4], can capture some steps that electronic devices cannot.

The success of the proposed deviation detection approach depends on the accuracy and the appropriateness of the process model as well. Creating a high-quality process model, however, can be challenging, given the complexity of some medical processes. We have been investigating techniques for eliciting and validating process models and have applied them successfully to several real-world processes[9,13,19]. We believe the criticality of certain processes warrants the time and effort needed to create high-quality process models that can in turn be leveraged to support continuous process improvement via various static analyses (e.g., model checking[20], fault-tree analysis[21], and failure-mode and effets analysis[22]) and to support deviation detection and other aspects of online process guidance (e.g., smart checklist[23]).

6 Conclusion and Future Work

This paper proposes an approach for online deviation detection to catch human errors in complex medical processes before harm is done. The approach relies on a mechanism for monitoring an executing process and on a detailed model that specifies how the process should be performed. We identify important issues related to the feasibility of this approach and evaluate that approach by applying it to detailed realistic models of two medical processes.

The initial evaluation of the deviation detection approach is promising. Delayed deviation detection rarely arose in the studied medical processes and when it did arise, no harm to the patient could occur as a result of the delay. The running time of the deviation detector is low relative to the human speed of performing process activities, suggesting that it could be used for online deviation detection. We also identify some vulnerabilities in the studied process models with respect to online deviation detection and suggest approaches for reducing such vulnerabilities.

The initial investigation of the deviation detection approach has suggested some interesting future research directions. When a deviation is detected, it might be useful to not only inform process performers about it, but to also provide some additional information that could help them identify potential error(s) and plan recovery actions. Examples of such information are possible indexes in the sequence of performed steps where an error might have occurred, as well as a ranked list of potential errors. We are exploring approaches for providing such information.

In this work, we studied the delayed deviation detection issue empirically. It will be useful, however, to also develop analytical approaches for determining an upper bound on possible deviation detection delays given a process model and assumptions about the errors that might occur. We are currently exploring such approaches. We are starting by assuming that only simple errors can occur (e.g., a single omission of a single step) before considering more complex errors and combinations of different errors.

Given the initial evaluation of the deviation detection approach and the recent developments in healthcare that could enable monitoring of executing processes, we believe the proposed approach can augment current approaches for online guidance (such as checklists) in medical processes and help catch errors before harm is done.

Acknowledgments

This material is based upon work supported by the National Science Foundation under awards IIS-1239334 and CMMI-1234070. The authors gratefully acknowledge the contributions of Lee Osterweil, Heather Conboy, Elizabeth Henneman, Jenna Marquard, and of many members of the staff of the D'Amour Center for Cancer Care, who graciously donated their time and expertise.

References

[1] Kohn LT, Corrigan JM, Donaldson MS, editors. To Err is Human: Building a Safer Health System. Washington, DC: National Academies Press; 1999.

[2] Stead WW, Lin HS, editors. Computational Technology for Effective Health Care: Immediate Steps and Strategic Directions. National Academies Press; 2009.

[3] James JT. A New, Evidence-based Estimate of Patient Harms Associated with Hospital Care. Journal of Patient Safety. 2013;9(3):122–128.

[4] Hafner K. A Busy Doctor's Right Hand, Ever Ready to Type. The New York Times. 2014 January 12;.

[5] Wilkinson JM, Leuven KV. Procedure checklist for administering a blood transfusion;.

[6] World Health Organization. Surgical Safety Checklist; 2008.

[7] Mertens WC, Brown DE, Parisi R, Cassells LJ, Naglieri-Prescod D, Higby DJ. Detection, Classification, and Correction of Defective Chemotherapy Orders Through Nursing and Pharmacy Oversight. Journal of Patient Safety. 2008;4(3):195–200.

[8] Henneman EA, Avrunin GS, Clarke LA, Osterweil LJ, Andrzejewski Jr C, Merrigan K, et al. Increasing Patient Safety and Efficiency in Transfusion Therapy Using Formal Process Definitions. Transfusion Medicine Review. 2007 January;21(1):49–57.

[9] Chen B, Avrunin GS, Henneman EA, Clarke LA, Osterweil LJ, Henneman PL. Analyzing medical processes. In: ICSE '08: Proceedings of the 30th International Conference on Software Engineering. ACM; 2008. p. 623–632.

[10] Peleg M, Tu S, Bury J, B MBC, Fox J, Greenes RA, et al. Comparing Computer-Interpretable Guideline Models: A Case-Study Approach. JAMIA. 2002;10:2003.

[11] Fitzgerald M, Cameron P, Mackenzie C, Farrow N, Scicluna P, Gocentas R, et al. Trauma Resuscitation Errors and Computer-Assisted Decision Support. Archives of Surgery. 2011;146(2):218–225.

[12] Avrunin GS, Clarke LA, Osterweil LJ, Christov SC, Chen B, Henneman EA, et al. Experience modeling and analyzing medical processes: UMass/baystate medical safety project overview. In: Proceedings of the 1st ACM International Health Informatics Symposium. IHI '10. New York, NY, USA: ACM; 2010. p. 316–325.

[13] Mertens W, Christov S, Avrunin GS, Clarke LA, Osterweil LJ, Cassells LJ, et al. Using Process Elicitation and Validation to Understand and Improve Chemotherapy Ordering and Delivery. The Joint Commission Journal on Quality and Patient Safety. 2012 November;38(11):497– 505.

[14] Cass AG, Lerner BS, Stanley M Sutton J, McCall EK, Wise A, Osterweil LJ. Little-JIL/Juliette: a process definition language and interpreter. In: ICSE '00: Proceedings of the 22nd International Conference on Software Engineering. ACM; 2000. p. 754–757.

[15] Christov SC, Avrunin GS, Clarke LA, Osterweil LJ, Henneman EA. A benchmark for evaluating software engineering techniques for improving medical processes. In: Proceedings of the 2010 ICSE Workshop on Software Engineering in Health Care. SEHC '10. New York, NY, USA: ACM; 2010. p. 50–56.

[16] DeMillo RA, Lipton RJ, Sayward FG. Hints on Test Data Selection: Help for the Practicing Programmer. Computer. 1978 Apr;11(4):34–41.

[17] Henneman EA, Blank FSJ, Gattasso S, Williamson K, Henneman PL. Testing a classification model for emergency department errors. Journal of Advanced Nursing. 2006;55(1):90–99.

[18] Hollnagel E. The phenotype of erroneous actions. Int J Man-Mach Stud. 1993 July;39:1–32.

[19] Christov SC, Chen B, Avrunin GS, Clarke LA, Osterweil LJ, Brown D, et al. Formally Defining Medical Processes. Methods of Information in Medicine Special Topic on Model-Based Design of Trustworthy Health Information Systems. 2008;47(5):392–398.

[20] Clarke MCEM, Grumberg O, Peled DA. Model Checking. MIT Press; 2000.

[21] Vesely WE, Goldberg FF, Roberts NH, Haasl DF. Fault Tree Handbook (NUREG-0492). U.S. Nuclear Regulatory Commission, Washington, D.C.; 1981.

[22] Stamatis DH. Failure Mode and Effect Analysis: FMEA from Theory to Execution. American Society for Quality; 1995.

[23] Avrunin GS, Clarke LA, Osterweil LJ, Goldman JM, Rausch T. Smart checklists for human-intensive medical systems. In: 42nd International Conference on Dependable Systems and Networks Workshops (DSN-W), Workshop on Open, Resilient, Human-aware, Cyber-physical Systems, IEEE/IFIP; 2012. p. 1–6.

Adapting a Clinical Data Repository to ICD-10-CM through the use of a Terminology Repository

James J. Cimino, MD; Lyubov Remennick, MD
Laboratory for Informatics Development, NIH Clinical Center, Bethesda, MD

Clinical data repositories frequently contain patient diagnoses coded with the International Classification of Diseases, Ninth Revision (ICD-9-CM). These repositories now need to accommodate data coded with the Tenth Revision (ICD-10-CM). Database users wish to retrieve relevant data regardless of the system by which they are coded. We demonstrate how a terminology repository (the Research Entities Dictionary or RED) serves as an ontology relating terms of both ICD versions to each other to support seamless version-independent retrieval from the Biomedical Translational Research Information System (BTRIS) at the National Institutes of Health. We make use of the Center for Medicare and Medicaid Services' General Equivalence Mappings (GEMs) to reduce the modeling effort required to determine whether ICD-10-CM terms should be added to the RED as new concepts or as synonyms of existing concepts. A divide-and-conquer approach is used to develop integration heuristics that offer a satisfactory interim solution and facilitate additional refinement of the integration as time and resources allow.

Introduction

Many patient diagnoses are recorded in electronic databases using codes from the International Classification of Diseases, Ninth Revision (ICD-9-CM, or I9 for short).[1] When these data are included in longitudinal databases, such as clinical data repositories, challenges arise with each annual I9 update. For example, when searching for patients with septic shock, a researcher looking for data prior to 2008 would need to use the I9 codes *785.59*[*] and *998.02* (*Shock without mention of trauma, not elsewhere classified* and *Postoperative shock, septic*, respectively). However, a longitudinal study of patient records coded with the I9 code *785.59* would show a sharp decrease in incidence in 2008. Unfortunately, this is not due to development of effective measures for the prevention of septic shock but rather due to the addition of the I9 code *785.52* (*Septic shock*), with a concomitant change in the implicit meaning of *785.59* and a corresponding decrease in its use to code data. Researchers are often unaware of how such changes impact their work.[2] Mitigating impact of updates on the use of data repositories requires special measures.[3]

Such challenges are likely to be even greater for US researchers when, in 2014, I9 is replaced by the International Classification of Diseases, Tenth Revision (ICD-10-CM, or I10 for short).[4] I10 is much larger than I9; some of this increase is due to more finely grained terms (for example, *Postoperative Shock, Septic* is being replaced by the three terms *Postprocedural septic shock, initial encounter*, *Postprocedural septic shock, subsequent encounter*, and *Postprocedural septic shock, sequela*). I10 also contains terms that are similar, but not synonymous, with I9 terms (for example, *Septic shock* is being replaced by *Severe sepsis with septic shock*). Past I9 updates, involving dozens or hundreds of changes pale in comparison to the change that I10 brings, with tens of thousands of new codes.

We have previously described a method for coping with annual I9 updates through the use of a controlled terminology resource that maps multiple external terminologies to a single, coherent terminology for use within a single institutional clinical data repository.[3] This paper describes the application of that methodology to adapt to the inclusion of I10 data into a repository that contains patient diagnostic data coded in part with I9.

Background

The Biomedical Translational Research Information System (BTRIS)

BTRIS is a repository of clinical research data collected at the Clinical Center, a 240-bed hospital on the National Institutes of Health campus in Bethesda, Maryland. It includes data from electronic health records and clinical trials management systems dating back to 1953. Along with clinical notes, laboratory test results, procedure reports, vital sign measurements, medication administration records, electrocardiograms, radiologic images, mass spectrograms, and exome data, BTRIS includes 1.9 million discharge diagnoses assigned to hospital admissions, coded in I9.[5]

[*] In this paper, codes and terms from the various versions of the International Classification of Diseases will be distinguished by the use of a different typeface.

The Research Entities Dictionary (RED)

The RED is a terminology repository modeled after Columbia University Medical Center's Medical Entities Dictionary (MED).[6] The RED contains all the terms from the various controlled terminologies that are used by each of the BTRIS data sources to code their data. Each term is mapped to a corresponding concept in the RED, usually in a one-to-one manner, although where terms from multiple sources are clearly synonymous, they may be mapped to single concepts. Each concept is assigned a unique identifier (the "RED Code"), which is stored with the source data in BTRIS.

RED concepts are organized in a directed acyclic graph of hierarchical is-a relationships that can be used to query BTRIS in a class-based manner. Using a single RED code, one can find all patients with a diagnosis of any infectious disease, any foodborne infectious disease, any disease caused by the infectious organism Salmonella, or a particular disease (e.g., typhoid fever). Terms from disparate data sources are organized into the same classification structure so that, for example, a query for a disease term will retrieve data coded with I9 codes from the current EHR and previous Clinical Center EHRs, rare disease terms coded by the Medical Records Department, and problem list terms from one of the clinical trials management systems. Figure 1 shows an example of the RED hierarchy.

RED maintenance is an ongoing task, involving two full-time ontologists. Changes to source terminologies must be characterized as changes in syntax and/or semantics. For example, if a data source changes the name of a term (syntax), a determination must be made as to whether the change reflects a minor name change (no change in meaning, requiring a simple update to the term information assigned to the corresponding RED concept) or a major name change (a change in meaning, requiring retirement of the existing RED concept the creation of a new one).[3]

The addition of a new data source to BTRIS usually requires addition of one or more new terminologies in the RED. Here the task is somewhat different. The first priority is to make sure there is a RED concept that corresponds to each source term; this allows data to be stored in the BTRIS database and supports simple queries (e.g., "get all the patient diagnoses"). The second priority is to place the new concepts into the existing RED hierarchy, or add new hierarchical concepts, as appropriate to the meaning of the terms; this supports the more sophisticated class-based queries (as described in Figure 1). Ideally, if the new term is synonymous with an existing source term, it will be added as a synonym to an existing RED concept. However, this usually requires time-consuming manual review and is actually the lowest priority. Redundant concepts clutter the RED and might confuse a user looking for terms with which to search BTRIS, but they do not adversely affect retrieval of data and can be corrected retroactively by merging concepts in the RED and replacing RED Codes in BTRIS.

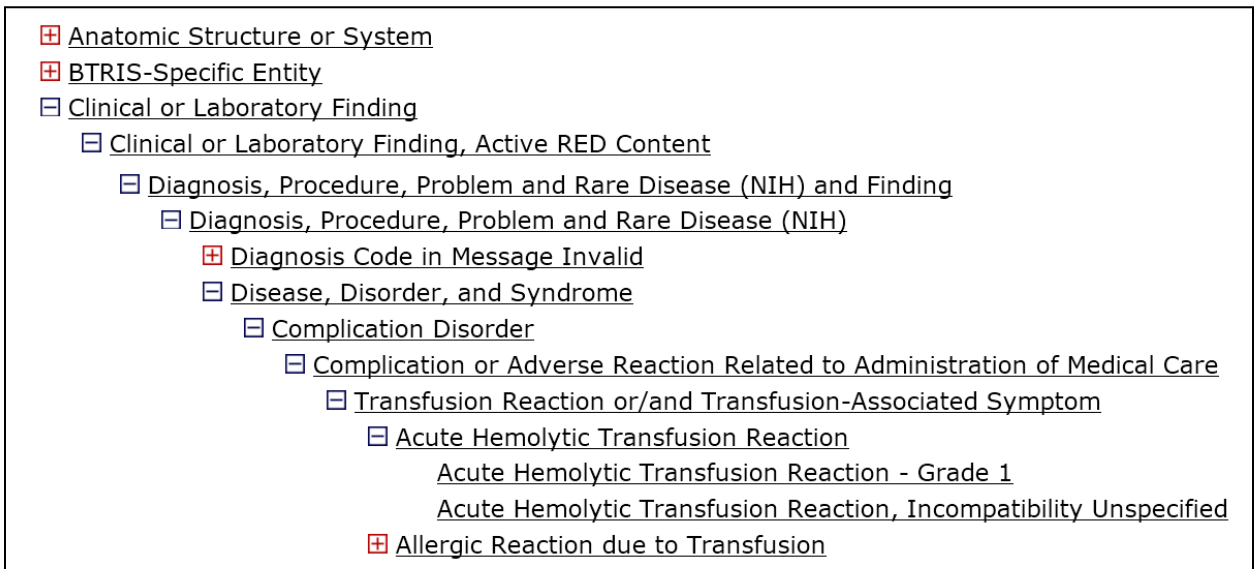


Figure 1: A sample of concepts in the RED hierarchy. The concept "Acute Hemolytic Reaction – Grade 1" corresponds to a term from the system in the Clinical Center's Department of Transfusion Medicine, while "Acute Hemolytic Transfusion Reaction, Incompatibility Unspecified" corresponds to the I9 term with the code *999.84*. Note that a BTRIS user wishing to retrieve data about acute hemolytic reactions can simply request the parent concept, "Acute Hemolytic Transfusion Reaction" and data coded by all three concepts will be retrieved.

General Equivalence Mappings (GEMs)

The National Center for Health Statistics (NCHS) has developed a bidirectional general equivalence mapping (GEM) system to assist users of I9 and I10 in deciding how to change their coding practices and understand how data coded in I9 and I10 relate to each other. NCHS, together with the Centers for Medicare and Medicaid Services (CMS) has published files that can be used as the basis for translating between I9 and I10 terms.[7]

The GEM mappings consist of two files, one in which each I9 code is listed at least once, with a set of flags associating it with one or more I10 codes, and one in which each I10 codes is listed at least once, with a set of flags associating it with one or more I9 codes.[8] The flags consist of five digits, with meanings as shown in Figure 2.

I10 Code	I10 Name	I9 Code	I9 Name	GEM	Meaning
A00.9	*Cholera, unspecified*	*001.9*	*Cholera, unspecified*	*00000*	Synonymous
A01.00	*Typhoid fever, unspecified*	*002.0*	*Typhoid fever*	*10000*	I10 term similar to I9 term
A01.01	*Typhoid meningitis*	*002.0*	*Typhoid fever*	*10000*	Another I10 term similar to the same I9 term
A02.1	*Salmonella sepsis*	*003.1*	*Salmonella septicemia*	*10111*	I10 term best expressed with two
A02.1	*Salmonella sepsis*	*995.91*	*Sepsis*	*10112*	I9 terms (one scenario)
R40.2131	*Coma scale, eyes open, to sound, in the field*			*01000*	No mapping in I9

Flag Position		Interpretation
1	"Approximate"	0=exact match (synonymous), 1=approximate match (similar meaning)
2	"No Map"	0=some plausible mapping; 1=no plausible mapping
3	"Combination"	0=mapping to single term, 1=mapping to multiple options ("scenarios"), multiple terms, or both
4	"Scenarios"	With the Combination flag=1, a coding option; may be one scenario (numbered "1") with multiple options or multiple scenarios (numbered "1", "2", etc.) each with one or more choices
5	"Choice"	With the Combination and Scenario flags, one or more options for a Particular scenario (numbered "1","2", etc.)

Figure 2: Understanding the General Equivalence Mapping Codes. Examples of the I10 to I9 GEM mappings are shown, with interpretation of each of the five flag positions. Note that GEM mapping files do not include term names; these were added here for clarity.

Methods

Our general approach to adding I10 terms to the RED is to examine the pairings with I9 codes provided in the GEMs and use the flags to suggest synonymy with or relationships to existing RED concepts. Relationships can include "child" (added to the hierarchy under the existing concept), "parent" (inserted into the hierarchy above the existing concept), or "sibling" (added to the hierarchy under the parent(s) of the existing concept). Using a "divide and conquer" approach,[9] we partitioned GEM mappings according to the following steps:

1. GEM flags were used to create four partitions: 00000 – *potential synonyms*; 10000 ("similar" flag set) – *potential children, parents or siblings*; 101xx ("combination" flag set) - *potential children, parents or siblings*; and 01000 ("no mapping" flag set) – *new concept*

2. We divided the 10000 ("similar") partition into four partitions based on the cardinality of terms involved the I10-to-I9 mappings: *one-to-one mappings*, *one-to-many mappings*, *many-to-one mappings* and *many-to-many mappings* (note that the "potential synonyms" partition is by definition one-to-one and the "combination" partition is by definition one-to-many or many-to-many but was not further subdivided; thus, the fourth and fifth flags were not considered further, as their complexity requires case-by-case review)

3. Each of the above partitions (except the *new concept* partition, in which there are no corresponding I9 terms) were further divided into based on the whether any of the I10 or I9 terms contained the word "other" or the phrase "not elsewhere classified"; this was done because the semantics of these terms can never be precisely determined (even the GEM mappings state that identical term names are not synonymous if they are "other" terms[10]); the phrases "not otherwise specified" and "other and unspecified" were excluded from this

consideration since they actually equivalent to the generic class terms; regardless of their original partition, "other" mappings were evaluated as *potential children, parents or siblings*

4. Within each of the resulting partitions (except the *new concept* partition, in which there are no corresponding I9 terms), we performed automated comparisons of the I10 and I9 terms names for similarity:
   - Name match: exact match between terms; treated as *potential synonyms*
   - Normalized name match: terms were normalized as follows prior to comparison: all letters capitalized, punctuation removed, words replaced with preferred forms using a set of previously described word synonyms,[11] stop words ("A", "AN", "AND", "OR", "THE", "UNSPECIFIED", and "WITH") removed, and remaining words sorted alphabetically; for example, "Typhoid meningitis" becomes "MENINGITIDES TYPHOID"; treated as *potential synonyms*
   - Non-name match: no exact match between original names or normalized names; treated as *potential children, parents or siblings*

We manually examined each partition to identify patterns of semantic relationships between the I10 and I9 terms. For example, we did not automatically assume that I10 and I9 terms were synonymous just because their GEM mapping was "00000" if the names (or normalized names) did not match. Similarly, we did not assume that terms were *not* synonymous just because their GEM mapping was "10000" if their names (or normalized names) matched and they were not "other" terms.

These manual examinations of each partition led to development of general decision rules, or heuristics, to apply to each I10-I9 pair. Based on the decions, the following actions were taken:

synonym - I10 term code and name added to existing I9 RED concept

child - new term added as child of existing I9 RED concept (new leaf node in hierarchy)

sibling - new term added as new leaf node child of the parent(s) of the existing I9 RED concept

parent - new term added as parent of existing RED concept (inserted into hierarchy above I9 concept).

Once the general rule was established, we began to manually review these decisions to find exceptions. This process allowed us to add I10 terms to the RED in bulk fashion with subsequent review and, where appropriate, correction (split synonyms into new and old concepts, merge synonymous concepts, and reclassify new concepts).

Results

General RED Inclusion Rules for Partitions of GEM Mappings

Table 1 shows the size of each of the partitions of GEM mappings. Also shown are the initial mapping decisions for each partition, plus any decisions made based on manual review to date. Table 2 shows examples of each of these mappings.

For example, all "non-other" I10-I9 pairs for which the GEM "synonymous" (00000) mapping were initially considered to be synonyms. After manual review, we agreed with these mappings for all pairs in which the names (or normalized names) matched. Figure 3 shows an example of a synonym mapping in the RED. However, our manual review of the "non-name match" partition discovered 95 pairs that we considered to be non-synonymous, such as the I10 term Newborn (suspected to be) affected by maternal hypertensive disorders (*P00.0*) and the I9 term *Maternal hypertensive disorders affecting fetus or newborn* (*760.0*); we understand the I9 term as a diagnosis that is applied to the parent of the newborn and the I10 term as a diagnosis that is applied to the newborn itself.

An example of combination mappings can be found with the I10 term *Salmonella sepsis* (*A02.1*), which has a GEM map of 10111 to the I9 term *Salmonella septicemia* (*003.1*) and a GEM map of 10112 to the I9 term *Salmonella sepsis* (*995.91*). In this example, there is only one scenario (fourth flag) with two choices (fifth flag). In this case, we determined that the I10 term should be added as a new concept in the RED and placed in the hierarchy under the existing concepts for both of the I9 terms.

Table 1: Partitions of GEM Mappings Based on Flag Pattern, Combinatorial Patterns, “Non-Other” versus “Other” Terms, and Type of Name Matching. RED addition decisions are shown as counts of synonyms, siblings and children; letters next to decision counts refer to examples shown in Table 2.

GEM Flags	I10 to I9 Mapping	"Non-Other" vs. "Other" Terms	Name Matching	Mapping Decision		
				Synonyms	Siblings	Children
00000 (Synonymous)	One to One n=3528	Non-Other n=3241	Name Match n=1554	1554 a	0	0
			Normalized Match n=262	262 b	0	0
			Non-Name Match n=1425	1251 c	85 d	89 e
		Other n=287	Name Match n=100	0	100 f	0
			Normalized Match n=31	0	31 g	0
			Non-Name Match n=156	62 h	76 i	18 j
10000 (Approximate)	One to One n=1012	Non-Other N=812	Name Match n=41	41 k	0	0
			Normalized Match n=37	37 l	0	0
			Non-Name Match n=734	0	647 m	87 n
		Other n=200	Name Match n=19	0	19 o	0
			Normalized Match n=7	0	7 p	0
			Non-Name Match n=174	0	173 q	1 r
	One to Many n=358 with 844 mappings	Non-Other n=592 mappings	Name Match n=28	28 s	0	0
			Normalized Match n=29	29 t	0	0
			Non-Name Match n=535	0	458 u	77 v
		Other n=252 mappings	Name Match n=14	0	14 w	0
			Normalized Match n=4	0	4 x	0
			Non-Name Match n=234	0	95 y	139 z
	Many to One n=57,236 (4192 I9 terms)	Non-Other n=46,323	Name Match n=83	83 aa	0	0
			Normalized Match n=292	292 bb	0	0
			Non-Name Match n=45,948	0	0	45,948 cc
		Other n=10,913	Name Match n=0	0	0	0
			Normalized Match n=0	0	0	0
			Non-Name Match n=10,913	0	0	10,913 dd
	Many to Many n=3221 with 7550 mappings	Non-Other n=5260 mappings	Name Match n=9	9 ee	0	0
			Normalized Match n=13	13 ff	0	0
			Non-Name Match n=5238	0	5238 gg	0
		Other n=2290 mappings	Name Match n=13	0	13 hh	0
			Normalized Match n=17	0	17 ii	0
			Non-Name Match n=2260	0	2260 jj	0
101xx (n=3808)	One to Many n=3808 with 7825 mappings	Non-Other n=5880 mallings	Name Match n=1	1 kk	0	0
			Normalized Match n=0	0	0	0
			Non-Name Match n=5879	0	0	5879 ll
		Other n=1945 mappings	Name Match n=0	0	0	0
			Normalized Match n=0	0	0	0
			Non-Name Match n=1945	0	0	1945 mm
None (n=669)	No mappings	No mappings	No mappings	0	669 nn	0

Table 2: Examples of Addition Decisions from Table 1

	I10 Code	I10 Name	I10 Flag	I9 Code	I9 Name	I9 Flag	Addition
a	A00.9	Cholera, unspecified	00000	001.9	Cholera, unspecified	00000	Synonym
b	G50.	Atypical facial pain	00000	350.2	Atypical face pain	00000	Synonym
c	A05.0	Foodborne staphylococcal intoxication	00000	005.0	Staphylococcal food poisoning	00000	Synonym
d	C91.10	Chronic lymphocytic leukemia of B-cell type not having achieved remission	00000	204.10	Chronic lymphoid leukemia, without mention of having achieved remission	00000	Sibling
d	P00.0	Newborn (suspected to be) affected by maternal hypertensive disorders	00000	760.0	Maternal hypertensive disorders affecting fetus or newborn	00000	Sibling
e	M72.0	Palmar fascial fibromatosis [Dupuytren]	00000	728.6	Contracture of palmar fascia	00000	Child
e	A31.0	Pulmonary mycobacterial infection	00000	031.0	Pulmonary diseases due to other mycobacteria	00000	Child
f	A02.8	Other specified salmonella infections	00000	003.8	Other specified salmonella infections	00000	Sibling
g	D56.8	Other thalassemias	00000	282.49	Other thalassemia	00000	Sibling
h	D49.7	Neoplasm of unspecified behavior of endocrine glands and other parts of nervous system	00000	239.7	Neoplasm of unspecified nature of endocrine glands and other parts of nervous system	00000	Synonym
i	A03.8	Other shigellosis	00000	004.8	Other specified shigella infections	00000	Sibling
j	A35	Other tetanus	00000	037	Tetanus	00000	Child
j	D26.0	Other benign neoplasm of cervix uteri	00000	219.0	Benign neoplasm of cervix uteri	00000	Child
k	A52.11	Tabes dorsalis	10000	094.0	Tabes dorsalis	10000	Synonym
l	A09	Infectious gastroenteritis and colitis, unspecified	10000	009.0	Infectious colitis, enteritis, and gastroenteritis	10000	Synonym
m	A08.11	Acute gastroenteropathy due to Norwalk agent	10000	008.63	Enteritis due to norwalk virus	10000	Sibling
n	A74.81	Chlamydial peritonitis	10000	079.88	Other specified chlamydial infection	[none provided]	Child
o	B10.89	Other human herpesvirus infection	10000	058.89	Other human herpesvirus infection	10000	Sibling
p	A18.13	Tuberculosis of other urinary organs	10000	016.30	Tuberculosis of other urinary organs, unspecified	10000	Sibling
q	A18.03	Tuberculosis of other bones	10000	015.60	Tuberculosis of mastoid, unspecified	10000	Sibling
r	J61	Pneumoconiosis due to asbestos and other mineral fibers	10000	501	Asbestosis	10000	Parent
s	B56.9	African trypanosomiasis, unspecified	10000	086.5	African trypanosomiasis, unspecified	10000	Synonym
t	A20.2	Pneumonic plague	10000	020.5	Pneumonic plague, unspecified	10000	Synonym
u	A15.5	Tuberculosis of larynx, trachea and bronchus	10000	011.30	Tuberculosis of bronchus, unspecified	10000	Sibling
v	A54.83	Gonococcal heart infection	10000	098.85	Other gonococcal heart disease	10000	Child
w	A07.8	Other specified protozoal intestinal diseases	10000	007.8	Other specified protozoal intestinal diseases	10000	Sibling
x	A08.39	Other viral enteritis	10000	008.69	Enteritis due to other viral enteritis	10000	Sibling
y	A07.8	Other specified protozoal intestinal diseases	10000	007.8	Other specified protozoal intestinal diseases	10000	Sibling
z	A54.83	Gonococcal heart infection	10000	098.85	Other gonococcal heart disease	10000	Child

Table 2 (continued): Examples of Addition Decisions from Table 1

	I10 Code	I10 Name	I10 Flag	I9 Code	I9 Name	I9 Flag	Addition
aa	A18.50	Tuberculosis of eye, unspecified	10000	017.30	Tuberculosis of eye, unspecified	10000	Synonym
bb	A01.00	Typhoid fever, unspecified	10000	002.0	Typhoid fever	10000	Synonym
cc	A01.01	Typhoid meningitis	10000	002.0	Typhoid fever		Child
dd	A01.09	Typhoid fever with other complications	10000	002.0	Typhoid fever		Child
ee	E46	Unspecified protein-calorie malnutrition	10000	263.9	Unspecified protein-calorie malnutrition	10000	Synonym
ff	A51.5	Early syphilis, latent	10000	092.9	Early syphilis, latent, unspecified	10000	Synonym
gg	A17.83	Tuberculous neuritis	10000	013.62	Tuberculous encephalitis or myelitis, bacteriological or histological examination unknown (at present)	[none provided]	Sibling
hh	A48.8	Other specified bacterial diseases	10000	040.89	Other specified bacterial diseases	10000	Sibling
ii	A54.39	Other gonococcal eye infection	10000	098.49	Other gonococcal infection of eye	10000	Sibling
jj	A18.89	Tuberculosis of other sites	10000	017.80	Tuberculosis of esophagus, unspecified	10000	Sibling
kk	A22.1	Pulmonary anthrax	10111	022.1	Pulmonary anthrax	10000	Synonym
ll	A02.1	Salmonella sepsis	10111	003.1	Salmonella septicemia	10000	Child
ll	A02.1	Salmonella sepsis	10112	995.91	Sepsis	[none provided]	Child
mm	A37.81	Whooping cough due to other Bordetella species with pneumonia	10111	033.8	Whooping cough due to other specified organism	[none provided]	Child
mm	A37.81	Whooping cough due to other Bordetella species with pneumonia	10112	484.3	Pneumonia in whooping cough	[none provided]	Child
nn	R40.2131	Coma scale, eyes open, to sound, in the field [EMT or ambulance]					

Repeating Patterns of Many-to-One Mappings

By far, the largest partition is the many-to-one approximate (10000) mappings, with 57,236 I10 codes mapping to 4192 I9 codes. Our initial assessment was that most of these mappings represent a refinement of an I9 term into many I10 terms. However, when we reviewed specific sets of mappings, we noticed recurring patterns. For example, 996 of the I9 codes map to exactly two I10 terms. In 261 cases, one of these is an "other" term and the other is not. More broadly, the mappings to 768 I9 terms (18.3%) include exactly one "non-other" I10 term. Examination of these sets shows that, in general, if the I9 term is represented as "X", one I10 term can be represented as "Other specified X" and the remaining I10 terms are more specific forms of X. If this is true for all such "one 'other' and many 'non-other'" patterns, we can with confidence add all of these I10 terms as children of their

Typhoid Fever **Code**: C4198842 **Preferred Term**: Typhoid Fever

Full_Syn	Typhoid fever
Syn_Source	SoftMed-CC
Syn_Type_Term:	PT
Syn_Source_Local_Code	002.0
Syn_Source_Domain	Diagnosis_ICD-9-CM
Full_Syn	Typhoid fever
Syn_Source	CRIMSON NIAID
Syn_Type_Term:	PT
Syn_Source_Local_Code	2068
Syn_Source_Domain	Problem
Full_Syn	Typhoid fever, unspecified
Syn_Source	SoftMed-CC
Syn_Type_Term:	PT
Syn_Source_Local_Code	A01.00
Syn_Source_Domain	Diagnosis_ICD-10-CM

Figure 3: Example of a RED concept that maps to an I9 term and an I10 – that is, the terms are considered synonymous. This is also a term used by an NIH clinical trials data management system (CRIMSON).

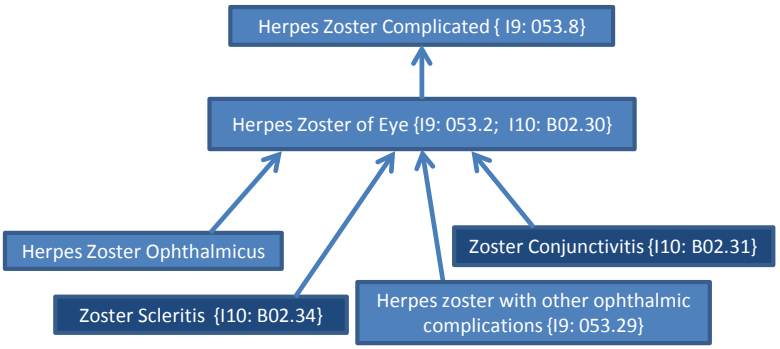


Figure 4: Example of RED hierarchy after inclusion of I10 terms. Light boxes represent concepts that were in the RED prior to inclusion of I10 terms; dark boxes represent newly added concepts. I9 and I10 codes are shown in {curly braces}.

corresponding I9 term. A similar rule may apply for the 2498 (59.6%) I9 terms that have no "other" I10 terms in their set of maps.

Class-Based Retrieval of BTRIS Data across ICD Versions

The ultimate purpose of all this work is to support the storage and retrieval of BTRIS data in as accurate and seamless a manner as possible. A perfect ontology is probably impossible, given all the subtle nuances of the differences among and between I9 and I10 codes. Recoding all the I9 data with I10 codes is not practical and would likely lead to inaccuracies. Requiring users seeking to retrieve data to select a code from each terminology version might be accurate but would be confusing and awkward. The I10 update we are applying to the RED provides a practical solution. Figure 4 shows an example of the new hierarchy of RED concepts that include mappings to I9 and I10 terms. BTRIS supports class-based queries so that a user could, for example, query for the RED concept Herpes Zoster Complicated and retrieve all data coded with the I9 codes 053.8, 053.2 and 053.29 and with the I10 codes B02.30, B02.31 and B02.34.

Discussion

A clinical data repository that combines I9 and I10 data without a coordinated terminology solution has two options: either re-code all the old I9 data with I10 or force users to query using both I9 and I10 terms. Our use of a single repository terminology, the RED, provides significant benefits to our users: they use a single terminology to perform queries that includes not only I9 and I10 terms but several other local diagnosis terminologies. The RED also provides the benefit of multiple hierarchies, so that users are not restricted to either of the single hierarchies of in I9 or I10, and they can still query for specific I9 or I10 terms if desired. The ontology approach – regardless of how the maintenance is accomplished - reduces the need to include multiple terms in queries. While this probably improves query execution somewhat, the real saving is a reduction in the BTRIS user's effort to construct a comprehensive query.

RED maintenance requires significant human effort. The manual integration of a new terminology of I10's magnitude would require many months of work, much of it mind-numbing and prone to error. The GEM mappings have been of tremendous benefit for providing a first approximation of synonymy and classification of I10 terms. The result is that I10 terms can be added quickly and be made useful immediately for both storing and retrieving patient data. Where revisions are required to merge redundant terms, split ambiguous ones, or reorganize the hierarchy, the amount of effort needed is no more that would be needed if we were to try to get everything "right" prior to loading I10 into the RED. Meanwhile, our use of GEM-based heuristics allows us to have a working, integrated repository while proceeding with "tuning" the RED. We now have the luxury of prioritizing our reviews based on the I10 terms as they actually start appearing in patient records.

In addition to continued work at reviewing and revising the I10 additions to the RED, we will continue to study the patterns of the many-to-one "similar" GEM mappings to determine if they can be used to improve the relationships between I9 and I10 terms. We will also need to make changes as updates to I10 occur; one of these has taken place already.

While the GEMs are obviously limited to use with I9 and I10, their existence raises the possibility of similar approaches to other mappings, such as between various drug terminologies or between I10 and the Systematized Nomenclature of Medicine (SNOMED-CT). The actual methods used and level of effort needed to develop the GEMs is not published, but it is interesting to consider whether similar tools could be developed for other term mappings.

When our ontology-based method for coping with I9 changes was originally published in 1996, Tuttle and Nelson praised it in an accompanying editorial but lamented the need for having to carry out "reverse engineering". They went on to say:

> Such methods should be viewed as necessary short-term expedients only, and all parties concerned should work toward an incremental plan by which the intent of changes to controlled health-care vocabularies can be made both explicit and machine processible. Only then can the comparability of patient descriptions be sustained.[12]

We, too, await such changes.[13] In the meantime, while the GEMs are not machine processable, they are at least machine readable and appear to be based on sound terminologic principles. Furthermore the ontologic approach of mapping terms to concepts and then coding data based on those concepts has proven to be a valuable precedent that continues to provide value in dealing with local and standard terminologies alike.

Conclusion

The practice of using an ontology that integrates disparate clinical terminologies has proven to be a powerful method for meeting the challenges of adapting a clinical data repository to include I10 data while maintaining the value of legacy I9 data. Our approach is able to take advantage of the GEM mappings from CMS to provide a rapid solution that can evolve gracefully.

Acknowledgments

Drs. Cimino and Remennik are supported in part by research funds from the NIH Clinical Center and the National Library of Medicine. The opinions expressed in this article are authors' own and do not reflect the view of the National Institutes of Health, or the Department of Health and Human Services

References

1. United States National Center for Health Statistics. *The International Classification of Diseases, 9th Revision, with Clinical Modifications*. Washington, DC; 1980.
2. Yu AC, Cimino JJ. A comparison of two methods for retrieving ICD-9-CM data: the effect of using an ontology-based method for handling terminology changes. J Biomed Inform. 2011 Apr;44(2):289-98.
3. Cimino JJ. Formal descriptions and adaptive mechanisms for changes in controlled medical vocabularies. Methods Inf Med. 1996 Sep;35(3):202-10.
4. Boyd AD1, Li JJ, Burton MD, Jonen M, Gardeux V, Achour I, Luo RQ, Zenku I, Bahroos N, Brown SB, Vanden Hoek T, Lussier YA. The discriminatory cost of ICD-10-CM transition between clinical specialties: metrics, case study, and mitigating tools. J Am Med Inform Assoc. 2013 Jul-Aug;20(4):708-17.
5. Cimino JJ, Ayres EJ, Remennik L, Rath S, Freedman R, Beri A, Chen Y, Huser V. The National Institutes of Health's Biomedical Translational Research Information System (BTRIS): Design, contents, functionality and experience to date. J Biomed Inform. 2013 Nov 19. pii: S1532-0464(13)00181-0. doi: 10.1016/j.jbi.2013.11.004.
6. Cimino JJ. From data to knowledge through concept-oriented terminologies: experience with the Medical Entities Dictionary. J Am Med Inform Assoc. 2000 May-Jun;7(3):288-97.
7. Ross-Davis SV. Preparing for ICD-10-CM/PCS: one payer's experience with general equivalence mappings (GEMs). Perspect Health Inf Manag. 2012;9:1e.
8. Centers for Medicare and Medicaid Services. Diagnosis Code Set General Equivalence Mappings: ICD-10-CM to ICD-9-CM and ICD-9-CM to ICD-10-CM 2009 Version: Documentation and User's Guide. Available at https://www.cms.gov/ICD10/11b1_2011_ICD10CM_and_GEMs.asp.
9. Gu H, Perl Y, Elhanan G, Min H, Zhang L, Peng Y. Auditing concept categorizations in the UMLS. Artif Intell Med. 2004 May;31(1):29-44.
10. Center for Medicare and Medicaid Services. *2014 General Equivalence Mappings (GEMs) – Diagnosis Codes and Guide*. https://www.cms.gov/Medicare/Coding/ICD10/Downloads/DiagnosisGEMs-2014.zip
11. Cimino JJ. Use of the Unified Medical Language System in patient care at the Columbia-Presbyterian Medical Center. Methods Inf Med. 1995 Mar;34(1-2):158-64.
12. Tuttle MS, Nelson SJ. A poor precedent. Methods Inf Med. 1996 Sep;35(3):211-7.
13. Cimino JJ. An approach to coping with the annual changes in ICD9-CM. Methods Inf Med. 1996 Sep;35(3):220.

Clinical Workflow Observations to Identify Opportunities for Nurse, Physicians and Patients to Share a Patient-centered Plan of Care

Sarah A. Collins, RN, PhD[1,2,3]; Priscilla Gazarian RN, PhD[2], Diana Stade[2], Kelly McNally[2], Conny Morrison[2]; Kumiko Ohashi PhD, RN[2]; Lisa Lehmann MD, PhD[2,3]; Anuj Dalal, MD[2,3]; David W. Bates MD, MSc[1,2,3]; Patricia C. Dykes, RN, PhD[2,3]

[1]Partners Healthcare Systems, Wellesley, MA; [2]Brigham and Women's Hospital, Boston, MA; [3]Harvard Medical School, Boston, MA

Abstract

Patient- and Family-Centered Care (PFCC) is essential for high quality care in the critical and acute-specialty care hospital setting. Effective PFCC requires clinicians to form an integrated interprofessional team to collaboratively engage with the patient/family and contribute to a shared patient-centered plan of care. We conducted observations on a critical care and specialty unit to understand the plan of care activities and workflow documentation requirements for nurses and physicians to inform the development of a shared patient-centered plan of care to support patient engagement. We identified siloed plan of care documentation, with workflow opportunities to converge the nurses plan of care with the physician planned To-do lists and quality and safety checklists. Integration of nurses and physicians plan of care activities into a shared plan of care is a feasible and valuable step toward interprofessional teams that effectively engage patients in plan of care activities.

Introduction

Patient- and Family-Centered Care (PFCC) is associated with better clinical outcomes in the critical care setting, improved decision-making for patient-centered goals of care, increased family satisfaction, decreased length of stay, and increased care-giving ability by family members post-discharge.[1–5] Effective PFCC in the acute and critical care setting requires the coordinated assessment and active consideration of the psychosocial needs and preferences of patients and families by the interprofessional care team through continuous engagement and involvement in discussions and decision-making.[2,6] Critical care patients are often too ill to advocate for themselves, making engagement of patients' families in decision-making discussions especially important; yet, few specific strategies to operationalize and sustain family-centered care in adult patients have been published.[1] We believe that a valuable and novel strategy is to provide hospitalized patients and families with a means of viewing their plan and providing feedback about their plan to providers in real-time.

Unfortunately, within the high technology and high acuity critical and specialty care environment, patient and family needs related to care planning and transitions of care, such as increased family care-giving or end-of-life decisions, are often overlooked.[2] Without appropriate support and information, family members experience undue burden associated with the increased anxiety, ineffective coping, spiritual distress, and impaired decision-making ability related to their family member's critical illness and end of life care. Likewise, patients in acute and critical care, when well enough to engage, seek information about their care and care team.[7,8] Engagement of patients and families occurs through the process of shared decision-making, which is a method to support patients and families with healthcare provider expertise while incorporating patient's preferences to reach patient-centered decisions for goals of care.[9] A multidisciplinary effort in 2007, spearheaded by the American College of Critical Care Medicine (ACCM), developed the clinical practice guidelines for support of the family in the patient-centered intensive care unit.[2] This guideline, consistent with other literature, specifically recommends: shared decision-making, early and repeated discussions with the patient and family at the first sign of ineffective coping, improved consistency in communication from the interprofessional care team, and staff education for PFCC.[2,10–12] Nurses and physicians, as the clinicians that interact most frequently and consistently with critical care patients and their families, are in an ideal position to engage patient/family members in shared plan of care discussions. However, first, nurses and physicians must share their plans of care with each other to establish a common plan for effective exchange and engagement with patients and their families. Critical and specialty care (e.g., oncology) are a highly collaborative

environments where clinicians seek out knowledge from other clinical professions and specialties, but there is a lack of tools for interprofessional exchange of plans of care. Failure for the care team to effectively communicate is associated with patient/family stress, nurse stress, patient safety errors, inefficiencies, and excessive lengths of stay.[1-5]

Ensuring that family members receive consistent and timely information from clinicians on the interprofessional care team is one of the greatest evidence-based practice (EBP) gaps within the model of family-centered care.[13] Poor team communication is associated with mismatched patient care goals.[14–16] For effective shared decision making conversations to occur between the patient/family dyad and the interprofessional care team, nurses and physicians must first establish common ground.[2] In prior work, we found that clinicians, including nurses and physicians, sought to understand care interventions delivered by other professions.[17] Active alignment of profession-specific care priorities with patient-centered priorities is a critical part of integrated care and the degree of knowledge sharing between professions increases as care becomes more patient-centered.[18,19] Currently, as a healthcare system we do a poor job at asking patients to provide their feedback with regard to their plan of care, particularly their problems, preferences, goals, schedule, and evaluation of their status. There are few use cases and even fewer evaluation studies in the literature of patient-centered knowledge sharing tools among patients, nurses, and physicians.[20,21] Sustainable and feasible patient-centered team behavioral interventions remain a challenge, but recent work points to active knowledge sharing as a critical component in effective models of patient-centered care.[19,22] This study is part of a large multi-year and multi-aim project that will provide hospitalized patients and their families with an iPad-based patient-centered toolkit (PCTK) to engage with their health data and plan of care as well as access information resources and communicate with their care team. The development of an electronic patient-centered toolkit (PCTK) for sharing the plan of care presents a ripe opportunity to: 1) understand the processes and plan of care activities that exist for patients, nurses, and physicians in critical and specialty care, 2) define opportunities for patient-centered engagement and alignment of plan of care activities, 3) define requirements for patient-centered plans that seize opportunities for clinician and patient engagement and avoids increasing clinician documentation requirements by considering existing processes and activities, and 4) provide patients with a novel mechanism to communicate real-time feedback and participate in the development of their plan of care.

The aims of this study are to: 1) Describe current clinical plan of care activities and workflow, 2) Identify opportunities to integrate the care teams' plan of care activities and workflow for care team engagement with patients and families, and 3) define workflow requirements for a plan of care that is integrated across the care team and shared with patients. This shared patient centered plan of care was deployed on the study units in May 2014.

Methods

This study took place at a large academic medical center in the Northeastern United States. We conducted observations on a Medical Intensive Care Unit (MICU) and Oncology Acute Care Unit (Oncology Unit) to understand the distinct and shared workflows of nurses and physicians and opportunities to engage with patients in development of the shared patient centered plan of care. The focus of this study was on nurses' and physicians' engagement with plan of care activities for the purpose of sharing an integrated team-based plan of care with the patient and family, to form a patient-centered plan of care. The study was reviewed and approved by the Partners HealthCare Human Research Committee.

Seven observations were conducted, 4 observations on the MICU and 3 observations on the Oncology Unit. Observations lasted between 2 and 4 hours. Each observation was conducted by 2-5 study investigators consisting of 2 nurse researchers, 1 physician, and 3 research assistants. The participants observed included nurses, residents and physicians assistants (PAs), attending physicians, fellows, charge nurses, staff nurses, pharmacists, patients and families. PAs managed patients in the Oncology Unit, but not in the MICU.

Study investigators took field notes during the observations; field notes were specifically targeted at activities and communication related to plan of care. Observational findings were validated using member checks during individual and group semi-structured interviews with clinicians. These interviews were conducted with 1-7 clinicians. During the interviews, clinicians were presented with 1) observational findings describing the current clinical plan of care activities and workflow, 2) Visio diagrams to confirm opportunities to integrate PCTK into plan of care workflow for patient and clinician engagement, and 3) user interface prototypes (paper-based screen shot mock-ups) to confirm workflow and documentation requirements for a shared patient centered plan of care. After each observation and interview session, the investigators met for peer debriefings to review field notes, reflect on

past observations, iteratively identify emergent themes, and collectively add to, form consensus on, and refine requirements.

RE-AIM Conceptual Framework

The RE-AIM (Reach, Effectiveness, Adoption, Implementation, and Maintenance) framework was used to understand existing processes and workflow related to plan of care activities, identify opportunities for patient-centered engagement, and avoid potential unintended consequences and poor adoption.[23] RE-AIM was developed specifically to direct the design and conduct of rigorous studies of real world implementations of efficacious interventions to facilitate their translation of research into practice and has been specifically applied to informatics implementations.[23] We used the RE-AIM framework to formulate the following targeted questions to consider during our observations and interview data collection:

- Reach: What are the documentation and workflow opportunities for nurses and physicians on the MICU and Oncology Units to engage with patients in a shared patient-centered plan of care?
- Effectiveness: What are the required plan of care activities and documentation for nurses and physicians?
- Adoption: What are the current clinical workflow processes related to plan of care that nurses and physicians engage in and must be considered for user acceptance and workflow fit?
- Implementation: What are the technical requirements for integrating nurses and physicians plan of care activities?
- Maintenance: Is the intervention sustainable overtime? [*Will be evaluated after 'go-live' implementation*].

Results

Care Unit Process Categories for Plan of Care Activity Workflow and Opportunities for Engagement

We found that plan of care activities were distributed among a number of clinical care unit process categories. To facilitate identification of opportunities to engage nurses and physicians with patients in plan of care activities, we used the clinical care unit process categories that emerged from the data to categorize the observational data. We identified six high-level clinical care unit process categories: Intake/Admission, 24 hours post-admission, Daily Rounds, Family Meetings, Discharge, Handoff/End-of-Shift Documentation.

Patient admission to a critical or specialty care unit is a busy time during which a patient is stabilized (if needed), assessed, and evaluated to determine their immediate needs/goal of care and a plan that meets those needs. Interview data confirmed that the plan of care may not be established immediately when a patient is admitted to the unit, but is established within the first 24 hours of admission. The first 24 hours after a patient is admitted was a critical opportunity for holding an initial plan of care discussion with patients and family so that the patient's goals of care could be established. The nurses interviewed conveyed that during an initial family meeting the patient and family are typically focused on critical medical issues, patient acuity, and seek information regarding the patient's treatments and logistics of the hospital stay.

After the initial 24 hours, Daily Rounds was the critical process used to evaluate plan of care progression over the past 24 hours and to revise the plan for the next 24 hours. Rounds were typically held outside patients' rooms and patients and families were not actively engaged in the rounds discussion on the two study units. Rather this time was afforded to the interprofessional care team to facilitate common ground and a shared understanding of the plan of care amongst the team members. Some clinicians noted that this workflow excludes the patient and family in a manner that was inconsistent with patient-centered care. Benefits cited by the clinicians of reserving a time for only clinicians to discuss the patient included the ability for the care team to efficiently and openly confirm, dispute, and ask questions about the patient's condition, diagnoses, therapeutic options, and share individual insights and communications that each clinician may have had with the patient and family or other clinicians.

Daily rounds on the MICU and the Oncology Unit have overlapping and distinct characteristics. Rounds on both units follow a consistent structure, which validated our prior work.[15,16,24] We observed and confirmed in interviews that the objective of rounds was to establish and discuss medically focused problems and goals for the next 24 hours. The documentation output from rounds included progress notes and to-do lists which were distributed and documented individually among care team members. In the MICU, the nurse was responsible for summarizing the team goals that were established during rounds and documenting those daily goals on a sheet that was posted outside the patient's door. Overall, the discussion was well-structured, but allowed for dynamic engagement of outstanding

issues highlighted by any team member. During rounds, the team referred to checklists to ensure certain items were discussed, including the patient's code status. However, discussions of other patient preferences beyond code status was rarely observed, but was validated as important when brought up by team members. When patient preferences *were* discussed, the attending physician and nurse were most frequently observed to initiate those discussions. In the MICU, there was a focus on daily checklists, sometimes referred to as "bundles," to ensure the patient's plan was aligned with known safety goals, evidence-based care, and quality metrics. Examples of the checklists ("bundles") include: the Ventilator Management Bundle, Catheter-associated Urinary Tract Infection (CAUTI) bundle, and the Central Line Bundle. The Quality and Safety Checklists used at MICU rounds also related to: restraints, nutrition, blood glucose checking, implementation of early physical therapy, deep vein thrombosis prevention, gastrointestinal ulcer prevention, communication with family, and code status.

The MICU nurses' summarization of the team goals was viewed as valuable for coordinated team-based care, yet the daily goals sheet that reflected this summarization was not part of the patient's formal medical chart and was discarded every 24 hours. For this reason, we conclude that there was no formal and permanent documentation that served as a central source of decisions or plan for reference and evaluation over time. When interviewed about the possibility of making the daily goals sheet computer-based for sharing among the interprofessional team and patient during and after rounds the nurses disagreed with the proposed workflow. The nurses stated that they use this daily goals sheet to quickly document the high level outcomes of rounds and quickly note tasks and To-Do's to complete that day. The nurses were not opposed to sharing the daily goals sheet with the team, as it was already available for others to view (posted outside the patient's room). Rather, the nurses were hesitant to make it computer-based; they referred to it as their "scut" sheet and felt that they needed a piece of paper to quickly take notes about the team's goals for the patient, tasks and To-Do's to refer to throughout the day. The nurses stated that all of the information on this daily goals sheet is formally documented after rounds, typically by the night shift nurse, in the nurses' plan of care. While the daily goals sheet is not recognized as a formal part of the patient's chart, the nurses' plan of care is a formal part of the patient's chart and may be a useful mechanism of communicating information that is meaningful to other members of the care team and patient.

Rounds on the oncology unit differed from the MICU in several ways. Only one team rounded in the MICU and the nurse cared for one to two patients per shift, facilitating the nurse's ability to participate as an integral part of the team during rounds. On the oncology units, several teams rounded simultaneously making it more difficult for the nurse to be present for each team's rounds. As a result, the PA or intern conducted a pre-rounding check with the nurse to identify important information to be shared on rounds. The oncology unit did not share a common checklist or "bundles," however, on the oncology unit many patients are admitted for specific treatment protocols that have a relatively well defined course, and our observations and interviews indicated that the team typically had a common understanding of the patient's anticipated progression. Informal handwritten notes were taken by PAs and interns during rounds, which were then entered into the online patient record and printed to be physically stored in the patient chart after rounds.

Family meetings are formal meetings scheduled with the family to discuss the patient's condition, reach common understanding about the goals of care, and agree on a plan. These meetings are held in the clinical unit's conference room and at a minimum include the attending physician or fellow, a nurse, the patient or health care agent, and family. The meetings are typically highly emotional, particularly meetings that are held "as needed" due to a patient's worsening condition. An attempt is made to hold a family meeting within the first 24 hours of admission and every couple of days after, although this is not always practical or achieved. Documentation of what was discussed is completed on a structured Patient/Family Communication Note that is part of the chart and may be completed by the social worker or attending physician. Our observations and interviews indicated that the Patient/Family Communication Note is often not completed and instead the attending physician may write a physician progress/daily note. A point of concern noted by a charge nurse is that these notes include variable and often insufficient levels of detail. For example, the charge nurse pointed out a family meeting note that stated: "DNR/DNI but will continue aggressive therapy" and explained that this plan is not clear because it does not specify which type of therapy, for what purpose, and the plan and schedule to reevaluate the patient. In addition to formal scheduled family meetings, informal meetings and updates are conducted with the family on a daily basis. These updates may be shorter, less formal discussions between the patient and/or family and the physician and/or nurse and they may occur over the phone, in the patient's room, or in the hallway of the clinical unit. The oncology unit has no specified plan for family meetings, using them only on an ad hoc basis. The PA teams return to the units in the afternoon to follow up with patients and families regarding changes or progress in the plan of care. Additionally,

as most acute care oncology patients are able to engage in their care, informal meetings with other providers (nurses, social workers, etc.) are regularly held at the bedside as needed.

Table 1. Care Unit Process Categories and Associated Opportunities for Engagement

Care Unit Process Categories	Ripe Opportunities for Clinician-Patient/Family Engagement with Shared Patient Centered Plan of Care
Admission	• *Introduction* of a shared patient-centered plan of care approach to patient/family

24hrs post admission	• *Teaching* by nurse with patient/family related to initial plan, medications, and procedures
	• *Identification* of
	  o Problems, preferences, and goals by patient/family
	  o Problems, goals, initial plan by care team

Daily Rounds	• *Data display* to care team during rounds captured from patient
	  o Patient identified problems, preferences, and goals
	  o Home/out-patient medications/therapy regimen
	• *Data capture* from care team during rounds to push to patient
	  o Leverage existing workflow to summarize goals and plan at end of rounds
	  ▪ Output to Nursing Shared Plan of Care
	  ▪ Output to Physician Shared "To-do" Lists and Safety and Quality Checklists
	• *Promote* discussion of
	  o Patient preferences among care team
	  o "Big picture" to consider if goals of care are aligned with patient condition and preferences
	  o Shared plan of care with patient/family in follow-up discussion after rounds

Family Meetings	• *Formalize* patient preferences and goals of care in context of treatment option • *Clarify* patient preferences and patient/family degree of certainty in decision-making • *Confirm* shared understanding of "big picture" of patient's condition • *Align* problems and specific plan with patient's goals of care and preferences • *Seek* patient/family decision logic and "if/then" plan (e.g., if patient does not respond to therapy after X time then re-examine goals of care) • *Provide* families with shared plan of care and sufficient time to consider plan • *Communicate* outcome of family meeting to care team that was not present
Discharge	• *Implement* patient progression framework prior to day of discharge
Handoff	• *Communicate* shared patient centered plan of care to next shift

During daily rounds, patients are deemed "cleared for discharge" if their condition permits. Patients on the MICU that are cleared for discharge typically wait about 6 to 24 hours for a bed to be available on step-down/acute care unit or a long term care facility. On the Oncology Unit a Patient Progression Model with Interprofessional Huddles was recently piloted due to delays in discharge and to facilitate team-based coordination and communication. The Patient Progression Model is a clinical process model that engages the team in actively and collaboratively identifying the expected date of discharge, location of discharge and any barriers to discharge, such as medications that need prior approval for oncology patients or physical therapy referrals. The hospital had determined that patient discharge was often delayed due to insufficient coordination of logistical reasons and dependencies, not medical reasons, and the Patient Progression Model is an organizational effort aimed at increasing the transparency of discharge dependencies through Interprofessional Huddles – frequent and short team-based discussions that may leverage a care coordinator and other members of the care team to coordinate and manage dependencies earlier during the patient's hospital stay. The Patient Progression Model with Interprofessional Huddles are promising workflows to engage the care team in a shared patient-centered plan of care. Yet, at this time, as a newly implemented pilot program, our ability to conclude how these workflows will be adopted and can be leveraged is limited.

In both the MICU and Oncology units, nurses document a plan of care on paper at the end of each shift. This is a continuous plan of care for the patient's stay on the unit as it is shared among each nurse caring for the patient and includes multiples shifts and days on the same paper. The nurses' plan of care includes a nursing problem list, relevant assessment data, goals, and interventions. The residents in the MICU update the "To-Do" list on a whiteboard every day after rounds in physician break-room. In addition to the To-Do list, physicians complete a

computer-based semi-structured medical problem list using an internally developed clinical information system. The medical problem list is refined during rounds and documented on each physician's daily note and used to inform the To-Do List. The "To-Do" list is shared by all members of the medical team, and primarily edited by the residents. The To-Do list is focused on items such as medications or procedures to order, laboratory results to review, consults to request, and family meetings to schedule. In our data collection, we observed that handoff at the end of a shift was used to convey the plan, any issues that arose during the past shift related to the plan, and outstanding issues that still needed to be addressed in the plan. These updates are incorporated into the nursing plan of care documentation and the physician "To-Do" list documentation.

Requirements for Nurse, Physician, and Patient Engagement with a Shared Patient Centered Plan of Care

Our observational and interview data indicated two existing documentation workflows that are feasible for data capture into a shared patient centered plan of care: 1) Nursing Shared Plan of Care documentation and 2) Physician Shared "To-Do" Lists and Safety and Quality Checklists (see Table 1). Both of these documentation workflows were paper- or whiteboard-based. The documentation was already shared among each profession, which eased some anticipated sociotechnical barriers to the development of a shared Patient-Centered Plan of Care. For example, we did not need to introduce the need for nurses to negotiate who is responsible for documenting the nursing plan of care and when the nursing plan of care documented should be completed. Likewise, the residents' and interns' responsibility for the Shared "To-Do" Lists and Safety and Quality Checklists had also already been negotiated. Therefore, the development of a shared patient centered plan of care that pulled from the nursing plan of care and physician checklists and "To-Do" Lists was seen as feasible and useful to the clinicians.

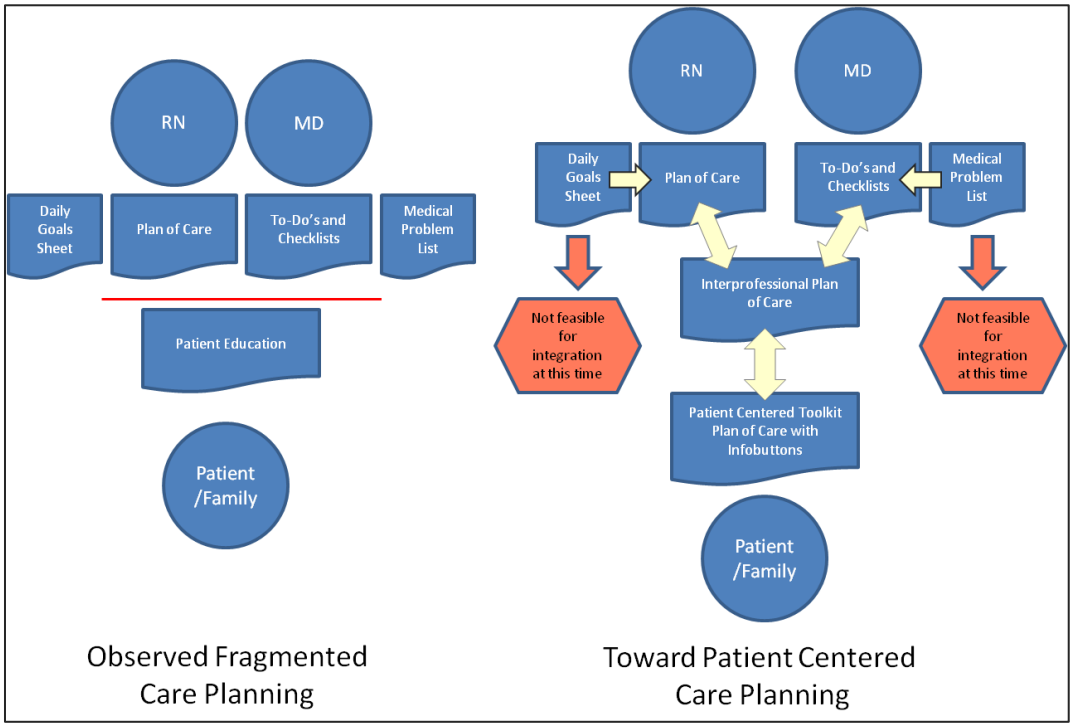


Figure 1. Toward Patient-Centered Care Planning Through Integration of Siloed Plan of Care Documenation

In the MICU, the residents and interns required the ability to document their "To-Do" List and Safety and Quality Checklists during rounds. The "To-Do" Lists include frequently documented tasks, such as the need to communicate with the surgical team or hold a family meeting. These frequently documented "To-Dos" could be structured and made available for selection, with the ability to add more detail or more "To-Dos" in free-text. The Safety and Quality Checklists required the ability to check off items daily during rounds and when appropriate provide a rationale for why an item was not completed. An analysis was required for checklist items that also existed on the nursing Plan of Care to determine the "source of truth" and level of sharing between each computer-based checklist. A core requirement for the To-Do list and Safety and Quality Checklists was the ability for the resident/interns to select "To-Do" and Checklist items to "push" to patient for viewing and that were Infobutton enabled.

During observations, Oncology clinicians identified rounds as a source of information exchange inefficiency, partially due to the informal process with which they recorded tasks. Thus, although a formal checklist and "To-Do" List was not part of the existing Oncology workflow; providers were willing to pilot a computer-based checklist during rounds that is being developed by the study team.

It is important to note that the physician/PA documentation related to the plan of care includes documentation of the medical problem list as well as the To-Do and Checklists. Based on clinical and technical stakeholder engagement, it was determined that it is currently not feasible to structure the medical problems entered by the physician/PA into the interprofessional and shared patient-centered plan of care due to hospital-wide clinical care dependencies on the existing applications used for problem list documentation. Modifying those existing applications was out of scope of this study. Hence, Figure 1 describes our initial model for working toward patient-centered care planning and conveys the documentation workflows that are feasible for integration in their current state. The Plan of Care and Daily Goals Sheet are included under "RN" in the figure to denote those are conducted by the nurse. The To-Do's and Checklists and Medical Problem List are included under the MD to indicate those are conducted by the Physician. While the Physician To-Do list is informed by the medical problem list and the Nurses Plan of Care is informed by the Daily Goals Sheet we acknowledge that the exclusion of the medical problem list and daily goals for feasibility reasons is not ideal. We envision an ideal future version of a shared patient-centered plan of care that fully integrates the medical problem list and daily goals documentation.

Table 2. Plan of Care Requirements based on Opportunities for Engagement and Current Documentation Artifacts

User	**Residents/Physicians' Assistant Shared Checklists and "To-Do" list**	**Nurses Shared Plan of Care**	**Patient/Family PRGS (Problems, Respect, Goals, Summary)**
User specific Requirements	• Documented during rounds • Structured commonly used "To-Do's" • Ability to free-text To-Dos • Structured checklists • Ability to electronically check off "To-Do" list • Ability to select To-Do's and checklist items to "push" to patient • Future state: integration of medical problem list	• Problem based • Structured problems • Ability to free text problems • Ability to select problems to "push" to patient • Identification of problem-based goals and interventions and status selection for each • Share common checklists with Residents/Physicians' Assistant checklists • Future state: integration of daily goals sheet	• Data Entry of PRGS o Problems o Respect for Preferences o Goals o Summary and Status • Schedule/planned procedures • Home/out-patient medications/therapy regimen
Shared Requirements*	Ability to view and edit Shared Patient Centered Plan of Care • Data pushed from o Nurses Shared Plan of Care o Residents/Physicians' Assistant Shared Checklists and "To-Do" list • Data pulled from Patient PRGS (Problems, Respect, Goals, Summary) • Ability to print • Ability to version • Ability to compare days		
*Shared Requirements pertain to users outlined in table as well as additional care team members such as Attending Physician and Charge Nurse.			

The nurses required the ability to view the plan of care from previous days and carry-over the problem list from shift-to-shift and revise as needed. The nurses shared plan of care documentation included a problem list and a separate system-based structure for the nurse to identifying goals and interventions that addressed each body system. The problem list used nursing terminology consistent with the Clinical Care Classification system, which facilitated the development of structured problems for selection with the ability to include free-text problems. There were a lack of linkages between the problems and daily goals/interventions which posed a challenge to pushing information

from the nursing plan of care to the patient. The nursing plan of care was restructured as a problem-based plan of care to facilitate alignment with the Patient-Centered Shared Plan of Care. Additional requirements included the ability to select the status of goals (Improved, Stabilized, or Deteriorated). Both the residents' To-Do list and Checklists and the nurses' Plan of Care required the ability to save new versions without overwriting old versions, compare past days to current day on the same screen, and print.

Our data indicated that a patient-facing care plan requires the ability for patients and families to identify problems, preferences for care, goals, and provide a summary of feedback about the status of care toward those goals. We coined the term for these requirements as the Patient's "PRGS: Problems, Respect for Preferences, Goals, and Summary/Status". As part of their plan, patients and family want to know the logistics for any planned procedures, such as the need to go to a different location in the hospital for a procedure, particularly in the context of coordinating family visits. This finding validated prior work that found patients wanted to view their daily schedule.[7] Finally, to ensure continuity of their plan, including information about their home and out-patient medications was important to patients. Including their out-patient therapy regimen was particularly important for oncology patients.

We also focused on the role of the attending and fellow physician and the charge nurse in the MICU and Oncology Unit. These roles were observed to actively drive the plan of care forward by focusing care team discussions and activities on the "big picture" (what are the care goals and are they aligned with current interventions and the patients' condition), long term goal clarification, care coordination, patient preferences, and the need for and outcomes of family meetings.

Discussion

This study confirms prior work that nurses and physicians document in silos (paper and electronic), despite engaging in shared and formal conversations at rounds.[15,16,25,26] While it is not possible to conclude that the clinicians interviewed are a representative sample, confirmation of findings with prior work lends confidence that the sample and data obtained are similar to other groups of clinicians in the acute and critical care setting. This study extends beyond describing the silos to a unique workflow analysis of how those silos can be integrated for shared planning. However, with critical clinician input we identified existing plan of care activity and workflow processes that provide feasible opportunities for nurses and physicians to contribute to a shared patient-centered plan of care without significant burden. In this study the distinct differences in documentation structure and concepts completed by nurses and physicians serve as a facilitator to integrate each professions' plan of care documentation into a unified view. The nurses' problem list and goal identification using standard nursing terminology is seen as a patient-friendly method of communicating important risks, clinical states, and conditions to the patient/family. The resident/PA "To-Do's" provide a mechanism of communicating concrete daily action items. The safety and quality checklists could leverage Infobuttons to educate patients about safety and quality care activities that they should be aware of and could help monitor.[27]

The option to develop a shared patient centered plan of care that pulled from the nursing plan of care and physician checklists and To-Do lists is seen as ideal because it does not require redundant documentation by the clinicians and fit with their current responsibilities for documenting plan of care activities. Converting the nursing shared Plan of Care documentation and physician shared "To-Do" Lists and Safety and Quality Checklists to a computer-based version was a requirement, but is endorsed by the clinicians and technically feasible. Converting the clinicians' plan of care activities' documentation to a computer-based format would provide an infrastructure to push structured and semi-structured plan of care concepts from the nurses' and physicians' documentation to the patient and receive structured and semi-structured plan of care documentation from the patient. Future work will focus on the feasibility of integrating the physicians' medical problem list and the daily goals sheet used by individual nurses' to take quick and informal notes of tasks and To-dos.

In our observations, patients and families were not actively engaged in rounds discussions. We also rarely observed the discussion of patient preferences (except for code status), confirming our team's prior work demonstrating the low frequency in which patient preferences are discussed during cardiac critical care and acute care rounds.[24] These practices are opposed to the principals of patient-centered care. However, our interviews confirmed that clinicians aimed to provide patient-centered care, yet believed rounds should be reserved for the clinical care team to discuss the patient case without the patient and family present. This observation raises critical questions: Within a PFCC model and a care setting that requires collaborative and coordinated team-based care for critical care and specialty

patients is there a need for clinicians to preserve formal time for discussion without the patient and family present? Does the acuity of the patient, level of uncertainly of diagnosis and prognosis, and role of teaching rounds influence the need for clinicians to preserve formal time for discussion without the patient and family present? PFCC supports the full inclusion of the patient and family. Further research is needed to understand the clinical process requirements for the interprofessional care team to achieve a shared understanding of the patient's state, plan, and preference if the patient and family continue to be excluded from the rounds discussion. It should be evaluated if the clinical workflow process change to *include* the patient and family in all shared discussions prevents open and efficient communication and disclosure of clinical issues, questions, and clarifications among members of the clinical care team.

Importantly, the attending physician and charge nurse were observed to function in critical roles for engaging the care team and patient in the plan of care, though this may not be reflected in any documentation. The attending physician and charge nurse focused on the "big picture" by challenging the care team to clarify long-term goals of care and validate that those goals were aligned with current interventions, the patients' condition, and patient preferences. Development of a shared patient centered plan of care based solely on the individuals that document plan of care activities will miss care team leaders that may serve as "silent" stewards of a patient-centered plan. It is critical that these roles are able to view and contribute to the shared Patient-Centered Plan of Care.

Limitations

This study is limited to two clinical settings, a MICU and an Oncology Unit in one tertiary care academic institution. Our findings require further validation to determine if they are generalizable to nursing and physician engagement with patients for shared patient-centered plans of care in other critical care and acute care environments. These two units were targeted due to some overlap in patient populations between the MICU and Oncology Unit, specifically for oncology patients that require critical care. This allowed for an analysis of two units with important differences but some known similarities to enable an analysis of standard versus distinct approaches to integrating plans of care based on unique workflows and needs of each unit. The clinical settings in this study did not have a fully integrated EHR. However, to our knowledge most EHRs currently do not have interprofessional plans of care that integrate nurses' and physicians' plans. Therefore, we see utility in our findings for care settings with and without fully integrated EHRs. Finally, our data is based on observations and interviews targeted at nurses, physicians, and patients and did not target all types of clinicians that are members of the care team. Certainly, other types of clinicians provide critical plan of care activities. Further studies should: 1) focus on the role of other members of the care team in planning care in the hospital setting, 2) evaluate opportunities for engagement with patients and families for patient-centered plans of care, and 3) assess the impact of a unified plan of care on the quality and efficiency of communication and ultimately on patient outcomes and quality of patient care.

Conclusion

We found that there are common Plan of Care concepts that should be shared among interprofessional care team members, but data entry requirements appear to require tailoring to fit with workflow, data reuse, and minimize duplication of current documentation. Plan of care activities are distributed among a variety of care processes and formal methods of clinical communication. This distributed nature of planning in the care setting and the distribution of plan of care identification among different members of the care team (professional and patient) require the ability to document plan of care components separately while preserving the ability to unify that data for an integrated shared patient centered plan of care.

Acknowledgements

The authors would like to thank the clinicians who provide ongoing feedback to inform development of a shared patient-centered plan of care. This study is part of the Brigham and Women's Hospital PROSPECT Project, which is part of the Libretto Consortium supported by the Gordon and Betty Moore Foundation.

References

1. Mitchell M, Chaboyer W, Burmeister E, Foster M. Positive effects of a nursing intervention on family-centered care in adult critical care. *Am J Crit Care*. 2009;18(6):543–52; quiz 553. doi:10.4037/ajcc2009226.
2. Davidson JE, Powers K, Hedayat KM, et al. Clinical practice guidelines for support of the family in the patient-centered intensive care unit: American College of Critical Care Medicine Task Force 2004 –2005. *Crit Care Med*. 2007;35(2):605–622. doi:10.1097/01.CCM.0000254067.14607.EB.

3. Henneman E, Cardin S. Family-centered critical care: a practical approach to making it happen. *Crit Care Nurse*. 2002;22(6):12–9.
4. Price AM. Intensive care nurses ' experiences of assessing and dealing with patients ' psychological needs. *Nurs Crit Care*. 2004;9(3).
5. Curtis JR, White DB. Practical guidance for evidence-based ICU family conferences. *Chest*. 2008;134(4):835–43. doi:10.1378/chest.08-0235.
6. Maizes V, Rakel D, Niemiec C. Integrative medicine and patient-centered care. In: *Commissioned for the IOM Summit on Integrative Medicine and the Health of the Public*.Vol 5.; 2009:277–89. doi:10.1016/j.explore.2009.06.008.
7. Caligtan C a, Carroll DL, Hurley AC, Gersh-Zaremski R, Dykes PC. Bedside information technology to support patient-centered care. *IJMI*. 2012;81(7):442–51. doi:10.1016/j.ijmedinf.2011.12.005.
8. Dykes PC, Carroll DL, Hurley AC, et al. Building and testing a patient-centric electronic bedside communication center. *J Gerontol Nurs*. 2013;39(1):15–9. doi:10.3928/00989134-20121204-03.
9. Morrow CE, Reed VA, Eliassen MS, Imset I. Shared decision making: skill acquisition for year III medical students. *Fam Med*. 43(10):721–5.
10. Balaban RB. A physician's guide to talking about end-of-life care. *J Gen Intern Med*. 2000;15(3):195–200.
11. Abbott KH, Sago JG, Breen CM, Abernethy AP, Tulsky JA. Families looking back: one year after discussion of withdrawal or withholding of life-sustaining support. *Crit Care Med*. 2001;29(1):197–201.
12. Adams JA, Bailey DE, Anderson RA, Docherty SL. Nursing Roles and Strategies in End-of-Life Decision Making in Acute Care: A Systematic Review of the Literature. *Nurs Res Pract*. 2011;2011:527834. doi:10.1155/2011/527834.
13. Davidson JE. Family-centered care: meeting the needs of patients' families and helping families adapt to critical illness. *Crit Care Nurse*. 2009;29(3):28–34; quiz 35. doi:10.4037/ccn2009611.
14. Baggs J, Schmitt M. Nurses' and resident physicians' perceptions of the process of collaboration in an MICU. *Res Nurs Health*. 1997;20(1):71–80.
15. Collins S, Bakken S, Vawdrey DK, Coiera E, Currie L. Model development for EHR interdisciplinary information exchange of ICU common goals. *Int J Med Inform*. 2011;80(8):e141–9. doi:10.1016/j.ijmedinf.2010.09.009.
16. Collins S, Bakken S, Vawdrey DK, Coiera E, Currie LM. Agreement between common goals discussed and documented in the ICU. *J Am Med Inf Assoc*. 2011;18(1):45–50. doi:10.1136/jamia.2010.006437.
17. Collins S, Currie LM, Bakken S, Cimino JJ. Information needs, Infobutton Manager use, and satisfaction by clinician type: a case study. *JAMIA*. 2009;16(1):140–2. doi:10.1197/jamia.M2746.
18. Boon H, Verhoef M, O'Hara D, Findlay B. From parallel practice to integrative health care: a conceptual framework. *BMC Health Serv Res*. 2004;4(1):15. doi:10.1186/1472-6963-4-15.
19. Gaboury I, Boon H, Verhoef M, Bujold M, Lapierre LM, Moher D. Practitioners' validation of framework of team-oriented practice models in integrative health care: a mixed methods study. *BMC Health Serv Res*. 2010;10(1):289. doi:10.1186/1472-6963-10-289.
20. Henry S, Holzemer WL. A comparison of problem lists generated by physicians, nurses, and patients: implications for CPR systems. In: *Proc Annu Symp Comput Appl Med Care*.; 1995:382–6.
21. Warren JJ, Collins J, Sorrentino C, Campbell JR. Just-in-time coding of the problem list in a clinical environment. In: *AMIA Annual Symposium proceedings*.; 1998:280–4.
22. Mazzocco K, Petitti DB, Fong KT, et al. Surgical team behaviors and patient outcomes. *Am J Surg*. 2009;197(5):678–85. doi:10.1016/j.amjsurg.2008.03.002.
23. Bakken S, Ruland CM. Translating clinical informatics interventions into routine clinical care: how can the RE-AIM framework help? *J Am Med Inform Assoc*. 2009;16(6):889–97. doi:10.1197/jamia.M3085.
24. Collins S, Hurley A., Chang F., Benoit A., Illa A., Laperle S. DP. Content and functional specifications for a standards-based multidisciplinary rounding tool to maintain continuity across acute and critical care. *J Am Med Inf Assoc*. 2013;in press.
25. Collins S, Mamykina L, Jordan D, et al. In Search of Common Ground in Handoff Documentation in an Intensive Care Unit. *J Biomed Inf*. 2012;45(2):307–315.
26. Collins S, Dykes PC, Sc DN, et al. Closed Loop Care Coordination: The Critical Linkages and Shared Concepts. In: *AMIA Annual Symposium proceedings [submitted for panel presentation]*.; 2013:3–5.
27. Del Fiol G, Curtis C, Cimino JJ, et al. Disseminating context-specific access to online knowledge resources within electronic health record systems. *Stud Health Technol Inform*. 2013;192:672–6.

Development and Implementation of a Real-Time 30-Day Readmission Predictive Model

Patrick R. Cronin, MA[1], Jeffrey L. Greenwald, MD[2], Gwen C. Crevensten, MD[2], Henry C. Chueh, MD, MS[1], Adrian H. Zai, MD, PhD[1]

[1]Laboratory of Computer Science and [2]Department of Medicine
Massachusetts General Hospital, Boston, MA

Abstract

Hospitals are under great pressure to reduce readmissions of patients. Being able to reliably predict patients at increased risk for rehospitalization would allow for tailored interventions to be offered to them. This requires the creation of a functional predictive model specifically designed to support real-time clinical operations. A predictive model for readmissions within 30 days of discharge was developed using retrospective data from 45,924 MGH admissions between 2/1/2012 and 1/31/2013 only including factors that would be available by the day after admission. It was then validated prospectively in a real-time implementation for 3,074 MGH admissions between 10/1/2013 and 10/31/2013. The model developed retrospectively had an AUC of 0.705 with good calibration. The real-time implementation had an AUC of 0.671 although the model was overestimating readmission risk. A moderately discriminative real-time 30-day readmission predictive model can be developed and implemented in a large academic hospital.

Keywords: Predictive Modeling, 30-Day Readmissions, Real-Time, Readmission Risk

Introduction

The affordable care act of 2010 established the Hospital Readmissions Reduction Program, requiring the Centers for Medicare and Medicaid Services to reduce payments to hospitals if the cost of readmission rates at a given hospital exceeded predicted costs[1]. As a result, hospitals, academic groups and independent private organizations have invested substantial resources to reduce readmissions[2]. One area of investment is the identification of patients at high risk for rehospitalization in the 30-days after an index hospitalization in order to enable targeted resource allocation to this population.

Identification of patients at high risk for early readmission using predictive modeling has been well documented in the literature. A 2011 systematic review of 26 validated readmission prediction models concluded that most had poor predictive ability but may be useful in certain settings[3]. Importantly, 23 models could not be implemented early during a hospitalization (i.e. in real-time) because data used in the prediction model was not available until after discharge or not available in a structured form. Three models were identified as usable in real-time [4-6]; however, two were predicting 12-month readmission[5,6], and the other was developed on a small population, 1029 patients with congestive heart failure[4]. Other early readmission prediction models were developed for cardiovascular events [7-12], pancreatitis[13], kidney transplant [14], and to identify avoidable early readmissions,[15] though there is no agreement in the literature on exactly how to define "avoidable." More recently developed early readmission models have improved discrimination, and two reported AUCs of 0.75 [16] and 0.77 [17] However, both models required

data that was not available in real-time at the MGH. We did not find any published 30-day real-time readmission prediction models that could be feasibly implemented at our hospital.

As such, we desired to create a model that included elements that would be commonly available on patients admitted to the hospital at the beginning of the admission and accessible electronically in real-time so that risk identification could begin prospectively early in the hospitalization. We defined *real-time* as being able to calculate the 30-day readmission risk the day after the patient was admitted.

Methods

This project was completed at the Massachusetts General Hospital (MGH). The MGH is an academic hospital with 957 licensed beds with approximately 48,000 admissions and 95,000+ emergency visits annually. Data for this project was extracted from existing databases maintained by the MGH for operational and research purposes. We extracted 45,924 hospital admissions for patients who arrived between 2/1/2012 and1/31/2013; of these, 5,570 (12.1%) were readmissions to MGH within 30 days of discharge. This data was merged with inpatient transactional data, emergency department historical data, billing data, laboratory tests, medication orders, and outpatient appointment history. Data was split 80:20 for developing and validating the predictive model.

Readmissions were identified by the presence of an inpatient record, with a subsequent admission date by the same patient (identified by Medical Record Number) within 30 days of the index admission discharge date. In the event that there were multiple encounters within 30 days of discharge, the readmission evaluated is the first one occurring after discharge. Any readmission can become an index admission if there is another encounter following which occurred within 30 days of the prior discharge. Please note, this only included readmissions to the MGH, and is not the same metric as the publicly reported readmission rates which include readmissions to other hospitals. Moreover, the following exclusions were applied to both the index admissions and the readmissions:

- *From the Index Admission and Readmissions: patients discharged to* rehabilitation and hospice
- *From the Index Admission Only:* discharge status of deceased, left against medical advice, transferred to another short term acute facility, discharged/transferred to a psychiatric hospital
- *From the Readmission Only*: Chemotherapy, radiation, dialysis, Obstetrics (birth/delivery)

We performed a literature review to identify predictors from published models. Then we reviewed the existing data infrastructure at the MGH to determine which variables were available and the "lag" on their availability for a real-time implementation. Two hospitalists and another physician were consulted regularly to provide feedback to identify which variables made clinical sense to be included in the model. The variables were then weighted based on their availability in real-time. We identified 40 variables, and used logistic regression to develop our predictive model. Variables that were not statistically significant or that did not meaningfully

improve the model's discrimination or calibration were removed from consideration. We proceeded iteratively with close consultation with three physicians. We purposely did not use a formal backwards elimination process because our goal was to implement a real-time model and factors other than statistical significance (i.e. resources required to access data in real-time) were essential to our predictor selection process. In addition, we aimed to maintain flexibility to perform data transformations to improve discrimination, calibration, and feasibility of implementation.

The two statistics used to measure the model's performance were the area under the ROC curve (AUC) and the calibration. We calculated the calibration by splitting the data into ten groups of lowest-to-highest risk and plotting the expected-versus-observed outcomes.

After we validated our model, we implemented it in TopCare [18], our hospital's population health management system. Using the developed model, we prospectively calculated the AUC, the calibration, and the 30-day readmission risk probability for all admissions between 10/1/2013 and 10/31/2013.

Results

Based on the data available one year prior to the admission at MGH, we split the derivation dataset into the following categories:

- No History: No prior admissions and no coded Elixhauser Comorbidities[19] from outpatient data
- Has Comorbidity: No prior admissions but at least 1 Elixhauser Comorbidity from outpatient data
- Has Inpatient History: Had at least 1 prior admission to the MGH

We developed three separate predictive models with this data and merged the results back together to calculate the AUC and calibration. The following variables were identified as potentially predictive but were not included because they did not meaningfully improve the final model: age, sex, race, language, insurance, presence of advanced directive, marital status, restraint use, means of arrival, count of emergency department visits, count of no-shows in the past year, count of prior admissions, testing positive for illegal drugs, warfarin, rituximab, oral hypoglycemic agents, oral antiplatelet agents, sotalol, oxycodone, fentanyl, hydrocodone, meperdine, morphine equivalent narcotic dose, sedative use, and presence of an emergency department psychiatric consult.

The following variables were included in the final model (Table 1): unplanned admission (Yes/No); admission source (i.e., physician's office, transfer from long-term care, transfer from rehabilitation or skilled nursing facility, ambulance, walk-in); had a notation of drug abuse (ICD-9 code) or homelessness (internal system flag) or had left against medical advice (identified from prior MGH discharges); count of Elixhauser comorbidities based on inpatient and outpatient ICD-9 codes in the past year; number of inpatient days at the MGH in the past year; insulin (Yes\No); antipsychotic (Yes\No); other medications(Yes\No) (furosemide $\geq$ 40 mg, metolazone, lactulose, cyclosporine, or heparin $\geq$ 3000 units). The list of medications was derived from a list of drugs related to adverse drug events[20] and supplemented with recommendations from the

Table 1: Comparison of Model Parameters in Retrospective and Prospective Validation Data sets

Variables	Retrospective Validation				Prospective Validation			
	N(%)		Readmit (%)		N(%)		Readmit (%)	
History with MGH (past year)								
None	3,583	(38%)	221	6.2%	1,347	(42%)	75	5.6%
Outpatient Only	2,554	(27%)	221	8.7%	693	(22%)	53	7.6%
Prior Admission	3,210	(34%)	666	20.7%	1,134	(36%)	165	14.6%
Admission Source								
Physician	4,345	(46%)	355	8.2%	1,431	(45%)	96	6.7%
Transfer	1,266	(14%)	135	10.7%	450	(14%)	48	10.7%
Other	1,942	(21%)	294	15.1%	675	(21%)	76	11.3%
Emergency Med Svc.	1,582	(17%)	276	17.4%	565	(18%)	68	12.0%
SNF or Rehab	212	(2%)	48	22.6%	53	(2%)	5	9.4%
Admission Type								
Planned	2,578	(28%)	181	7.0%	821	(26%)	46	5.6%
Unplanned	6,769	(72%)	927	13.7%	2,353	(74%)	247	10.5%
Behavioral Issue (Past Year)								
No	9,009	(96%)	1,027	11.4%	3,025	(95%)	274	9.1%
Yes	338	(4%)	81	24.0%	149	(5%)	19	12.8%
Elixhauser Comorbidity Count (past year data)								
0	3,949	(42%)	272	6.9%	1,517	(48%)	86	5.7%
1	1,290	(14%)	111	8.6%	357	(11%)	31	8.7%
2	948	(10%)	94	9.9%	295	(9%)	27	9.2%
3	759	(8%)	109	14.4%	222	(7%)	24	10.8%
4	619	(7%)	119	19.2%	190	(6%)	20	10.5%
5	483	(5%)	87	18.0%	154	(5%)	22	14.3%
6	393	(4%)	71	18.1%	128	(4%)	25	19.5%
7	324	(3%)	79	24.4%	86	(3%)	18	20.9%
8+	582	(6%)	166	28.5%	225	(7%)	40	17.8%
Inpatient Bed Days (past year)								
0	6,137	(66%)	442	7.2%	2,040	(64%)	128	6.3%
1-5	862	(9%)	122	14.2%	395	(12%)	35	8.9%
6-10	719	(8%)	134	18.6%	241	(8%)	32	13.3%
10-20	721	(8%)	154	21.4%	229	(7%)	45	19.7%
20+	908	(10%)	256	28.2%	269	(8%)	53	19.7%
Insulin								
No	6,951	(74%)	719	10.3%	2,423	(76%)	188	7.8%
Yes	2,396	(26%)	389	16.2%	751	(24%)	105	14.0%
Antipsychotic								
No	7,854	(84%)	877	11.2%	2,399	(76%)	218	9.1%
Yes	1,493	(16%)	231	15.5%	775	(24%)	75	9.7%
Other Medication								
No	7,830	(84%)	812	10.4%	2,798	(88%)	238	8.5%
Yes	1,517	(16%)	296	19.5%	376	(12%)	55	14.6%

physician collaborators. Please note, all aforementioned medications were ordered during the patient's index admission.

Using the training data set of 36,462 records, the overall model had an AUC of 0.705 (95% CI 0.697 to 0.713). Odds ratios were calculated for each of the sub-models in Figure 1 and the model factors are illustrated in Table 2. The logistic coefficients were applied to the validation data set of 9,325 records and the model had an AUC of 0.714 (95% CI 0.698 to 0.730) and the calibration is in Figure 2. The model was then implemented in real-time for 3,074 admissions between 10/1/2013 and 10/31/2013. It had an AUC of 0.671 and the calibration is in Figure 3. The breakdowns of the readmission rates by each model variable for the retrospective and prospective validation is in table 1.

Figure 1: Odds Ratios for Model Parameters

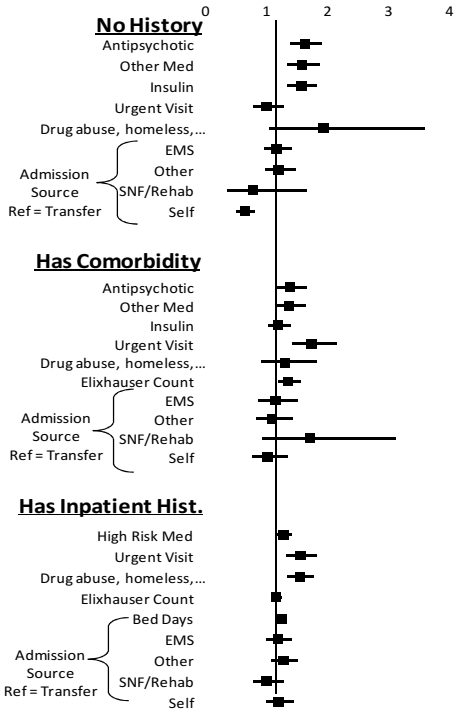


Figure 2: Retrospective Validation ROC Curve and Calibration

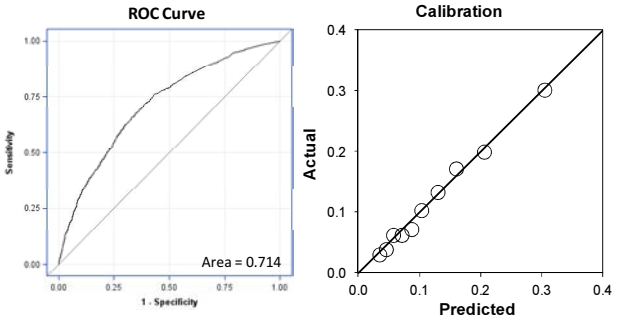


Figure 3: Prospective Validation ROC Curve and Calibration

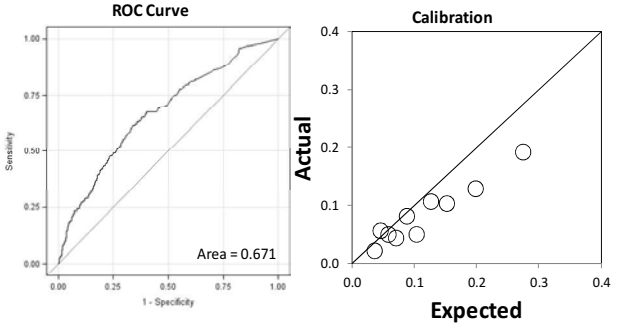


Table 2: Maximum Likelihood Estimates (MLE) of Model Parameters

Variables	No History		Outpatient Only		Has Prior Admission	
	MLE ± Error	P	MLE ± Error	P	MLE ± Error	P
Intercept	-2.904 ± 0.147	<.01	-3.189 ± 0.159	<.01	-2.944 ± 0.108	<.01
Admission Source						
Transfer (ref)	-0.091		0.145		0.110	
Emergency Med Svc.	0.227 ± 0.105	0.03	-0.024 ± 0.092	0.80	0.050 ± 0.048	0.30
SNF or Rehab	-0.193 ± 0.318	0.54	0.378 ± 0.234	0.11	-0.122 ± 0.085	0.15
Physician	-0.384 ± 0.117	<.01	-0.140 ± 0.092	0.13	0.061 ± 0.051	0.23
Other	0.259 ± 0.105	0.01	-0.070 ± 0.085	0.41	0.121 ± 0.045	<.01
Visit Type						
Urgent Visit (Yes)	-0.012 ± 0.127	0.93	0.544 ± 0.106	<.01	0.433 ± 0.080	<.01
History (Past year)						
Behavioral Issue	0.652 ± 0.318	0.04	0.245 ± 0.176	0.16	0.422 ± 0.072	<.01
Elixhauser Count			0.293 ± 0.068	<.01	0.132 ± 0.037	<.01
Inpatient Days					0.216 ± 0.017	<.01
Inpatient Medications						
Insulin	0.442 ± 0.090	<.01	0.305 ± 0.091	<.01	0.236 ± 0.053	<.01
Antipsychotic	0.478 ± 0.084	<.01	0.316 ± 0.090	<.01		
High Risk Drug	0.437 ± 0.079	<.01	0.160 ± 0.079	0.04		

Discussion

This project was a practical demonstration of developing and implementing a real-time 30-day readmission prediction model for clinical operational purposes in a large academic medical center. The model was moderately discriminative, and was successfully implemented in TopCare, our hospital's population health management system. The infrastructure now exists at the MGH to notify providers early about patients who are at high risk for readmission within 30-days of discharge.

The AUC of the prospective real-time validation was similar to the retrospective validation. Although the model overestimated the outcome in the prospective validation, we attributed this in large part to the difference between the rates of readmission rates in the two groups (9.23% in the validation set and 12.1% in the derivation cohort). We also identified an issue where bed days in the prior year were being over-represented due to duplicates in the database, resulting in an overestimation of readmission risk.

As we developed and implemented this real-time model, we learned three important lessons essential to a successful implementation of a 30-day real-time readmission predictive model. First, risk scores had to be calculated daily for all inpatients beginning no later than one day after the patients' admission date. The timeliness of this calculation was required because post-discharge interventions associated with reducing readmission had to be planned soon after a patient's admission. As a result, we were forced to exclude many popular readmission predictors including length-of-stay and discharge diagnosis, as those variables are not available upon admission. Second, complexity had to be limited to readily available technical capabilities. For example, we excluded predictors identified using keyword searches because highly accurate

parsing could not be implemented easily. Third, the model had to make sense to hospitalists. Certain hospitalists had trouble believing the validity of the prior implemented model because it did not make sense clinically. We addressed this issue by inviting hospitalists to assist us with the development of the model. The success of the implementation of a real-time model required a fine balance between clinical believability, statistical requirements, availability of real-time data, and technical capabilities.

Limitations: The model was designed with data from a single institution from data for a single year, and may not be replicable at other institutions with different available data sources. In addition, we did not try to determine if a patient was readmitted at a non-MGH institution and are therefore underestimating total readmissions.

Conclusion

A moderately discriminative real-time 30-day readmission predictive model can be successfully implemented in a large academic hospital using existing data. Developers of real-time clinical predictive models need to consider more than the discrimination when developing models. An implementable real-time model balances clinical priorities, statistical requirements, availability of real-time data, and technical requirements.

References

1. Joynt KE, Jha AK. Characteristics of hospitals receiving penalties under the Hospital Readmissions Reduction Program. *JAMA : the journal of the American Medical Association.* Jan 23 2013;309(4):342-343.
2. Evans M. Healthcare's 'moneyball'. *Predictive modeling being tested in data-driven effort to strike out hospital readmissions.* 2011. http://www.modernhealthcare.com/article/20111010/MAGAZINE/111009989#. Accessed 3/12/2014.
3. Kansagara D, Englander H, Salanitro A, et al. Risk prediction models for hospital readmission: a systematic review. *JAMA : the journal of the American Medical Association.* Oct 19 2011;306(15):1688-1698.
4. Amarasingham R, Moore BJ, Tabak YP, et al. An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. *Medical care.* Nov 2010;48(11):981-988.
5. Billings J, Dixon J, Mijanovich T, Wennberg D. Case finding for patients at risk of readmission to hospital: development of algorithm to identify high risk patients. *BMJ.* Aug 12 2006;333(7563):327.
6. Billings J, Mijanovich T. Improving the management of care for high-cost Medicaid patients. *Health Aff (Millwood).* Nov-Dec 2007;26(6):1643-1654.
7. Wasfy JH, Rosenfield K, Zelevinsky K, et al. A prediction model to identify patients at high risk for 30-day readmission after percutaneous coronary intervention. *Circulation. Cardiovascular quality and outcomes.* Jul 2013;6(4):429-435.
8. Brown JR, Conley SM, Niles NW, 2nd. Predicting Readmission or Death After Acute ST-Elevation Myocardial Infarction. *Clinical cardiology.* Oct 2013;36(10):570-575.
9. Watson AJ, O'Rourke J, Jethwani K, et al. Linking electronic health record-extracted psychosocial data in real-time to risk of readmission for heart failure. *Psychosomatics.* Jul-Aug 2011;52(4):319-327.

10. Wallmann R, Llorca J, Gomez-Acebo I, Ortega AC, Roldan FR, Dierssen-Sotos T. Prediction of 30-day cardiac-related-emergency-readmissions using simple administrative hospital data. *International journal of cardiology.* Apr 5 2013;164(2):193-200.
11. Hammill BG, Curtis LH, Fonarow GC, et al. Incremental value of clinical data beyond claims data in predicting 30-day outcomes after heart failure hospitalization. *Circulation. Cardiovascular quality and outcomes.* Jan 1 2011;4(1):60-67.
12. Au AG, McAlister FA, Bakal JA, Ezekowitz J, Kaul P, van Walraven C. Predicting the risk of unplanned readmission or death within 30 days of discharge after a heart failure hospitalization. *American heart journal.* Sep 2012;164(3):365-372.
13. Whitlock TL, Tignor A, Webster EM, et al. A scoring system to predict readmission of patients with acute pancreatitis to the hospital within thirty days of discharge. *Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association.* Feb 2011;9(2):175-180; quiz e118.
14. McAdams-DeMarco MA, Law A, Salter ML, et al. Frailty and early hospital readmission after kidney transplantation. *American journal of transplantation : official journal of the American Society of Transplantation and the American Society of Transplant Surgeons.* Aug 2013;13(8):2091-2095.
15. Donze J, Aujesky D, Williams D, Schnipper JL. Potentially avoidable 30-day hospital readmissions in medical patients: derivation and validation of a prediction model. *JAMA internal medicine.* Apr 22 2013;173(8):632-638.
16. He D, Mathews SC, Kalloo AN, Hutfless S. Mining high-dimensional administrative claims data to predict early hospital readmissions. *Journal of the American Medical Informatics Association : JAMIA.* Sep 27 2013.
17. van Walraven C, Wong J, Forster AJ. LACE+ index: extension of a validated index to predict early death or urgent readmission after hospital discharge using administrative data. *Open medicine : a peer-reviewed, independent, open-access journal.* 2012;6(3):e80-90.
18. TopCare Powered by Blender Patient Population Management Software. Fort Lauderdale, FL, FL: SRG Tech, Inc.: SRG Technology; 2013.
19. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. *Medical care.* Jan 1998;36(1):8-27.
20. Budnitz DS, Lovegrove MC, Shehab N, Richards CL. Emergency hospitalizations for adverse drug events in older Americans. *The New England journal of medicine.* Nov 24 2011;365(21):2002-2012.

Acknowledgements

The authors thank Steven Wong, for his invaluable assistance with collecting data for this study, and Tina Rong, for her critical role with implementation of the pilot study.

A Semantic-based Approach for Exploring Consumer Health Questions Using UMLS

Licong Cui[1], PhD, Shiqiang Tao[1,2], MS, Guo-Qiang Zhang[1,2], PhD
[1]Department of EECS, Case Western Reserve University, Cleveland, OH
[2]Division of Medical Informatics, Case Western Reserve University, Cleveland, OH

Abstract

NetWellness is a non-profit web service providing high quality health information. It has been in operation since 1995 with over 13 million visits per year by consumers across the world in recent years. Consumer questions in NetWellness have been answered by medical and health professional faculties at three Ohio partner universities: Case Western Reserve University, the Ohio State University, and University of Cincinnati. However, the resident interface in NetWellness is ineffective in searching existing questions that have already been carefully answered by experts in an easy-to-understand manner. In our previous work, we presented a Conjunctive Exploratory Navigation Interface (CENI) reusing NetWellness' 120 pre-defined health topics in assisting question retrieval. This paper presents a novel semantic-based search interface called Semantic Conjunctive Exploratory Navigation Interface (SCENI), using UMLS concepts as topics. 60,000 questions were tagged by UMLS Concept Unique Identifies (CUIs), with each question allowing multiple possible tags. Using a slightly modified 5-point Likert scale for relevance, SCENI reveals improved precision and relevance (precision: 93.47%, relevance: 4.31) in comparison to CENI using NetWellness' pre-defined topics alone (precision: 77.85%, relevance: 3.3) and NetWellness' resident search interface (precision: 50.62%, relevance: 1.97), on a set of sample queries.

Introduction

Although a substantial amount of consumer health information is available online [1], it is not necessarily easy for general consumers to access such information. For example, a study reported in JAMA [2] by Berland et al. found that accessing health information by use of search engines (e.g., Google or Yahoo!) and simple search terms was not sufficient. Only less than a quarter of links on the search engines first pages of search results led to relevant content.

To improve health information retrieval, we developed a Conjunctive Exploratory Navigation Interface (CENI [3]) for exploring NetWellness [7] health questions with health topics as dynamic and searchable menus complementing lookup search. The efficacy of CENI was evaluated by comparing it with a similar search interface with keyword-based search only, as well as the existing search mode using Google search or NetWellness advanced search. The evaluation was conducted through crowdsourcing, a valuable method for gathering data when human participation is needed. Our crowdsourced evaluation of CENI with a comparative study of search interfaces with anonymous, paid participants recruited from an online labor marketplace called Amazon Mechanical Turk (AMT) showed a nearly 2-1 ratio in preference of CENI for 9 carefully designed search tasks. Participants indicated that CENI was easy to use, provided better organization of health questions by topics, allowed users to narrow down to the most relevant contents quickly, and supported the exploratory navigation by non-experts or those unsure how to initiate their search.

However, a limitation of CENI is that the health topics are predefined, reusing those from NetWellness in order to have a side-by-side comparative study. Now that the effectiveness of the CENI interface has been established, this paper presents our work in using UMLS concept labels as topics, resulting in a semantic-based system called Semantic Conjunctive Exploratory Navigation Interface (SCENI). This way, SCENI shares CENI's benefits in its use of topics as dynamic and searchable menus for consumer health information retrieval and navigation, and in allowing users to quickly narrow down to the most relevant results, without the restriction of a relatively small set of predefined topics.

1 Background

1.1 Unified Medical Language System (UMLS) and MetaMap

The Unified Medical Language System (UMLS) [4], developed by the US National Library of Medicine (NLM), is perhaps the largest integrated repository of biomedical vocabularies. The 2014AA release of UMLS covers over 2.9 million concepts from more than 150 source vocabularies. Vocabularies integrated in the UMLS Metathesaurus include the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT), Consumer Health Vocabulary (CHV), National Center for Biotechnology Information (NCBI) taxonomy, the Medical Subject Headings (MeSH),

RxNorm, and International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM). UMLS also provides a set of broad subject categories, or semantic types, that allows for the semantic categorization of all concepts. Each UMLS concept has a unique concept identifier (CUI), and is assigned one or more semantic type(s) of the 135 semantic types (e.g., Disease or Syndrome, Body Location or Region).

MetaMap [5, 6], also provided by NLM, is the state-of-the-art solution for the annotation of UMLS Metathesaurus concepts and semantic types. It is available via web access, a downloadable Java implementation (MMTx), an Application Programming Interface (API), and a downloadable version of the complete Prolog implementation of MetaMap itself. MetaMap associates biomedical terms to UMLS concepts and assigns a CUI to every term that it can identify.

1.2 NetWellness

NetWellness [7, 8] is one of the first consumer health websites that has been in existence for 19 years. It is a community service providing high quality, unbiased health information created and evaluated by medical and health professional faculties at Case Western Reserve University, the Ohio State University, and University of Cincinnati. These health professionals include physicians, nurses, pharmacists, dietitians and dentists, and serve as "experts" in NetWellness' popular "Ask an Expert" service. Health questions in NetWellness have been generated by consumers and answered by experts after editorial review to make them readable and understandable to consumers.

NetWellness has over 60,000 consumer questions categorized into 120 health topics. Table 1 shows the top 10 health topics ranked by the number of questions related to each topic. Each question in NetWellness consists of four major components: Health Topic, Subject, Question, and Answer (See Figure 1 for a sample question). Each question is assigned one health topic by a consumer when asking the question. The Subject is the title of a question and usually contains the key topical information of a question. Question provides more details on the subject. Answers are given by an expert in the related health topic area.

Table 1: Top 10 health topics ranked by the number of questions in NetWellness.

Topics	**# of Questions**
Pharmacy and Medications	3802
Ear, Nose, and Throat Disorders	3481
Pregnancy	3209
Children's Health	2827
Myasthenia Gravis	2470
Diet and Nutrition	2335
Women's Health	2310
Eye and Vision Care	2249
Lung and Respiratory	1674
Kidney Diseases	1624

1.3 Conjunctive Exploratory Navigation Interface (CENI)

Since each question is assigned a single topic in NetWellness, this presents an impediment to access health questions through multiple pathways in navigational exploration. In [9], a multi-topic assignment method using formal concept analysis was proposed to categorize a question into multiple relevant topics. It achieved a 36.5% increase in recall with virtually no sacrifice in precision compared to NetWellness' original single topic assignment. Based on the results of the multi-topic assignment approach, a novel Conjunctive Exploratory Navigation Interface called CENI [3] was developed for supporting effective retrieval of consumer health questions using health topics in NetWellness. The effectiveness of the CENI interface was evaluated through crowdsourcing and received a nearly 2-1 ratio of preference compared to two other search modes.

However, CENI is limited in the number of health topics that are adopted from NetWellness' existing ones, which may not represent the best choices of potential health topics. In this paper, we present semantic-based SCENI by mining a relatively large collection of UMLS concepts as health topics and tagging each question with the most relevant topics.

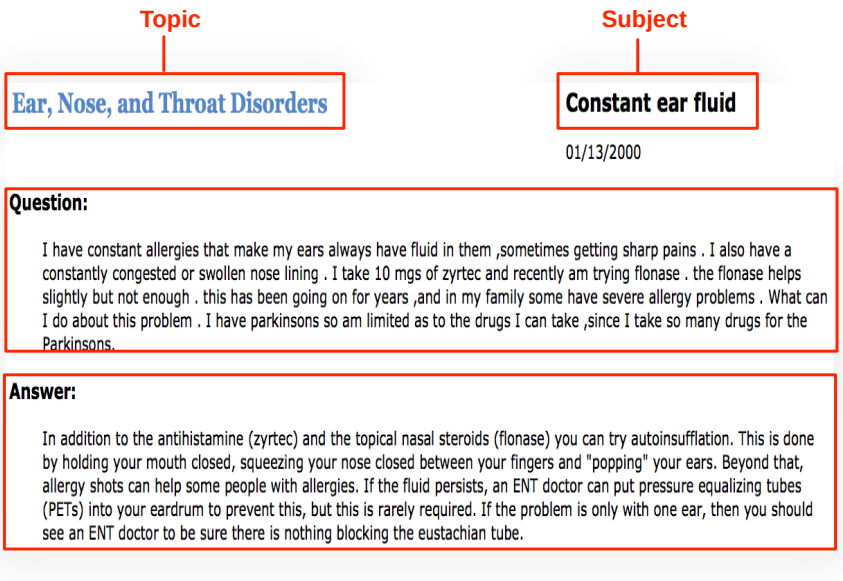


Figure 1: A sample question with the four major components: Health Topic, Subject, Question, and Answer.

2 Methods

Figure 2 depicts the higher level architecture of the proposed UMLS-concept-based tagging approach to facilitating consumer health questions retrieval in NetWellness. First, all the health questions in NetWellness are processed and tagged with UMLS concepts, allowing the handling of synonyms. Second, the most relevant concepts for each question are selected using concept TF-IDF and consumer health vocabulary. Then, resulting concepts for all the questions constitute a concept cloud, which are indexed for retrieving questions in the semantic search interface SCENI.

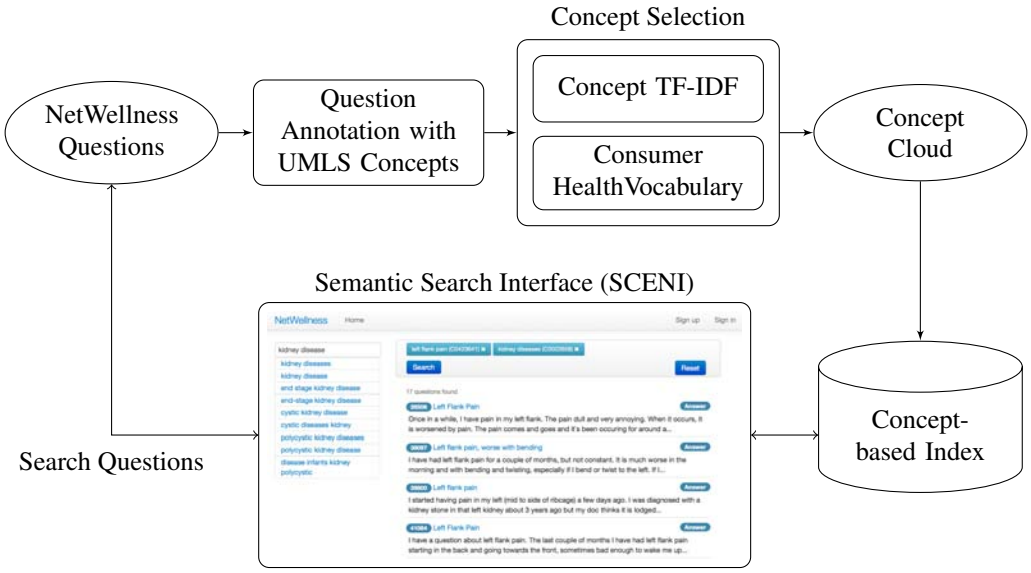


Figure 2: Overview of semantic-based approach to exploring consumer health questions using UMLS concepts.

2.1 Question Annotation with UMLS Concepts

For each health question in NetWellness, the text in its topic, subject and question are processed using MetaMap and mapped to UMLS concepts (CUIs) and semantic types. Since there is a large number of CUIs involved and not all of

them are relevant to consumer heath, we manually identified 31 semantic types that are less meaningful for retrieving health questions, including "Qualitative Concept," "Quantitative Concept," "Intellectual Product," "Regulation or Law," and "Geographic Area." Concepts with such semantic types were filtered out. Moreover, we only consider non-negated concepts in order to remove the noise that negated concepts brought for retrieving related questions based on concepts. After this process, each question is annotated with a set of UMLS CUIs representing the question's relevant concepts. For example, the question shown in Figure 1 is annotated with a set of 15 CUIs shown in Table 2, where each CUI takes the form of "CUI: Preferred concept name," and each number in the parenthesis represents the count of occurrences of the CUI in this question (the parenthesis is omitted if the CUI occurs only once).

Table 2: A set of 15 UMLS concepts annotated for the question shown in Figure 2. The number of occurrences of each CUI is indicated in parenthesis.

C0013447: Ear Diseases	C0028432: Nose Diseases	C0544058: Throat Disorders
C0013443: Ear (2)	C0005889: Body Fluids (2)	C0020517: Allergy
C1881534: Make	C0455270: Sharp pain	C0240577: Swollen nose
C0162723: Zyrtec	C0286677: Flonase (2)	C2945656: Severe Allergy
C0033213: Problem	C0439801: Limited	C0013227: Drug (2)

2.2 Concept Selection

For each question, we select a smaller set of the most relevant concepts adopting the idea of the Term Frequency-Inverse Document Frequency (TF-IDF) metric [10], a most common term weighting scheme in information retrieval and text mining. In this paper, we calculate TF-IDF based on CUIs instead of terms.

Hence the term frequency (TF) for a CUI c in a question q is calculated as the number of occurrences of the CUI in the question and it's topic and subject ($f(c, q)$), normalized by the number of all CUI occurrences in that question ($\sum\{f(w, q) : w \in q\}$), as follows:

$$\text{tf}(c, q) = \frac{f(c, q)}{\sum\{f(w, q) : w \in q\}}.$$

The inverse document frequency (IDF) is used to measure the importance of a CUI c in a corpus of questions Q. It is calculated as the logarithm of the quotient of the number of all questions ($|Q|$) and the number of questions containing the CUI ($|\{q \in Q : c \in q\}|$), as follows:

$$\text{idf}(c, Q) = \log \frac{|Q|}{|\{q \in Q : c \in q\}|}.$$

To avoid division-by-zero, we adjust the denominator to $1 + |\{q \in Q : c \in q\}|$. The TF-IDF weight, calculated as $\text{tf-idf}(c, q, Q) = \text{tf}(c, q) \times \text{idf}(c, Q)$, is used to determine the importance of the CUI c for the question q.

For each question, its top 5 CUIs ranked by TF-IDF weight are selected for tagging the question. We also keep the CUIs identified for the topic and subject of a question, since they contain key information of the question in very short text. Furthermore, CUIs not included in the Consumer Health Vocabulary (CHV) [11, 12] are discarded.

After this process, each question is annotated with a smaller set of CUIs representing the most relevant concepts. For example, Table 3 displays the resulted set for the question in Figure 1. Combining the most relevant CUIs of all the questions with duplicates removed results in a concept cloud.

Table 3: A set of 7 most relevant UMLS concepts representing the question shown in Figure 1.

C0013447: Ear Diseases	C0028432: Nose Diseases	C0544058: Throat Disorders
C0013443: Ear	C0005889: Body Fluids	C0240577: Swollen nose
C0286677: Flonase		

2.3 Concept-based Indexing and Semantic-based Conjunctive Search

The concept cloud is indexed using Picky [13], a Ruby-based semantic text search engine for categorized data. Both CUIs and the strings they represent are indexed to support quick response of concept searching by users. For each health question, the most relevant CUIs obtained after the concept selection process are indexed to support question retrieval using a semantic search interface SCENI. SCENI is implemented in Ruby on Rails and reused the interface design of CENI. SCENI is designed to deal with conjunctive search in a large number of questions and concepts. In SCENI, given a set of concepts to search, their synonyms were also used to better retrieve related questions.

2.4 Evaluation

We evaluated the proposed semantic-based approach in two ways. One is to evaluate if the concept selection process obtains the most relevant concepts for tagging questions. The other is to compare the proposed concept-based conjunctive search with two other existing search modes of NetWellness questions.

2.4.1 Concept Selection

Since no well-established reference standard is available, two experts in health informatics created a reference standard of most relevant concepts for a set of 50 health questions in NetWellness. In [9], a set of randomly selected 300 health questions in NetWellness were used for evaluating the multi-topic assignment method, and resulted in a subset of 278 questions with good quality as the reference standard. In this paper, we randomly chose 50 questions among these 278 and developed a web-based annotation interface for the annotators to review questions and tagging them with the most related concepts. Since it is hard for the annotators to come up with UMLS concepts by themselves, we provided a baseline set of UMLS concepts for each question identified by MetaMap after removing concepts with less meaningful semantic types. The two annotators reviewed the 50 questions together, resolved any disagreements, and created the reference standard.

2.4.2 Semantic-based Conjunctive Search

To evaluate the semantic-based conjunctive search, we designed 5 search tasks (see Table 4) to compare three search modes: the key-word based search in NetWellness official website (NWO) [7], the topic-based conjunctive search in CENI [3], and the proposed UMLS concept-based conjunctive search in SCENI.

Table 4: List of five search tasks.

Search Task ID	Search Task Description
1	Can anti-epileptic medications be taken during pregnancy?
2	Is colon cancer an inherited disease?
3	Is it safe to take birth control pills when breastfeeding?
4	Is it possible to contract HIV from toilets?
5	Can hypertension cause heart attack?
6	Does Keppra cause hair loss?
7	Does drinking alcohol afffect emphysema?
8	What diet would help with gastroesophageal reflux disease (GERD)?
9	What are possible causes of infant sleep apnea?
10	Does toothpaste cause allergy?

For each search task in Table 4, an expert in health informatics used each search mode to retrieve a list of questions, and gave a relevance score for each of the retrieved question using a slightly modified 5-point Likert scale [14], where 0 indicated not relevant at all, and values between 1 and 5 were considered relevant (1 indicated weakly relevant and 5 indicated strong relevant).

3 Results

3.1 Concept Statistics

Over 60000 health questions in NetWellness were processed. MetaMap identified 32195 distinct relevant UMLS CUIs for all questions, their topics and subjects. 23955 CUIs were obtained after filtering out uninformative ones using their semantic types. 21365 CUIs remained left after performing CUI TF-IDF. Removing CUIs that were not in CHV resulted in 18538 CUIs, which constituted a cloud of concepts. Figure 3 displays the top 15 concepts and the numbers of questions they occur in.

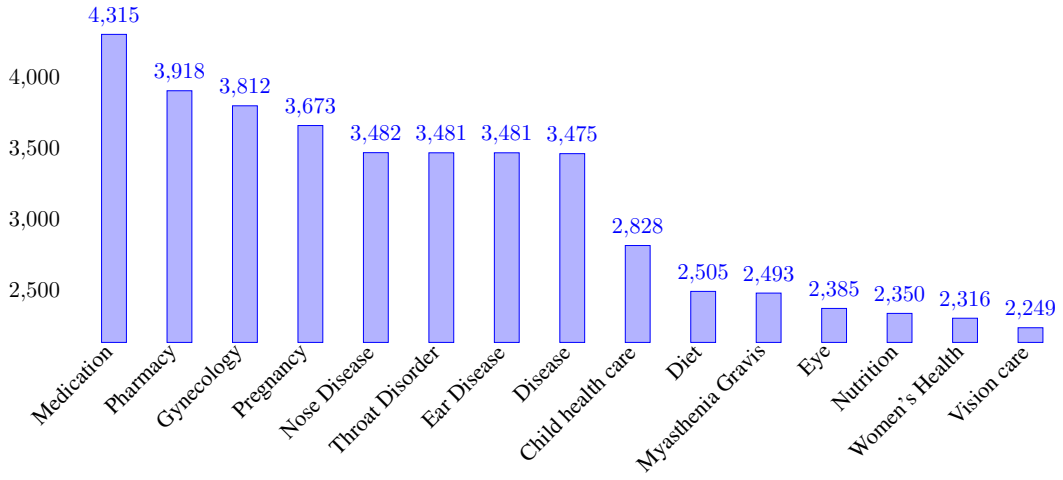


Figure 3: Top 15 UMLS concepts occurring in the NetWellness questions. The x-axis indicates the UMLS concept name, and the y-axis indicates the number of questions each concept occurs in.

3.2 Semantic Conjunctive Exploratory Navigation Interface (SCENI)

The general layout of SCENI interface is illustrated in Figure 4. There are four general areas of SCENI:

1. *Area for searching concepts* allows a user to type a search string into a search box and retrieve concepts of interest among the concept cloud;
2. *Area for displaying candidate concepts matching the search string* lists the matched concepts for a user to select from. Clicking any concept in the list automatically adds a concept tag in the area for displaying selected concepts (red arrow);
3. *Area for displaying selected concepts* shows user-selected concepts. Each selected concept is displayed as a tag in the form of "concept name (CUI)." The "Reset" button is used to start over a new query by clearing the specified concepts;
4. *Area for displaying health questions related to all the selected concepts* is automatically updated when the user clicks a candidate concept, the "Search" button, or the "Reset" button. By default, all questions are displayed if no concept is selected, consistent with convention.

3.3 Evaluation of Concept Selection

To evaluate the results of concept selection, example-based measures for multi-label classification problems [15, 16] were used as the evaluation metrics to avoid a few questions dominating the values of the metrics. Let R be the reference standard consisting of $m = 50$ questions $\{(q_i, Y_i) \mid i = 1, \ldots, m\}$, where Y_i is the set of most relevant concepts tagged for the question q_i. Let Z_i be the set of predicted concepts for q_i. The example-based *precision* (P),

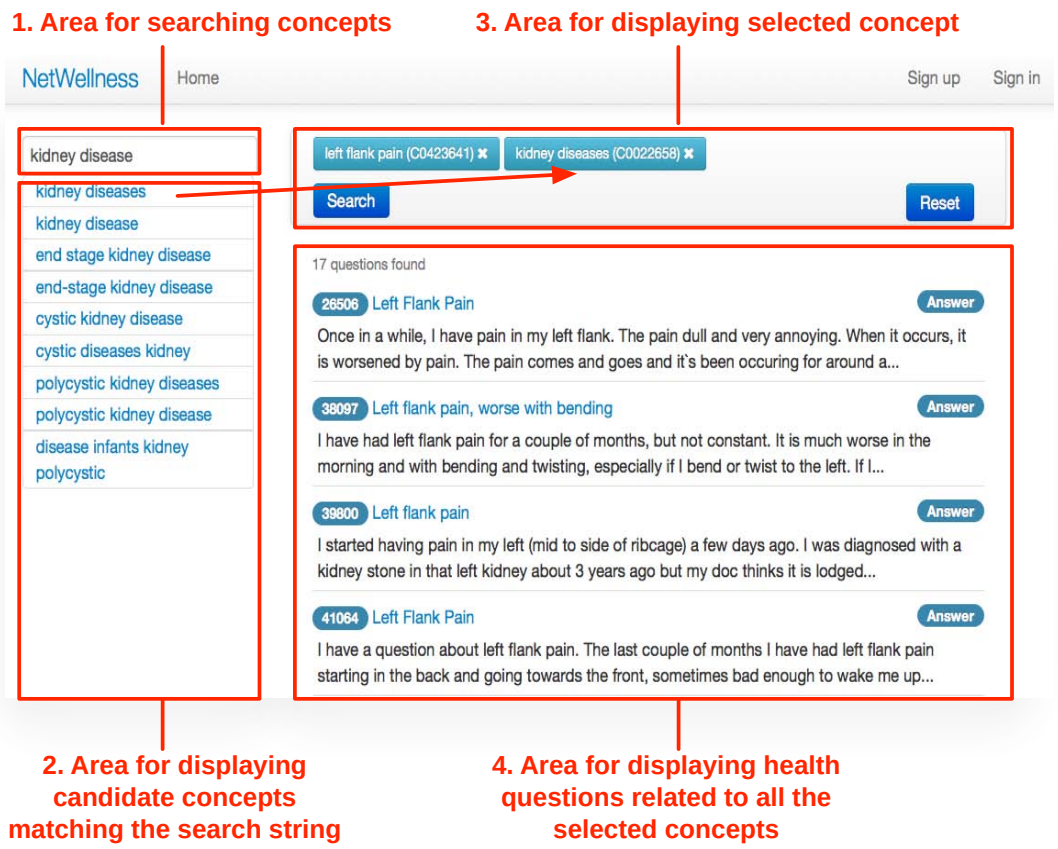


Figure 4: A screenshot of the semantic conjunctive exploratory navigation interface SCENI to retrieve health questions relating to "left flank pain" and "kidney disease." These selected concepts (3. Area for displaying selected concepts) are obtained by typing an unstructured query string into the search box (1. Area for searching concepts) and clicking desired concepts (2. Area for displaying candidate concepts matching the query string).

recall (R) and F_1 *measure* (F_1) were calculated as follows:

$$P = \frac{1}{m}\sum_{i=1}^{m}\frac{|Y_i \cap Z_i|}{|Z_i|},$$

$$R = \frac{1}{m}\sum_{i=1}^{m}\frac{|Y_i \cap Z_i|}{|Y_i|}, \text{and}$$

$$F_1 = \frac{1}{m}\sum_{i=1}^{m}\frac{2|Y_i \cap Z_i|}{|Z_i| + |Y_i|}.$$

Table 5 shows the values of these metrics for the results of three methods: the baseline MetaMap, the TF-IDF method, and the TF-IDF method followed by filtering out concepts not in CHV (TF-IDF + CHV, for short). The TF-IDF + CHV method achieved the best F_1 measure of 0.806, an increase of 40% compared to the baseline MetaMap.

Table 5: The example-based precision, recall, and F_1 measure for the results of concept selection using baseline MetaMap, TF-IDF, and TF-IDF + CHV.

Method	Precision	Recall	F_1
Baseline MetaMap	0.446	1.0	0.576
TF-IDF	0.743	0.932	0.802
TF-IDF + CHV	0.807	0.859	0.806

3.4 Evaluation of Semantic-based Conjunctive Search

To evaluate concept-based search using SCENI, the scores of retrieved questions for search tasks in Table 4 given by the evaluator were used to calculate *precision* and *relevance*. For each search task, precision was calculated by dividing the number of relevant questions (score of 1 or higher) by the number of retrieved questions; relevance was calculated by averaging scores given to each retrieved question:

$$\text{precision} = \frac{\text{\# of relevant questions}}{n}$$

$$\text{relevance} = \frac{\sum_{i=1}^{n} \text{score}_i}{n}$$

where n is the number of retrieved questions. We did not calculate *recall* since it would require the evaluation of each health question in NetWellness for each search task, which is too labor-intensive to perform manually.

Table 6 shows the detailed results of precision and relevance for the two existent search modes (NWO and CENI), and the proposed SCENI. The precision was on average 50.62% for NWO, 77.85% for CENI, and 93.47% for SCENI. The relevance was on average 1.97 for NWO, 3.3 for CENI, and 4.31 for SCENI. This shows the improvement using semantic-based approach.

Table 6: Results of precision and relevance for three search modes of retrieving health questions in NetWellness. NWO: the NetWellness Official website [7], CENI: the Conjunctive Exploratory Navigation Interface in [3], and SCENI: the proposed semantic CENI. Relevance was measured using a slightly modified 5-point Likert scale, where 0 indicated not relevant at all, 1 indicated weakly relevant, and 5 indicated strong relevant.

Query ID	**Precision(%)**			**Relevance (scale 0-5)**		
	NWO	CENI	SCENI	NWO	CENI	SCENI
1	42.86	58.33	87.5	2.14	2.82	4.38
2	22.22	100	100	0.33	5	5
3	33.33	90.91	100	1.11	3.82	3.88
4	0	100	100	0	5	5
5	70	75	88.89	2.5	1.83	2.22
6	100	100	100	4	4.5	4.75
7	33.33	42.86	66.67	1.67	2.14	3.33
8	76	85.71	91.67	2.64	3.79	4.58
9	60	40	100	3	2	5
10	68.42	85.71	100	2.32	2.14	5
Average	50.62	77.85	93.47	1.97	3.3	4.31

4 Discussions

4.1 Performance Analysis

We manually reviewed some health questions and found a number of factors affecting the performance of the proposed semantic-based approach:

1. Spelling errors in health questions. For example, “Larynx” was misspelled as “Larnyx,” and was not identified by MetaMap.
2. Incorrect UMLS concepts identified by MetaMap. Consumers sometimes describe health questions using abbreviations, such as “MG” for “Myasthenia Gravis” and “mg” for “Milligram.” In such cases, MetaMap sometimes incorrectly identify the concept “C2346927: Magnesium Cation” for them.
3. Relatively short description of health questions. Some questions are described in only one or two sentences, which may weaken the advantage of concept selection using TF-IDF and result in less relevant concepts. For instance, for the following question

"If I had a hiatal hernia, would an upper GI series be able to identify it?"

"able" was recognized as the concept "C1299581: Able (finding)" and selected as one of the most relevant concepts due to a small number of concepts identified from such a brief question.

4.2 Limitations

Our evaluation of the semantic-based search SCENI is limited in the number of search tasks performed and the number of evaluators involved in reviewing results. More search tasks and evaluators would avoid bias and provide more statistical power. Since each search task may result in 10s if not 100s of resulting questions, manually evaluating the precision and recall performance is infeasible for larger number of questions. The alternative would be to make the SCENI interface available to the public, with a feedback mechanism built-in the interface. Although this could potentially achieve feedback in scale, it could still introduce biases in feedback responses.

Using UMLS concepts as topics introduces additional challenges. First, concepts with same CUI may have multiple labels which introduces overhead on the interface. On the other hand, terms similar to each other may have distinct CUIs providing an opportunity to apply subsumption reasoning, which we have explored in this work.

5 Conclusion

We have presented a semantic-based approach for exploring consumer health questions using UMLS concepts. SCENI shares the established benefits of CENI without the restriction of a smaller predefined set of topics for supporting menu-driven conjunctive navigation. Our preliminary evaluation shows a slight improvement in precision, and points to the need to alternative forms of systematic, larger scale evaluation.

Acknowledgement. We thank Susan Wentz for providing the consumer health questions from NetWellness. This publication was made possible by the Clinical and Translational Science Collaborative of Cleveland, UL1TR000439 from the National Center for Advancing Translational Sciences (NCATS) component of the US National Institutes of Health and NIH roadmap for Medical Research. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

References

[1] Luo W, Najdawi M. Trust-building measures: a review of consumer health portals. *Communications of the ACM*. 2004;47(1):108-113.

[2] Berland GK, Elliott MN, Morales LS, *et al.* Health information on the internet accessibility, quality, and readability in English and Spanish. *JAMA*. 2001;285(20):2612-2621.

[3] Cui L, Carter R, Zhang GQ. Evaluation of a Novel Conjunctive Exploratory Navigation Interface for Consumer Health Information: A Crowdsourced Comparative Study. *Journal of Medical Internet Research*, 2014;16(2):e45.

[4] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res*. 2004;32:D267-D270.

[5] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *In: Proceedings of the American Medical Informatics Association (AMIA) Symposium*. 2001. p. 17-21.

[6] Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*. 2010;17(3):229-236.

[7] http://www.netwellness.org/. Accessed August 3rd, 2014.

[8] Morris TA, Guard JR, Marine SA, *et al.* Approaching Equity in Consumer Health Information Delivery NetWellness. *Journal of the American Medical Informatics Association*. 1997; 4(1):6-13.

[9] Cui L, Xu R, Luo Z, Wentz S, Scarberry K, Zhang GQ. Multi-topic Assignment for Exploratory Navigation of Consumer Health Information in NetWellness using Formal Concept Analysis. *BMC Medical Informatics and Decision Making*. 2014;14:63.

[10] Jones KS. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*. 1972; 28(1):11-21.

[11] http://www.consumerhealthvocab.org/. Accessed August 3rd, 2014.

[12] Zeng Q, Tse T. Exploring and developing consumer health vocabularies. *Journal of the American Medical Informatics Association*. 2006;13(1):24-29.

[13] http://florianhanke.com/picky/. Accessed August 3rd, 2014.

[14] Kwak M, Leroy G, Martinez JD, Harwell J. Development and evaluation of a biomedical search engine using a predicate-based vector space model. Journal of biomedical informatics. 2013;46(5):929-939.

[15] Tsoumakas G, Katakis I. Multi-label classification: an overview. *International Journal of Data Warehousing and Mining (IJDWM)*. 2007;3(3):1-13.

[16] Tsoumakas G, Katakis I, Vlahavas I. Mining multi-label data. *Data Mining and Knowledge Discovery Handbook*, pp. 667-685. Springer US, 2010.

Semantic Processing to Identify Adverse Drug Event Information from Black Box Warnings

Adam Culbertson M.S., Marcelo Fiszman, M.D., Ph.D., Dongwook Shin, Ph.D., Thomas C. Rindflesch, Ph.D
Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, MD

Abstract

Adverse drug events account for two million combined injuries, hospitalizations, or deaths each year. Furthermore, there are few comprehensive, up-to-date, and free sources of drug information. Clinical decision support systems may significantly mitigate the number of adverse drug events. However, these systems depend on up-to-date, comprehensive, and codified data to serve as input. The DailyMed website, a resource managed by the FDA and NLM, contains all currently approved drugs. We used a semantic natural language processing approach that successfully extracted information for adverse drug events, at-risk conditions, and susceptible populations from black box warning labels on this site. The precision, recall, and F-score were, 94%, 52%, 0.67 for adverse drug events; 80%, 53%, and 0.64 for conditions; and 95%, 44%, 0.61 for populations. Overall performance was 90% precision, 51% recall, and 0.65 F-Score. Information extracted can be stored in a structured format and may support clinical decision support systems.

Introduction

The prescription drug market in the United States is the largest in the world at $266 billion dollars per year in annual sales in 2010 [1]. These lifesaving drugs help save patients' lives and improve the quality of life for millions, but there are trade-offs. Each year there are as many as two million injuries, hospitalizations, and deaths from adverse drug events (ADEs) [2]. In fact, one study has listed ADEs of hospitalized patients as high as the 4th leading cause of deaths in the United States [3]. Data suggest that patients who experience ADEs spend an average of 8.25% longer in the hospital with greater morbidity [4]. In addition to the higher human suffering, these events lead to a substantial financial cost to an already financially stretched healthcare system. The total cost for ADEs in the United States is estimated at $75 billion annually [2]. The cost savings from these unnecessary tragedies could be invested in other potential cost saving technologies, such as health information exchange.

An adverse drug event is defined by the World Health Organization (WHO) as "any untoward medical occurrence that may present during treatment with a pharmaceutical product but which does not necessarily have a causal relationship with this treatment [5]." An ADE could best be described as an unintended negative outcome that occurs as a result of taking a drug. These could range from rash and hives for a mild reaction to severe reactions, such as respiratory complications or death. Information on the nature of the ADEs caused by a drug is essential in preventing negative outcomes. Furthermore, information that surrounds the ADE such as populations and conditions at greater risk, and what to do in case of an event are also of great importance in preventing the events themselves.

Given the magnitude of adverse drug events on our health care system, one proposed approach to improving efficacy and safety is to make better use of drug safety information. Currently, accurate, relevant, and easy to read drug information is difficult to acquire. Much of the essential information is found in the US Food and Drug Administration (FDA) structured product labels [6]. The labels contain details about pharmacology, warnings, precautions, contraindications, and especially, the black box warnings [7]. However, this information is locked in free text, making its use by automated systems impossible. It would be useful to have this data in a structured format, which would allow access, for example, by clinical decision support systems. Several studies have attempted to extract information from structured product labels using both statistical data mining techniques and natural language processing. However, the scope of these methods is limited in that focus is only on extraction of the actual adverse drug events.

In this paper, we test the feasibility of using SemRep (a semantic based natural language processing system) [8] as a general method to extract information from FDA black box warnings labels, to include adverse drug events, conditions at-risk, and susceptible populations. Our preliminary work on this topic can be found here [9].

Background

Black box warnings

Currently, the US Food and Drug Administration defines and detects adverse events for drugs that are currently on the market. The FDA maintains the Adverse Event Reporting System (AERS), which contains drug-use profiles, population databases, and active surveillance systems [10]. In addition to this knowledge source, the FDA and the National Library of Medicine manage the DailyMed Web site, which currently contains 47,385 drug labels stored in an electronic format known as the structured product label (SPL) [7]. This format is the approved standard of the Health Level 7 (HL7) consortium and has been adopted by the FDA for exchanging product and facility information [11]. The labels include a wide range of information on approved drugs: Description of the drug, Clinical pharmacology, Indications and Usage, Contraindications, Warnings, Precautions, Adverse Reactions, Overdosage, Dosage & Administration, Patient Counseling Information, Supplemental Patient Material, Patient Package Insert, Highlights, Full Table of Contents, Medication Guide, and Black Box Warnings (see Figure 1 for a black box warning example). The information contained in the labels is independently vetted by the FDA when the drug is approved and is considered to be the gold standard for drug information. The FDA can force changes to the labels after the drug is approved, if new safety concerns or information warrant such change. Not all drugs contain black box warnings and this section is one of the most stringent sections of the label and is used for drugs that have the potential for serious adverse drugs events.

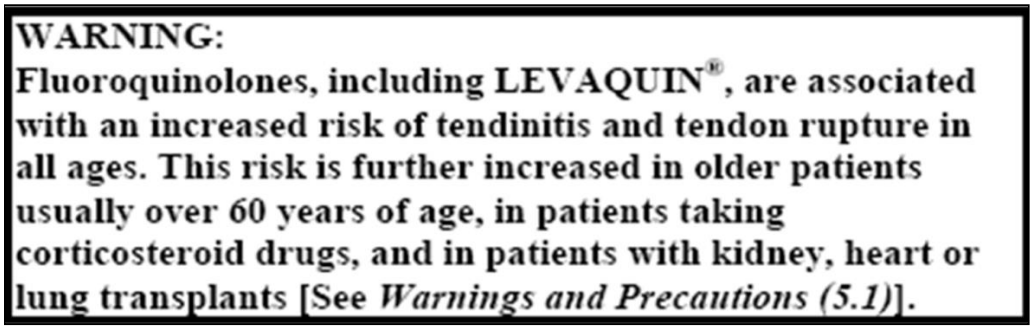


Figure 1. Sample black box warning label for Levaquin® (generic name: levofloxacin.)

Black box warnings include a variety of pertinent information for the prescribing physician. Therefore mining them may have significant clinical impact. Examples of pertinent information include severe adverse drug events, potential susceptible populations such as pregnant women, and conditions that would put a patient at particular risk of adverse events. Given the richness and importance of this information, we chose to analyze black box warning labels.

Automatic methods for identifying adverse drug events

Data mining and natural language processing have been used to extract information about adverse drug events from existing sources of biomedical information. Bates et al. reviewed the research on automatic extraction of adverse drug events. They claim that information technology, especially natural language processing, can be effectively used to identify adverse events in electronic medical records. They suggest that natural language processing will likely become prevalent as new sources of health information continue to become available [12].

Harpaz et al. have explored data mining to discover new adverse drug events[2]. They utilized data-mining analysis and evaluated potential new data sources for ADE discovery such as electronic medical records, administrative claims data, and the biomedical literature. Others have tried knowledge discovery from social media sites. There has been a growing appreciation that social media serves as an innovative method to harness data directly from the consumer. Social media offers great promise as a 'non-traditional source' of data and may serve as an effective tool for the identification of new adverse drug events. One such study by Nikfarin et al. sought to mine the social media Web site DailyStrength for patients self-reporting adverse drug events. They utilized NLP rules-based methods to identify adverse events that may be under reported via traditional methods [13].

Other researchers have focused their attention on SPLs. Rubrichi et al. utilized conditional random fields and compared it to support vector machines for extracting information from SPLs [14]. They sought to automatically extract active ingredient and interaction effects. The work demonstrates the ability of automated methods to successfully pull information from drug labels. Though the paper had promising results, the approach focused only

on the extraction of active ingredient and interaction effects. Fung et al. focused on extracting drug indications from DailyMed [15]. They utilized natural language processing and compared their final results to the national drug file-reference (NDF-RT) and the Semantic MEDLINE Database [16]. The study demonstrates the ability of NLP tools to extract information from SPLs.

Although finding previously unidentified adverse drug events from emerging sources of health data is important, there is still the problem of accessing currently available ADE data from curated sources such as DailyMed. Several research groups have evaluated methodologies to utilize this resource, with a fair amount of success. Bisgin et al. used a topic modeling approach to analyze black box warnings, adverse events, and the warnings and precautions sections of prescription drug labels [17]. They used a statistical hierarchical method. The study produced insight into the connection between various drugs and common adverse drug events such as liver failure and hepatic injury. Another approach was developed by Freidlin and Duke. They created a natural language processing application (SPLICER) to extract adverse events from SPLs. From the output of their system, a standardized ADE knowledge base was created [18]. The system was effective in extracting a total of 534,125 adverse drug events from a total of 5602 product label, with precision of 95% and recall of 93%. SPLICER uses a rule-based and algorithmic approach after adjustments with training data and works in a three step process. It first parses the text, after which the ADE extractor finds patterns likely to contain adverse events. Finally, terms are mapped to the Medical Dictionary for Regulatory Activities (MEDRA). The extracted data is then translated to SNOMED CT codes that can be stored as a value set for clinical decision support. This could lead to better clinical decisions by identifying problematic adverse drug events before the patient receives that drug. The focus of this research is only on adverse drug events; though useful for the clinical decision support, it would be ideal if a knowledge resource contained more general information about the ADE such as at-risk conditions and susceptible populations. Semantic approaches to natural language processing may provide the ability to create a more general approach to extract a broad array of information from structured product labels.

SemRep

SemRep is a semantic natural language system developed at the National Library of Medicine [17]. The system first tokenizes input text and then performs a lexical lookup to the SPECIALIST Lexicon [19]. A parser identifies noun phrases for each sentence, and these are mapped to Unified Medical Language System (UMLS) concepts by MetaMap [20]. SemRep extracts semantic predications which consist of three parts: a subject, an object, and a predicate, which indicates the relationship between the subject and object. For example, from (1) SemRep produces the predication in (2).

(1) Lactic acidosis is a metabolic complication due to metformin.

(2) Metformin CAUSES Acidosis, Lactic.

In addition to CAUSES, SemRep predicates used in this study were PREDISPOSES (identifies risk factors), ADMINISTERED_TO (e.g. “Drug ADMINISTERED_TO PATIENT”), and PROCESS_OF (e.g. “Disease PROCESS_OF Human”). We tested the ability of SemRep to extract information from FDA structured product labels. We focused on black box warnings due to the fact that they are the most serious warning that the FDA issues to drug manufacturers. It is believed that this approach (utilizing semantic predications) can unlock a richer, more generalizable, and deeper understanding of the corpus of drug data than by the many of the previously mentioned methods.

Methods

The DailyMed site housed at the National Library of Medicine provides a rich repository of structured product labels. For this study we only processed the black box warnings of the SPL. Figure 2 gives an overview of the study.

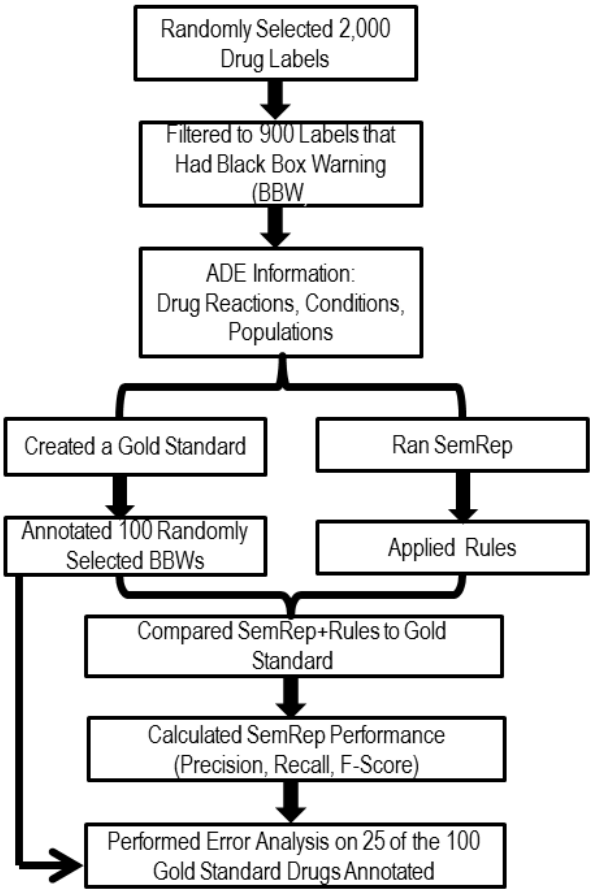


Figure 2. The overall process used to analyze black box drug warning labels with SemRep.

2,000 drug structured product labels were randomly selected from the DailyMed website. Out of those, 900 had black box warnings, which were identified by using the LOINC code 34066-1 (BOXED WARNING SECTION) in the XML text.

We decided to extract three items of information from the free text of the black box warning: the adverse event itself, conditions which put patients at risk for events, and populations at risk for events. These are defined as:

- Adverse event: An adverse event to the drug. (e.g. "Lactic acidosis is a metabolic complication due to metformin.")
- Condition: Disorders which put the patient at greater risk for adverse events. (e.g. "Patients with cardiovascular disease may be at greater risk…"
- Population: Groups at greater risk for an adverse event. (e.g. "Pregnant women should not receive methotrexate.")

Rules applied to SemRep output

Three rules were developed to systematically extract the three items of information from SemRep predications:

1. Adverse event: Object of CAUSES OR Object of PREDISPOSES

2. Condition: Subject of PROCESS_OF

3. Population: (Object of ADMINISTERED_TO OR Object of PROCESS_OF) IF the object Concept is not general, such as "Patient" "Individual," and "Personnel"

1. To identify adverse events, objects of "CAUSES," or objects of "PREDISPOSES" were exploited. An example for the event category is "Lactic acidosis is a metabolic complication due to metformin." The predication extracted from the sentence output was "Metformin CAUSES Acidosis, Lactic." Lactic acidosis, as the object, was retained as an adverse event.

2. To identify conditions, subjects of PROCESS_OF were retained. For example, from the text "Patients with cardiovascular disease may be at greater risk…," SemRep extracted the "predication Cardiovascular Disease PROCESS_OF Patients." Since "Cardiovascular Disease" is the subject, it was retained as a condition for which the drug would put a patient at a greater risk of an adverse event.

3. To identify the population, the object of PROCESS_OF was retained, if it was not "Patient," "Individual," or "Personnel." This was done because these concepts are not specific enough to identify a (useful) population that would be at greater risk for taking the drug. For example, the predication "Acidosis, Lactic PROCESS_OF Pregnant Women" was extracted from the text "Fatal lactic acidosis has been reported in pregnant women", "Pregnant Women," as the object of PROCESS_OF, was retained as an at-risk population.

Evaluation

To build a gold standard, we used 100 randomly selected black box warnings from the 900 extracted from the DailyMed website. The second author (MF) identified at-risk conditions, adverse events, and susceptible populations manually in each of these 100 warnings. We ran SemRep on the same 100 warnings and the rules described above were applied to the output to automatically identify at-risk conditions, susceptible populations, and adverse events. Output from the system (SemRep + rules) was then compared to the gold standard. Performance metrics were calculated for conditions, adverse events, and populations using precision, recall and F-Score.

Results

The results demonstrated that SemRep performed well as a novel tool to identify information from black box warnings (Table 1). Precision was 95% for susceptible populations, with recall 44%, and an F-Score of 0.61. For at-risk conditions, precision was 80%, recall was 53%, and the F-Score was 0.64. For adverse events, precision was 94% with recall of 52%, and F-Score of 0.67.

Table 1. Performance metrics on the ability of semantic processing to extract information from FDA black box warning labels. N = Total number of instances.

	Condition	Population	Adverse event	Overall
Precision	80%	95%	94%	90%
Recall	53%	44%	52%	51%
F-Score	0.64	0.61	0.67	0.65
N	84	83	317	484

Discussion

Results were promising in this study to determine the feasibility of using SemRep and post-processing rules to extract adverse event information (including susceptible populations and at-risk conditions) from black box warnings. Since recall was lower than precision, we performed an error analysis concentrating on the false negatives. Twenty-five random false negatives were analyzed. The most common error type was due to anaphora. Information needed for its resolution often does not appear in the sentence being processed. The following text provides an example.

"*Only physicians experienced in immunosuppressive therapy for organ transplant patients should prescribe Cyclosporine…The drug increased susceptibility to infection and the possible development of lymphoma."*

In order to resolve the sortal anaphoric element *the drug*, SemRep would require access to its antecedent *cyclosporine* in the preceding sentence, which it currently does not have. A future version of SemRep will have the ability to resolve this type of anaphora.

A second common cause of error was that some semantic interpretation depends on inferencing, as in the following sentence:

"There are serious and life-threatening events associated with tamoxifen. Uterine malignancies consist of both endometrial adenocarcinoma and uterine sarcoma."

In order to determine the adverse events associated with tamoxifen, it is necessary to infer that uterine malignancies and endometrial adenocarcinomas are life threatening. This is a challenging problem for natural language processing, which SemRep does not address.

Since precision was lower for at-risk-conditions, we also looked at false positives in this category. The method uses a simple rule to extract conditions-at-risk with the assumption that in the black box warnings the subjects of PROCESS_OF would be mostly these conditions. In spite of 80% precision, there were eleven false positives where the condition mentioned was not the condition-at-risk, but an indication of the drug. The following sentence provides an example.

"Fluvoxamine Maleate Tablets are not approved for use in pediatric patients except for patients with obsessive compulsive disorder"

In this case, the antidepressant fluvoxamine can be used in children with obsessive compulsive disorder. The rule erroneously extracted obsessive compulsive disorder as condition-at-risk.

The ability of a natural language processing system to extract adverse events, but also other types of information, from black box warnings, such as susceptible populations and at-risk conditions is essential for clinical applications such as clinical decision support systems. The information extracted from black box warnings and stored in structured format can be matched against the electronic health record and prevent serious adverse events. For example, if a clinician is considering prescribing an angiotensin-converting-enzyme inhibitor such as enalapril for a patient who is pregnant, information extracted from the black box warning (i.e. Drug: enalapril, Population: pregnant women, Adverse event: fetal death), could provide an alert concerning the serious potential consequences.

Given the number of patients affected by adverse drug events, there is a need to produce general structured data that can be used to provide clinicians, patients, and informatics applications with readily accessible information. It would be particularly valuable to have accurate, up-to-date information for clinical decision support systems available to the clinician at point of care. This could be accomplished by giving the clinical decisions support tool access to the data from drug warning labels that would, for example, help physicians identify patients most at risk for adverse events, especially those with relevant preexisting conditions.

In the future, we would like to expand the information that we extract from black box warnings beyond populations, conditions, and adverse events. Information such as what to do in the event of an adverse event, contra-indications, warnings, precautions, and drug dosage would all be useful expansions to the current research. It is likely that writing rules that apply to SemRep output, and that capture this additional information, will be straightforward. Additionally, we would like to process the entire corpus of prescription drug labels on the DailyMed site. Lastly, we need to modify SemRep to address the linguistic errors identified in the error analysis.

Limitations

This study has limitations. First, the sample size for analysis was rather small with 100 labels evaluated. It remains to be seen how the performance of the system will scale in the larger set of 900 labels that we currently have identified. Second, the gold standard was created by one physician who has experience with SemRep. It would be ideal if physicians not involved in natural language processing system were involved in gold standard creation. Finally, we did not compare our methods to other approaches (i.e. statistical and machine learning), since in this exploratory study our major objective was to determine the feasibility of a rule-based approach.

Conclusions

Having reliable automatic access to drug and prescribing information from FDA package insert labels provides notable benefit to health care. Such access would likely translate into better health care, fewer deaths, and fewer adverse events to prescription drugs. While previous approaches such as topic modeling showed promise to evaluate package insert labels, the additional information available in semantic search technologies will likely underpin significant improvements to mining drug labels. Overall, SemRep was useful as a novel tool to extract information from FDA black box warning labels. The lower recall of the system was largely due to deficiencies in anaphora resolution and the challenging problem of inferencing.

Acknowledgements

This research was supported in part by an appointment to the National Library of Medicine Research Participation Program administered by the Oak Ridge Institute for Science and Education through an inter-agency agreement between the US Department of Energy and the National Library of Medicine.This study was supported in part by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

References

1. United States Commercence Department, Census Bureau. Statistical Abstract of the United States. 131th edition. Census Bureau; 2012.
2. Harpaz R, DuMouchel W, Shah N, Madigan D, Ryan P, Friedman C. Novel. Data-Mining Methodologies for Adverse Drug Event Discovery and Analysis. Clinical Pharmacology & Therapeutics. 2012; Aug;91(6).
3. Lazarou J PBHCPN. Incidence of adverse drug reactions in hospitalized patients: A meta-analysis of prospective studies. JAMA. 1998 April;279(15):1200-5.
4. Rehan HS, Chopra D, Kakkar AK. Physician's guide to pharmacovigilance: terminology and causality assessment. Eur J Intern Med. 2009 January; 20(1):3-8.
5. Nebeker JR, Barach P, Samore MH. Clarifying adverse drug events: a clinician's guide to terminology, documentation, and reporting. Annals of internal medicine. 2004 May;140(10):795-801.
6. United States Food & Drug Administration. Center for Drug Evaluation and Research. Center for Biologics Evaluation and Research. Guidance for Industry Warnings and Precautions, Contraindications, and Boxed Warnings Sections of Labeling for human Prescription Drug and Biological Products-Content and Format. Food & Drug Administration; 2011.
7. National Library of Medicine, DailyMed. 2013. Available from: http://dailymed.nlm.nih.gov/dailymed/about.cfm.
8. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. Journal of Biomedical Informatics. 2003;36(6):462-77.
9. Culbertson A, Fiszman M, Shin D, Rindflesch TC. Semantic processing to identify adverse drug event. Semantic processing to identify adverse drug event information from black box warnings. AMIA Annu Symp Proc. 2013:266.
10. Trontell AE. How the US Food and Drug Administration defines and detects adverse drug events. Current Therapeutic Research. 2001 September;62(9):641-9.
11. U.S. Food & Drug Administration. Structured Product Labeling Resources. 2013. Available from: http://www.fda.gov/ForIndustry/DataStandards/StructuredProductLabeling/default.htm.
12. Bates DW, Evans RS, Murff H, Stetson PD, Pizziferri L, Hripcsak G. Detecting adverse events using information technology. Journal of the American Medical Informatics Association. 2003 October;10(2):115-28.
13. Nikfarjam A, Gonzalez GH, editors. Pattern Mining for Extraction of mentions of Adverse Drug Reactions from User Comments. AMIA Annu Symp Proc. 2011;2011:1019-26
14. Rubrichi S, Quaglini S. Summary of Product Characteristics content extraction for a safe drugs usage. Journal of Biomedical Informatics. 2012 April;45(2):231-9.
15. Fung KW, Jao CS, Demner-Fushman D. Extracting drug indication information from structured product labels using natural language processing. J Am Med Inform Assoc. 2013 20(3):482-8.
16. Kilicoglu H, Fiszman M, Rodriguez A, Shin D, Ripple A, Rindflesch TC, editors. Semantic MEDLINE: a web application for managing the results of PubMed Searches. Proceedings of the third international symposium for semantic mining in biomedicine. 2008:69-76
17. Bisgin H, Liu Z, Fang H, Xu X, Tong W. Mining FDA drug labels using an unsupervised learning technique--topic modeling. BMC bioinformatics. 2011;12 Suppl 10:S11.
18. Duke JD, Friedlin J. ADESSA: A Real-Time Decision Support Service for Delivery of Semantically Coded Adverse Drug Event Data. AMIA Annu Symp Proc. 2010:177-81.
19. McCray AT, Srinivasan S, Browne AC. Lexical methods for managing variation in bio-medical terminologies. Proc Annu Symp Comput Appl Med Care 1994;:235-9.
20. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. AMIA Annu Symp Proc. 2001:17-21.

Information Requirements for Health Information Exchange Supported Communication between Emergency Departments and Poison Control Centers

Mollie R. Cummins, PhD, RN, FAAN, Barbara I. Crouch, PharmD, MPH, Guilherme Del Fiol, MD, PhD, Brenda Mateos, Anusha Muthukutty, MS, Anastasia Wyckoff, MA
University of Utah, Salt Lake City, UT, USA

Abstract

We analyzed audio recordings of telephone calls between emergency departments (EDs) and poison control centers (PCCs) in order to describe the information requirements for health information exchange. Analysis included a random sample of 120 poison exposure cases involving ED-PCC communication that occurred during 2009. We identified 52 information types characterized as patient or provider information, exposure information, ED assessment and treatment/ management, or PCC consultation. These information types constitute a focused subset of information that should be shared in the context of emergency treatment for poison exposure. Up to 60% of the information types identified in the analysis of call recordings can be represented using existing clinical terminology. In order to accomplish standards-based health information exchange between EDs and PCCs using data coded according to a standard clinical terminology system, it is necessary to define appropriate terms, information models and value sets.

Introduction

Poison control centers (PCCs) routinely collaborate with emergency departments (EDs) to provide care for poison exposed patients. Of approximately 2.3 million poison exposures reported to U.S. PCCs in 2012, 27% were managed in a healthcare facility.[1] The PCC and ED share information about the patient, the patient's status, and circumstances surrounding the poison exposure throughout the ED visit. As poison exposure cases evolve, the PCC and ED are in regular communication via telephone for the purpose of collaborative care planning and on-going consultation.[2] The ED care providers verbally share clinical information with a PCC specialist, including patient symptoms, general condition, and the results of certain laboratory tests; the PCC frequently updates treatment recommendations as additional information becomes available. Multiple handoffs can occur for both the PCC and ED.[2] Information may be communicated to one or multiple ED care providers depending on the workload in the ED and the status of the poisoned patient or other patients in the ED. Crucial information documented during this complex process remains isolated within the respective information systems of the ED and PCC and is not shared across organizational boundaries.

In preliminary studies, we found that the current process of telephone-based communication and information sharing between emergency departments and poison control centers is vulnerable to miscommunication, data loss, and error.[2] Examples of safety vulnerabilities include: difficulty establishing synchronous verbal communication via telephone, discussion of multiple patients during the same telephone conversation, and communication with non-clinical staff members.[1] Additionally, any information moved among patient care settings via phone may or may not be documented for continued use by the recipient health care facility. Verbal communication, though expressive, is fraught with pitfalls that affect patient care.[3,4]

The current process of telephone-based communication is also vulnerable in disaster scenarios. PCCs manage calls from *both* public and health care providers, and routinely see increases in call volume due to infectious disease outbreaks and other non-poisoning public health concerns. A process based solely on verbal, telephone communication is fragile. Studies of poison control center operations during call surges are scarce. However, in a previous study, we found that even routinely encountered busy times tax the ability of PCC specialists to communicate.[5] A 2007 article by Vassilev and colleagues describes an incident in which a poison control center experienced compromised operations due to an unexpected 148% increase in call volume.[6] In the setting of mass scale poisonings due to natural disasters, accidents or terrorism, a system based entirely upon telephone communication is fragile.

Given existing technology to support the electronic exchange of data and information across organizational boundaries, there is potential to reduce medical error, reduce time to treatment, and improve continuity of care through a more efficient and structured process of ED-PCC collaboration. In order to design a more robust, structured process for ED-PCC collaboration, supported by health information exchange, it was necessary to determine and characterize the information requirements We analyzed audio recordings of current telephone-based communications between EDs and PCCs in order to describe the information requirements for health information exchange between poison control centers (PCCs) and emergency departments (EDs). Our goal was to identify a focused subset of available health information, most relevant to emergency treatment of poison exposure, in order to support generalizable process re-design. The first research assistant analyzed the transcripts, creating a formative list of information types evident in the communication.

Methods

Setting. We conducted the study at a single poison control center in the intermountain west. The poison control center has a very high utilization rate compared to national data, and so it constitutes a rich source of data describing interactions between PCCs and EDs. In 2011, the poison control center managed 42,544 human poison exposures, with approximately 13% of cases managed in the emergency department setting and entailing collaboration between an emergency department and the poison control center.

Sample. We used a saturation sampling approach, based upon an initial random sample of 500 poison exposure cases, managed in the emergency department setting and involving ED-PCC communication, during the calendar year 2009. As described in Cummins et al (2013), we sequentially analyzed the calls until we failed to identify new information types in 40 additional, randomly selected cases.[2] This saturation sampling approach ensured an adequate sample size to describe the types of information routinely shared in ED-PCC communication, while conserving the resources necessary to analyze calls and minimizing the use of patient data. Ultimately, we analyzed calls corresponding to a sample of 120 poison exposure cases.

Data Preparation and Analysis. As described in Cummins et al (2013), telephone calls between the poison control center and its collaborating EDs are routinely recorded and stored on a secure server, and we previously developed a process for linking cases to analog call recordings.[2,7] We applied this procedure in 20 case increments according to the saturation sampling plan, linking calls to cases as necessary until saturation was reached. For each 20 case increment used in the analysis, we linked call recordings to cases, verified the linkage, exported the call recordings, and converted them to digital format. We transcribed the call recordings and removed identifiers of personnel, patients, and health care facilities. Two research assistants, graduate students in biomedical or nursing informatics, analyzed the data. The first research assistant analyzed the transcripts, creating a formative list of information types evident in the communication. The second research assistant independently reviewed the transcripts, in order to validate the information types. Questions or difficulties in conceptualizing and categorizing information types were resolved through discussion by the research team, and in cases of disagreement, the principal investigator made final determination. Given the lack of pre-determined concepts and categories, this process was formative and iterative. This process resulted in a list of information types and the frequency of their occurrence per poison exposure case.

Representation in Clinical Terminology and Coding Systems. A research assistant manually searched for inclusion of each information type in existing clinical terminology or coding systems, with review and supervision by the principal investigator. These systems included AAPCC (American Association of Poison Control Centers) substance codes, CPT (Current Procedural Terminology), ICD-9 and ICD-10 (International Classification of Diseases, Ninth and Tenth Revisions), NDC (National Drug Code), RxNorm, Poisindex, SNOMED-CT (SNOMED Clinical Terms), LOINC (Logical Observation Identifiers Names and Codes), DEEDS (Data Elements for Emergency Department Systems) 1.0, and NPDS (National Poison Data System).[8,9,10,11] Information types were insufficiently granular for precise mapping at the concept level, but the capability to represent each information type could be assessed. In these cases, the principal investigator reviewed the nature and scope of each terminology system with respect to the information type, and determined whether the system could be used to represent the information type.

Results

Identification of Information Types. We reached sampling saturation upon analysis of calls corresponding to 120 cases. Inter-coder disagreement was infrequent and resolved through team discussion and review of transcripts. The research team reviewed the information types, and where appropriate, aggregated duplicate or nested concepts. After validation and aggregation, we identified 52 information types listed and categorized in table 1. The information content of analyzed calls included essential identifying information – information identifying both

health care provider (location, type of provider, name) and patient (name, age, gender). It also included essential health information about the patient, including current medications, allergies, and health history. Many information types described the poison exposure incident. Information was exchanged about the poison, its characteristics and effects, and clinical treatment. Information was also exchanged about the poisoning scenario, including important circumstances that bear upon decision making related to care of the patient.

Narrative information, the poisoning "story", included details important for discerning whether the poisoning was intentional (overdose or suicide attempt) or unintentional, and details that help to establish the certainty, dose, and timing of the exposure. As a fictional example, a narrative might include a description of the parent's estimate of the number of tablets remaining in a full prescription bottle, and the time at which a child was found eating tablets from that open bottle. Both the PCC and ED collected and shared these types of information. The ED care providers, who assess the patient in person, shared information about the physical exam and appearance of the patient, clinical findings, and the results of any diagnostic testing. They also shared information about the patient's plan of care, and treatment or management. The poison control center, acting as consultant, provided feedback on clinical findings as well as treatment and monitoring recommendations. This frequently entailed general communication about the type of poison, its characteristics, effects, and treatment in a type of communication that appears to establish common ground for the discussion of the specific patient and exposure at hand. Communication also frequently included requests for information from the other party.

Table 1. Types of information shared during telephone calls between EDs and PCCs, excluding patient/ provider identifying information.*

Exposure information	
Exposure Type	E.g., polypharmacy, accidental, overdose.
Certainty of Formulation	In a circumstance where the poison is a therapeutic agent, the level of certainty about the precise formulation, based on the subjective/ objective information provided.
Chronicity	Duration of exposure (acute vs. chronic)
Establishing background/certainty	Narrative information about the poison exposure scenario specifically relevant to determining the certainty of ingestion/exposure, and general information helpful in constructing a clinical picture of the patient and treatment plan.
Substance class	General grouping or characterization of poison (e.g., beta-blocker)
Substance information	General characteristics of a poison (e.g., half-life, mechanism of action, peak effect)
Substance name (generic)	Generic name of poison that is also a therapeutic agent
Substance amount	Dosage or amount of substance (e.g., 500mg metformin)
Substance name (brand)	Brand name of poison that is also a therapeutic agent
Substance description	Informal description ranging from general class of drug (e.g., antiarrhythmic) to intended purpose (e.g., a chemical used to clean carburetors)
Substance form	E.g., tablet, pill, powder, lozenges.
Substance formulation	E.g., extended release vs. Rapid release
Substance identification rationale	Narrative describing reasoning of PCC specialist in identifying a substance based on described characteristics of the poison (e.g., "A blue pill in that shape can be a few different things, but it is probably Viagra")
Substance-nonpharmacological	Name of non-pharmacological substance
Substance type	E.g., solid, inhalant, liquid.
Time since ingestion	Amount of time elapsed since initial poison exposure (e.g., "It has been about 5 hours since she took the pills.")
Patient health history	
Medical history	Information about patient's past medical history.
Patient medications	Medications that the patient currently takes at home.
Subjective & objective Information	
Chief complaint/ reason for	Patient's chief complaint upon presenting to emergency department or

visit	calling poison control center.
Absence of clinical effects	The absence of any signs or symptoms attributable to poison exposure.
Mental status	Information about patient's mental status.
Caller reported symptoms	Symptoms as reported by patient or caller to PCC.
Physical exam findings	Signs observed by ED health care providers.
Unrelated symptoms	Health care provider or PCC assessment that a symptom is unrelated or likely unrelated to the poison exposure. For example, a patient may exhibit tremors related to underlying Parkinson's disease, unrelated to an acute narcotics overdose.
Vital signs	Information describing patient's blood pressure, heart rate, and/or respiratory rate along with the time the vital signs were obtained.
PCC recommendations &toxicology information	
PCC recommendations for treatment and discharge parameters	PCC recommendations for treatment and/or duration of direct observation prior to discharge from the ED.
Toxic dose	Specific dose or amount of exposure, at which a substance becomes toxic.
Toxicity levels	Circulating blood level of a substance considered toxic.
Clinical effects of substance	Potential or expected clinical effects of a given poison exposure.
Worst case scenario	Description of the most harmful clinical effects and poorest outcome that patient might experience (a description of risk), usually PCC to ED.
ED treatment / management information	
Confirmation that was treatment given	ED staff member indicates whether or not a PCC recommended treatment has been administered to a patient.
Patient discharge medications	Medications prescribed for patient at time of discharge from ED.
Patient status	ED description of patient's clinical condition, particularly whether clinical effects of exposure are observed.
Plan of care	Refers to ED health care provider plan of care for patient (e.g., treatments, procedures, length of stay, parameters for discharge).
Diagnostic test results	Inquiry about results of diagnostic tests (e.g., laboratory, ECG, imaging).
Time next laboratory tests will be ordered	Information about the ED's planned or recommended timing of subsequent diagnostic testing (e.g., "We'll get another level at 4 hours [after ingestion].").
Time laboratory test was performed/drawn	Information about the timing of a treatment or a lab test, usually in relation to the time of ingestion (e.g., "That level was drawn 2 hours after ingestion").
Treatment performed	Information that a treatment, directly related to the poison exposure, was administered.
Information types idiosyncratic to telephone communication	
Ambiguous Test Result	Characterization of the results of laboratory or non-laboratory diagnostic testing
Confirmation Patient Arrived at Hospital	Confirmation by an ED staff member that a given patient is under the care of that ED
Request for Chronicity	Inquiry about duration of exposure
Patient Status Request	Inquiry into patient's status, both clinical condition and whether the patient is still under ED care
Request for clinical effects information	PCC inquiry as to clinical effects effects observed by ED health care provider
Request for Lab/Test Results	Inquiry as to results of laboratory diagnostic testing
Request for Test Results	Inquiry as to results of non-laboratory diagnostic testing

*PCCs collect information for entry into the National Poison Data System (NPDS) and consequently, some information types resemble NPDS data elements. However, NPDS data elements are defined differently, as described in a published coding manual.[11]

We compared these information types to commonly accepted clinical terminology and coding systems using the UMLS Metathesaurus, terminology browsers or coding manuals. In this preliminary screening, none of the terminology systems provided complete coverage of the identified information types. NPDS (National Poison Data System) represented 38/52 information types (73%). Clinical terminologies LOINC and SNOMED-CT represented approximately half of the information types. A specialized set of data elements designed for the emergency department setting (DEEDS), represented 31/52 information types (60%). See figure 1 for results. In figure 2, we show the categories of information types along with the coding or terminology systems that best represented the information types within those categories.

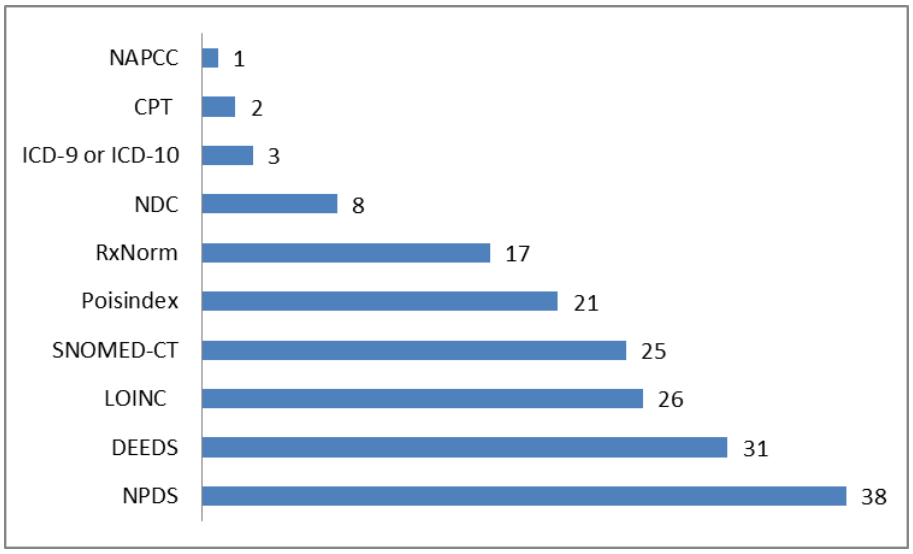


Figure 1. Representation of information types by commonly accepted clinical terminology and coding systems.

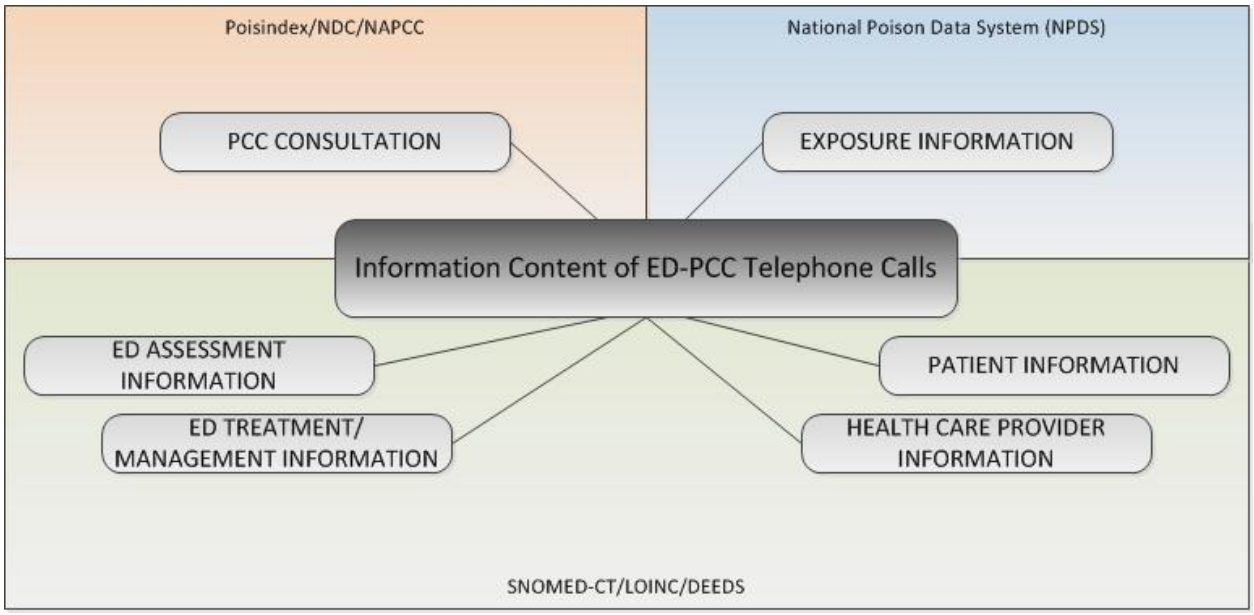


Figure 2. Terminology coverage of information types used in ED-PCC communication.

Discussion and Implications

Approximately 50-60% of the information types identified in the analysis of call recordings map to the widely used clinical terminologies LOINC and SNOMED-CT, and a specialized set of data elements designed for the emergency department setting, DEEDS. However, many of the identified information types are highly specialized to the context of poisoning scenarios and are not found in standard clinical terminology systems. ED-PCC communication about poison exposed patients involves types of information that are not commonly represented in standard clinical terminology systems. In order to accomplish standards-based health information exchange using data coded according to a standard clinical terminology system, additional terms must be proposed and adopted.

The National Poison Data System (NPDS) data elements mapped most successfully to the information types. However, NPDS is structured to support case-level data vs. patient-level data, and is not structured as a clinical terminology. NPDS data elements were designed for the purpose of surveillance and not to support clinical information systems or patient care. Although used by all U.S. poison control centers, NPDS data elements are not currently mapped to any standard clinical terminology system and preliminary attempts at mapping have yielded disappointing results, so they do not facilitate interoperability with emergency departments.[12] Moving toward interoperability, it is important to create a mapping of NPDS data elements to standard clinical terminology systems or expand existing clinical terminologies. It is also necessary to identify standard information models (e.g., CDA) that would be used to instantiate the standard codes, since those codes need to be bound to slots in standard information models. Standard value sets (allowed set of codes for a given data element) should be used in the standard information models, and the NLM (National Library of Medicine) value set authority center could be used as the central resource for submitting and managing these value sets.

Representing Information Content with Existing Clinical Terminology and Coding Systems. One key challenge in designing a health information exchange supported collaboration process, particularly in the context of an emergency poisoning event, is to focus attention on information immediately relevant to the health event and decision making at hand. In the context of ED-PCC collaboration to care for poison exposed patients, we identified 52 information types that should be addressed in the design of health information exchange activities. However, it is currently not possible to represent all of these information types using standard clinical terminologies.

Patient and Health Care Provider Information. Basic information about patients and health care providers (name and gender, for example) is commonly collected in health care settings, and so it is commonly included in standard clinical terminologies and standard clinical information models. Similarly, information about health care providers can be represented using multiple clinical terminology systems and information models.

Emergency Department information. ED assessment information, including the initial history & physical examination and subsequent clinical observations, can be represented using concepts found in DEEDS (Data Elements for Emergency Department Systems)/ LOINC and SNOMED-CT. Similarly, ED treatment or management information can be described using concepts found in DEEDS/LOINC and SNOMED-CT. DEEDS is a terminology system designed to support emergency department operations, and closely fits the type of assessment and observational data exchanged in ED-PCC telephone calls. DEEDS is included in LOINC, a much larger clinical terminology system.

Exposure. Exposure information closely matches NPDS data elements, a logical circumstance given that collection of these data elements is one purpose of the telephone calls. To some extent, these information types can also be represented using DEEDS/ LOINC. However, the coverage of exposure information by DEEDS/LOINC is incomplete or insufficiently granular. This is not surprising, given the highly specialized nature of poison control center operations. PCCs collect information about poison exposure at a much more granular level of detail, a level of detail that supports surveillance and public health intervention. NPDS is case-based, collected for the purpose of surveillance, and NPDS data elements are not structured as a clinical terminology. New information models such as CDA templates would need to be created to represent PCC exposure information.

Substance. Substance information can be represented using several terminology or coding systems. However, the nature and extent of information varies. Poisindex, a proprietary, subscription-based toxicology database, best matches the depth of information shared in telephone calls between EDs and PCCs. Poisindex codes are currently used to identify a very specific type, brand, and formulation of a substance, pharmaceutical or non-pharmaceutical. NAPCC substance codes are also currently used to characterize the type of substance in a meaningful way from the perspective of poisoning epidemiology. Drugs manufactured for therapeutic purposes are often involved in poisoning and if the poison is a manufactured therapeutic agent (Vicodin, for example), it could be

coded using NDC or RxNorm. However, many poisonings involve naturally occurring substances not indexed in NDC or RxNorm (snake venom or toxins from wild mushrooms, for example). Many more poisonings involve inhaled gases (e.g. carbon monoxide), illicit recreational drugs, industrial chemicals, household cleaning agents, and other non-therapeutic chemicals not indexed in NDC or RxNorm. A specific value set for substances could include a subset of RxNorm codes (e.g., ingredients and branded drugs) and a combination of codes from other non-drug terminologies. Additionally, these codes must be used in a way that clearly establishes the context as poison exposure vs. pharmacotherapy.

Limitations

The findings of this study are observational in nature, describing information that is currently shared using telephone communication. In current process re-design efforts, we are obtaining clinician input about additional, potentially helpful information types as well as prioritization of information types. As a single site study, the findings possibly include some information types that are idiosyncratic to the region (intermountain west) or poison control center and emergency departments studied. However, the communication process used by this poison control center is highly similar to the process used by other U.S. poison control centers, and the specific information types found in this study appear to reflect universal aspects of emergency care for poison exposure. Consequently, the findings provide a reasonable basis for standards development and process improvement efforts. Comparison of information content according to ED characteristics was outside the scope of this study. However, the emergency departments served by the poison control center are diverse and include both small, rural hospitals and urban academic medical centers. Additionally, information content may also have varied during surge conditions.[5] The primary purpose of our study was to inventory information types, and the presence of an information type in a single transcript was sufficient for inclusion in the inventory.

Conclusion

Given evidence of inefficiencies and safety vulnerabilities in the current telephone-based process of ED-PCC communication and information sharing, new models of collaboration must be considered. We are currently developing a health information exchange supported collaboration process, designed to support dynamic, real time, bidirectional communication between EDs and PCCs to support care of poison exposed patients. This study identified the subset of health information that should be included in such a process. Moreover, the results indicate that telephone-based communication about poison exposed patients entails the sharing of information types that are not currently represented in standard clinical terminology systems. In order to accomplish standards-based health information exchange between EDs and PCCs using data coded according to a standard clinical terminology system, additional terms must be proposed and adopted. Corresponding information models, such as CDA templates, require development and specific value sets must be identified.

Acknowledgement

This study was supported by the US Department of Health and Human Services, Agency for Healthcare Research and Quality, grant 1R21HS018773-01.

References

1. Mowry JB, Spyker DA, Cantilena LR, Jr., Bailey JE, Ford M. 2012 Annual Report of the American Association of Poison Control Centers' National Poison Data System (NPDS): 30th Annual Report. Clin Toxicol (Phila). 2013 Dec;51(10):949-1229. PubMed PMID: 24359283.
2. Cummins MR, Crouch B, Gesteland P, Wyckoff A, Allen T, Muthukutty A, et al. Inefficiencies and vulnerabilities of telephone-based communication between U. S. poison control centers and emergency departments. Clin Toxicol (Phila). 2013 Jun;51(5):435-43. PubMed PMID: 23697459.
3. Bhasale AL, Miller GC, Reid SE, Britt HC. Analysing potential harm in Australian general practice: an incident-monitoring study. Med J Aust. 1998 Jul 20;169(2):73-6. PubMed PMID: 9700340.
4. Donchin Y, Gopher D, Olin M, Badihi Y, Biesky M, Sprung CL, et al. A look into the nature and causes of human error in the intensive care unit. Qual Saf Health Care. 2003;2003(12):143-7.
5. Caravati EM, Latimer S, Reblin M, Bennett HK, Cummins MR, Crouch BI, et al. High call volume at poison control centers: identification and implications for communication. Clin Toxicol (Phila). 2012 Sep;50(8):781-7. PubMed PMID: 22889059.

6. Vassilev ZP, Kashani J, Ruck B, Hoffman RS, Marcus SM. Poison control center surge capacity during an unusual increase in call volume--results from a natural experiment. Prehosp Disaster Med. 2007 Jan-Feb;22(1):55-8. PubMed PMID: 17484364.
7. Poynton M, Jasti S, Ellington L, Dudley W, Crouch B, Caravati M, et al. Matching waveform audio files with toxicall data: Record linkage in a poison control center. Stud Health Technol Inform. 2006;122:849. PubMed PMID: 17102421.
8. Regenstrief Institute Inc. LOINC. 2013.
9. International Health Standards Development Organization. SNOMED-CT. 2013.
10. Poisindex. Updated periodically. ed. Greenwood Village, Colo: Thomson Reuters (Healthcare) Inc.
11. American Association of Poison Control Centers. National Poison Data System (NPDS)© Reference ManualPart 2 - System Information Manual. 2009.
12. Cummins M, Doing-Harris K, Passman J, Mateos B. Automated mapping of NPDS data elements to the UMLS Metathesaurus. Proceedings of AMIA 2013: American Medical Informatics Association (AMIA) Annual Symposium. 2013.

An Analysis of Medication Adherence of Sooner Health Access Network SoonerCare Choice Patients

Nicholas A. Davis, PhD[1], David C. Kendrick, MD, MPH[1]

[1]University of Oklahoma School of Community Medicine, Tulsa, OK

Abstract

Medication adherence is a desirable but rarely available metric in patient care, providing key insights into patient behavior that has a direct effect on a patient's health. In this research, we determine the medication adherence characteristics of over 46,000 patients enrolled in the Sooner Health Access Network (HAN), based on Medicaid claims data from the Oklahoma Health Care Authority. We introduce a new measure called Specific Medication PDC (smPDC), based on the popular Proportion of Days Covered (PDC) method, using the last fill date for the end date of the measurement duration. The smPDC method is demonstrated by calculating medication adherence across the eligible patient population, for relevant subpopulations over a two-year period spanning 2012 - 2013. We leverage a clinical analytics platform to disseminate adherence measurements to providers. Aggregate results demonstrate that the smPDC method is relevant and indicates potential opportunities for health improvement for certain population segments.

Introduction

Medication adherence generally describes whether patients fill their medication prescriptions at the designated frequency (e.g. monthly). Other related concepts include *compliance* (taking the medication as directed) and *persistence*, which deals with the overall duration of a drug therapy. While these terms are sometimes used interchangeably in the literature, *adherence* is commonly applied when measuring patient use of one or more concurrent prescriptions.

The rate at which patients fill new prescriptions is called *primary adherence*. An acceptable level of medication adherence, considered high adherence, has been generally established at 80% or above[7]. However, some medications, such as antiretrovirals for HIV/AIDS, may require a higher threshold at 90% or greater[17]. Non-adherence occurs whenever patients fail to fill their prescriptions entirely, or delay the act of filling. In recent years, non-adherence is an area of active research and mounting concern for many parties in the healthcare industry, including clinicians, healthcare organizations, and payers[7]. As stated by former US Surgeon General C. Everett Koop, "Drugs don't work in patients who don't take them."

Several studies have been published that demonstrate an association between medication adherence and health outcomes. In cardiovascular disease patients, high adherence to antihypertensive medications has been associated with higher odds of blood pressure control compared with patients having medium or low adherence[10]. Non-adherence to cardiovascular drugs has been linked with increased risk of morbidity and mortality[11]. Low adherence to heart failure drugs has been associated with an increased number of cardiovascular-related emergency department visits[12]. Patients with chronic diseases such as hypertension and diabetes have an alarmingly high rate of non-adherence, 28.4% and 31.4%, respectively[13]. Medication non-adherence is prevalent among patients with diabetes mellitus[14] and psychiatric illness[15]. Among patients with psychoses, the mean rate of adherence is 58%, and among those with depression the mean rate is 65%[15].

With a growing percentage of the population taking prescription drugs, medication adherence has been called "the next frontier in quality improvement"[8]. According to a recent Mayo Clinic study of drug prescriptions in a representative sample of the US population (n=96,953), over 68% of the study participants were taking at least one prescription drug[9]. Additionally, the study found that the percentage taking prescription drugs is rising, from 44% in 1999-2000 to 48% in 2007-2008[9]. Thus, adherence is an issue affecting the majority of US citizens.

In addition to the adverse effect on health outcomes, non-adherence also has a severe economic impact. Several studies have shown the increased costs to the healthcare system for patients with low adherence rates[14,16]. A recent estimate of the impact of non-adherence is $170 billion annually in the US alone[13]. Thus, medication non-adherence is a major impact to our national health outcomes and economy, and clinical interventions to improve adherence are critical.

In this study, we seek to characterize medication adherence behaviors of Oklahoma Medicaid patients whose providers also participate in the Sooner Health Access Network (HAN) by using pharmacy insurance claims data from the Oklahoma Health Care Authority (OHCA), with an aim to understand opportunities across demographic and drug categories. In addition to utilizing existing clinical analytics tools, we have developed an alternative method and corresponding software for calculating adherence. We present the results from initial analyses on the OHCA Medicaid pharmacy records and identify room for improvement across a number of patient demographics. Leveraging this data on local medication adherence rates may enable clinicians to have a positive impact on patient health outcomes.

Background

After a review of the literature on medication adherence, *proportion of days covered* (PDC) arose as the preferred method used to calculate adherence. A closely related and often-used method is *medication possession ratio*, or MPR. One particular improvement with PDC over MPR is its more conservative estimate of adherence. For example, a patient who fills a prescription early during the measurement interval may result in an "extra fill" in the calculation, leading to an overestimated medication adherence. MPR has been criticized for its tendency to overestimate the true rate of adherence. Additionally, PDC correctly accounts for cases where patients switch medications during a calculation interval, and there is overlap between the prior and new medication. PDC has been endorsed by the Pharmacy Quality Alliance (PQA), Centers for Medicare and Medicaid (CMS), National Committee for Quality Assurance (NCQA) and National Quality Forum (NQF)[17,18] as the preferred method for calculating medication adherence.

Whereas MPR sums the days supplied for a medication over the measurement interval, PDC employs the concept of time arrays to reflect the dates encompassed by each prescription fill. For example, a thirty-day supply on June 1st would generate a time array that would cover June 1 - 30. After all time arrays have been created for a particular medication or class, these can then be aggregated and used to determine the number of days that were covered by at least one array. This comprises the numerator of the measure. The denominator of PDC is the number of days between the first prescription fill of the medication and the end of the measurement period. This number is typically expressed as a percentage, so is multiplied by 100 prior to display. Below is the equation used to calculate adherence according to the PDC method:

$$\text{PDC} = \frac{\text{Number of days covered by at least one drug in the class}}{\text{Number of days in measurement period}} \times 100$$

Additionally, when medication adherence is calculated across an entire population, or a segment of a population, the rate may be expressed as a performance measure. In this approach, each individual patient's adherence is formulated using the method described above. The population's or segment's adherence rate is then calculated by counting the number of patients with a PDC greater than or equal to 80%, and then dividing by the total number of eligible patients.

The School of Community Medicine at the University of Oklahoma (OUSCM) Tulsa was established to address the tremendous disparities in health outcomes and healthcare access in Oklahoma. OUSCM has closely tracked the shift in payment models from strictly "fee for service" to the more risked-based shared savings models. OUSCM has committed to supporting a shift in delivery models from individual physicians practicing largely autonomously to patient-centered medical home teams[3], which employ a team of care providers (including clinicians, nurses, social workers, pharmacists and other staff) to maximize the efficacy and efficiency of a patient's treatment. A key strategy in the school's mission is the application of health information technology (HIT) to addressing these issues. HIT can help clinicians to provide timely, equitable, and affordable treatment options.

Crucial in the deployment of HIT is the adoption of electronic health records (EHRs). While EHRs provide compelling advantages on an individual patient basis, they are often not designed to enable an aggregate view of

population health and care team management outcomes. Thus, an often-required parallel component in the deployment of HIT systems is the creation of a data warehouse and analytics tools for storage and analysis of patient EHR data.

Complementary to the OUSCM clinical data available from the EHR is the availability of insurance claims data from the Oklahoma Health Care Authority (OHCA)[1], Oklahoma's Medicaid Agency. In 2009, OHCA converted their largest covered population, SoonerCare Choice members, into a Patient Centered Medical Home delivery model and in 2010 added the Health Access Network program to provide support for the Medicaid providers and their patients transitioning to the PCMH model. The largest of three health access networks, the Sooner Health Access Network (HAN)[2], is operated by a team from OUSCM and offers care management services, PCMH tier advancement support, quality improvement, and health data analytics support to Medicaid providers throughout Oklahoma. The HAN receives claims data feeds for the Medicaid patients attributed to the Medicaid providers enrolled in the HAN. OHCA sends claims data of three types in three different formats: UCE 1500 for professional fees; UB 92 for facility fees (i.e. hospitals and hospital-based ambulatory and lab fees), and character-delimited plaintext files based on the NCPDP format for pharmacy claims. The HAN analytics team processes the claims data onsite. Pharmacy claims data are stored in a SQL Server 2008 database built for this purpose. Updated records for the pharmacy data set are sent monthly from OHCA, and this comprises the data used for this research in medication adherence.

OUSCM utilizes Pentaho, an open source Java-based business intelligence (BI) platform, for its clinical analytics system. Pentaho provides analytics, reporting, and health performance dashboards. Pentaho is a component of an enterprise data warehouse that stores historical data on care management, quality indicators, research studies, and clinical and insurance data sources. The clinical and claims data feeds arrive in their native formats and must be processed in a sequence of steps. Via Pentaho Data Integration (PDI), HAN's business intelligence developers construct intricate data extract, transform, and load (ETL) pipelines using a GUI-based visual design tool. As described in Methods, this environment was augmented to create an ETL workflow and analytics dashboard for medication adherence data.

The following materials and software were used for this research:

- Sooner HAN SoonerCare Choice pharmacy claims data on 133,757 patients, for a total of 3,926,135 individual records. This was filtered to 2012-2013 data and using the algorithm described in Methods.
- RxNorm database version August 5, 2013 Full Update release version
- NPPES National Provider Identifier (NPI) database, July 2013
- R 3.0.2 and RStudio 0.98
- Pentaho 4.5 GA Release, including Analysis Server, Pentaho Data Integration, Schema Workbench
- SQL Server 2008 R2 and Oracle SQL Developer 4.0

Methods

While PDC has been defined concisely as described above, there remains a degree of freedom when employing the algorithm. Specifically, the duration of the measurement period used in the denominator is left to the researcher to determine. Some studies use a period of one year or six months, but this is usually chosen to be appropriate to the particular population segment in question. One challenge with conventional PDC durations is the selection of the measurement period to accurately calculate adherence for a variety of medications and patients. In the literature, studies often calculate medication adherence separately for patients in different segments, such as disease categories. For our purposes, it is desirable to calculate adherence across several segments at once, so that the behavior of the adherence calculation can be studied and its correlation to clinical outcomes evaluated.

Therefore, we set out to calculate adherence *en masse* using a robust approach. We employ the last prescription fill date (the day the patient filled their prescription at a pharmacy) for each medication corresponding to an individual patient. Our approach is a slight modification of the standard PDC method described above that we term Specific Medication PDC, smPDC. While it is true that the patient possesses a full cycle of medication at the last fill date, we have no mechanism, such as a subsequent fill event, to accurately determine whether the patient has indeed taken the medication. A reasonable proxy is the end of a measurement period (i.e. the last prescription filled by a patient occurs prior to the end of measurement), but the assumption required is that the patient has taken the medication.

Thus, to be conservative and avoid an overestimated adherence, we select the last fill date as the end of the measurement period. By utilizing the last fill date, we avoid an unnecessary gap near the end of a static measurement duration (e.g. end of calendar year) that can occur if a patient discontinues a medication. This gap may result in an adherence calculation that underestimates the true value of medication adherence. In addition, the termination of a prescription may be the result of a change in the patient's enrollment status or their PCP leaving the HAN program, but this information is missing from our OHCA pharmacy data set.

A patient may fill a prescription early by a few days. When this occurs, the patient possesses a quantity greater than a single day's supply on the days prior to the end of the original prescription. This would extend the true count of the days covered by a medication by the number of early fill days. The standard PDC algorithm neglects these early fills days in calculating adherence, but there is a modified version that accounts for early fills. Using Specific Medication PDC, medication adherence is calculated as follows:

$$\text{smPDC} = \frac{\text{Number of days covered by at least one drug in the class}}{\text{Number of days between last and first fill}} \times 100$$

An example will serve to illustrate the method. Our patient, Mitch, has a prescription of a renin-angiotensin inhibitor for a cardiovascular condition. Each prescription is a 30 day supply (one pill per day), with refills provided on a monthly basis. Mitch fills the first prescription on March 15, a second on April 18, a third on May 18, and the final on June 23 as shown in Figure 1. To calculate Mitch's adherence to the renin-angiotensin inhibitor, we first determine the duration (representing the denominator in the equation above). Duration is calculated using the number of days between the first and last fill date, in this example between March 15 and June 23, which is 100 days. The numerator is determined based on the number of days covered by the prescription, which is 30 * 3 = 90. Thus, Mitch's adherence to the medication over the duration is 90/100 * 100 = 90%. This can also be calculated in an alternative manner by taking perfect adherence, 100 days, and subtracting the gaps (representing non-adherence) in the prescription fills as illustrated in Figure 1. 100 - 4 - 6 = 90, which coincides with the 90% adherence previously calculated.

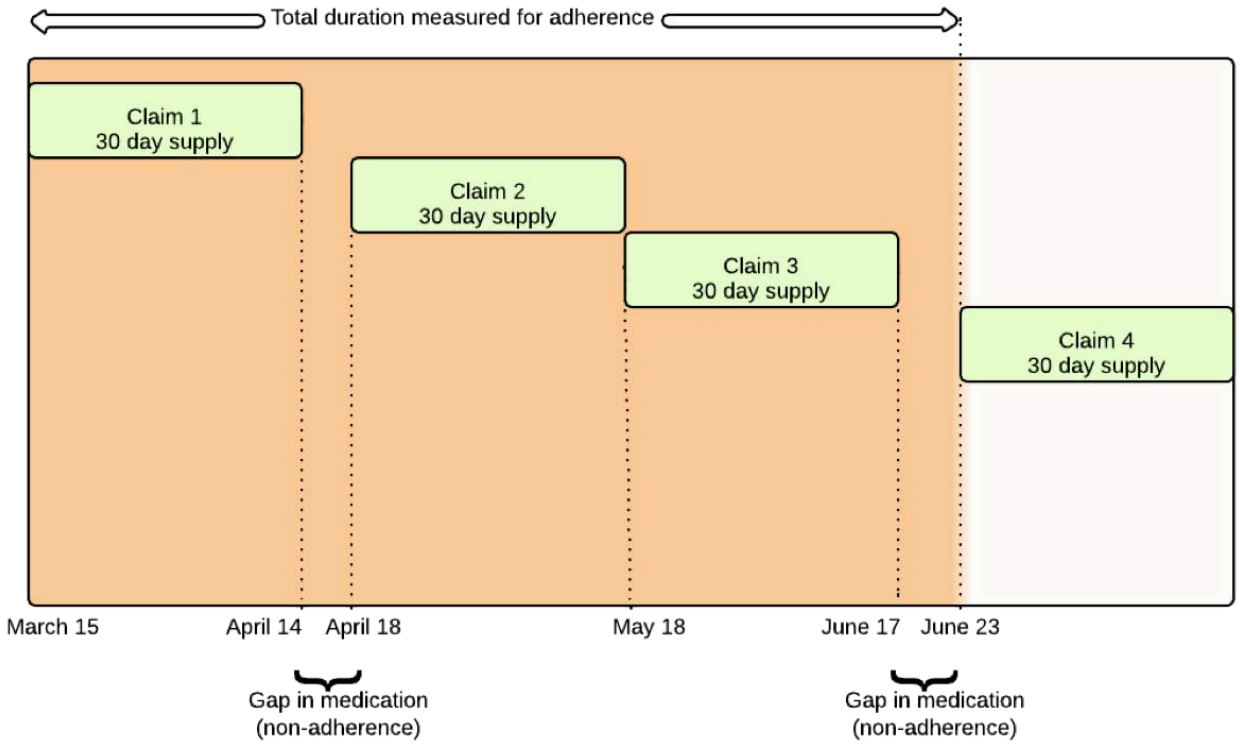


Figure 1. Example medication adherence calculation diagram

A series of filters is applied to the original OHCA pharmacy claims data set. Patients with two or more fills for a particular medication are selected from the entire set of records. A least two prescription fills are required for each patient/medication combination in order to determine the duration between successive prescriptions using the smPDC method. Patients with a single prescription fill of a particular medication are thus excluded from the analysis. Next, records that occur in the 2012 - 2013 timeframe are preserved and the remainder filtered. This two-year study period focuses the analysis on the most current full subset of prescriptions. To be included, the initial prescription fill must occur on or after 1/1/2012 and the last prescription fill on or before 12/31/2013.

Additionally, the claims records received from OHCA included denied claims and reversals. Reversals occur when an original paid claim is later modified or reversed to a denied status. If a claim is denied, it is unlikely that the patient has been issued the medication from the pharmacy (otherwise the pharmacy would be ineligible for reimbursement from Medicaid). Reversals are essentially updates to an existing record, and usually involve correcting clerical errors on a previous claim. For instance, the pharmacist may inadvertently specify the wrong quantity of pills. Reversals are discrete records, thus the final paid claim is used when one or more reversed claims exist for a particular prescription. We exclude both denied claims and reversals from the analysis.

To complement the Medicaid pharmacy claims data set, a couple of external data sources were integrated into the OUSCM enterprise data warehouse. RxNorm[19] from UMLS is used to relate a specific drug code (NDC or National Drug Code[6]) descriptor to an RxNorm concept identifier. RxNorm is further used to identify a VA drug class for each medication, using an approach similar to the algorithm described in Parthak et al[5]. Although each pharmacy record has an existing class provided by OHCA, the VA drug class is more granular in its categories, and is also familiar to clinicians at OUSCM. Provider details are also incorporated from the CMS NPPES National Provider Identifier (NPI) database. The prescribing provider NPI is a data element in the OHCA pharmacy records, and this is used to capture additional details including the first and last name of the provider. Both RxNorm and NPI data sources have been loaded from the provided source files in a SQL Server database and integrated in the analysis.

R is used to perform the initial adherence calculations using the smPDC method. R is an open source statistical framework and language with a multitude of add-on packages and libraries providing external functionality, with the added benefit of being used by a large community of researchers and scientists. R lends itself to trivial parallelism in many cases as well. The pseudocode of the medication adherence calculation is as follows:

1. Select data from SQL database using RODBC package. Denied claims and reversals are excluded, as are patient/medication combinations with fewer than 2 prescriptions. Data elements include patient ID, sex, age, drug code, drug name, drug quantity, days supplied, drug class, provider NPI. Additional data sources are joined and returned in the SQL query, including RxNorm, NPI, and OHCA drug class.
2. Trim the beginning and end of string columns from the raw data (inconsistent extra spaces confound the matching algorithm)
3. Partition the filtered data for parallel processing by patient ID and drug code (data is "embarrassingly parallel")
4. Run parallel algorithm (parallelized for a multicore machine) using mclapply. For each partition of the data:
    a. Calculate medication adherence
        i. Use smPDC as described above
        ii. Exclude durations <= 7 days (filter spurious records with fills within a week of each other for the same patient/medication)
        iii. Exclude first fills before 1/1/2012 and last fills after 12/31/2013
        iv. Set any medication adherence value to 1.0 (100%) if calculated value > 1
    b. Use early fills to extend the number of days covered

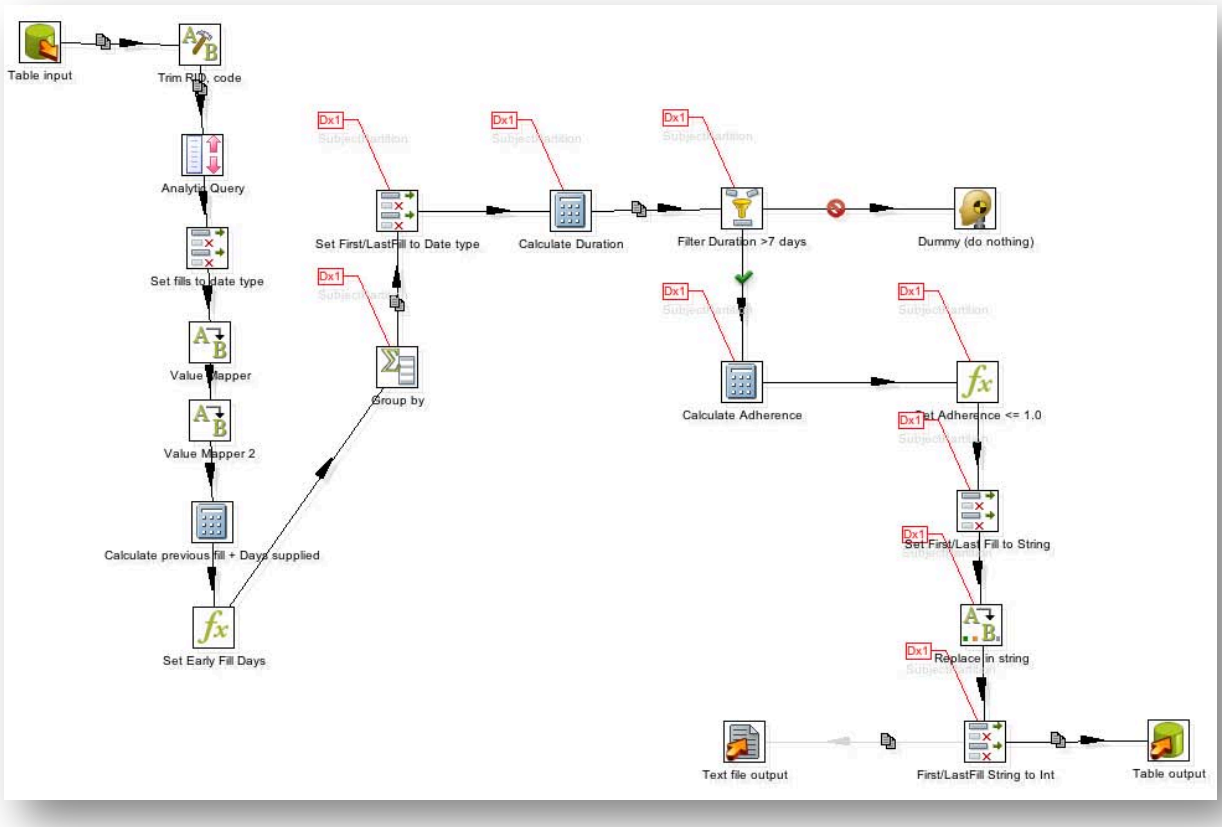


Figure 2. Pentaho Data Integration (PDI) transformation

As OUSCM clinicians, administrators, and staff use Pentaho for clinical reporting and analytics on patient treatment, quality of care, and other areas, we sought next to implement the code in this existing environment. The R code was ported to Pentaho Data Integration (PDI) for selection, filtering, and calculation of adherence results. A PDI transformation was created to perform this ETL step and write the results to a table in a SQL database. Figure 2 shows the resulting transformation used to calculate adherence, and matches the pseudocode previously described. This transformation is subsequently incorporated in a PDI job, which is then scheduled to run on a monthly basis to concur with the availability of the pharmacy claims data from OHCA. Note that this recurring workflow is no longer restricted to the two-year study period used for this research. Thus, historical adherence trends can be observed.

Schema Workbench, a GUI component of the Pentaho suite, is lastly used to create a Data Source that provides an interactive dashboard for medication adherence in Pentaho Analysis Server. The tool connects to the table storing the medication adherence results from the PDI calculations, as well as a number of existing data sources. Relevant data elements are included to allow exploration of the data by a number of attributes. Calculated measures are used to provide a mechanism to determine the mean adherence across a segment of patients or other categories.

Results

OHCA pharmacy claims data records exist from 1994 to the present, received on a monthly basis by the Sooner HAN. While the time span of these records is significant, > 97% of the data is very recent, from 2009 onward. In the unfiltered, original data there are over 3.9 million individual records for approximately 134,000 unique patients. After filtering the denied and reversed claims, as well as patient/medication combinations with fewer than two prescriptions, 2,206,026 records remain. A further filter for the two-year period from 2012 - 2013 produces 961,883 records covering 55,795 patients. This subset of data provides the basis used to calculate medication adherence for all 56K patients using the smPDC method.

Each medication is linked to a corresponding RxNorm concept identifier (RXCUI) using the RxNorm tables. The UMLS RxNorm source files provide the schema and data for a series of nine tables. RXNSAT (Simple Concept and Atom Attributes) is used to match a National Drug Code (NDC) from the OHCA pharmacy claims data to a corresponding RxNorm Concept Unique Identifier (RXCUI). A matching concept is identified for 97.8% (940,882) of the records in the filtered data.

Based on the identified RXCUI, we then match a drug with its VA drug class name. Each pharmacy record arrives encoded with a drug class code, assigned by OHCA, represented as an integer from 0 - 99, for a total of 100 categories. These codes are mapped to a table with class names, provided from OHCA as a data dictionary. The OHCA drug classes are assigned from First Data Bank (FDB)[4]. There are 411 distinct VA drug class categories in the RXNSAT table, providing a finer classification granularity for medications. For the matching algorithm, three tables are used: RXNSAT, RXNCONSO (Concept Names and Sources), and RXNREL (Related Concepts). Of the filtered data set, 97.2% (935,426) of the records map to a corresponding VA drug class.

Provider details are incorporated to allow mapping to prescribing physicians. Each pharmacy record contains the prescribing provider NPI, which is leveraged to identify the physician's first and last name. Nearly all records (>99.9%, 961,344) have a value for NPI. In the source table for this data, NPI, several rows are missing values for provider first and last name. Despite that, 99.7% (959,035) of records from the filtered data set have matching provider names when joined from the NPI table based on the prescribing provider NPI.

From the input filtered data set (~962K records/56K patients), the algorithm further filters out prescriptions of the same patient/medication with a duration less than or equal to seven days. With this filter applied in the medication adherence calculation, the patient count is 46,568. The adherence calculations are written to an Adherence table that includes patient ID, drug code, first and last prescription fill dates, duration, OHCA and VA drug class names (two distinct columns), provider NPI and first and last names, and of course the calculated adherence score. Adherence has a total of 138,686 rows, with each row representing a unique patient/medication combination.

Table 1. Medication adherence calculated with smPDC by segment, 2012 - 2013

Segment	Sample size (n)	Mean adherence
All	46,568	56.08%
Sex		
Female	28,364	56.33%
Male	18,197	55.64%
Unknown	7	56.47%
Age		
< 18 years old	27,602	48.23%
18 - 25	5,573	54.29%
26 - 35	5,791	60.33%
36 - 45	3,265	65.24%
46 - 64	4,302	68.36%
65+	36	73.56%
Race		
American Indian/Alaskan Native	3,428	57.06%
Asian	526	50.84%
African American	9,766	53.04%
Native Hawaiian/Pacific Islander	96	52.11%
White	29,731	57.40%
Drug class		
Psychostimulants-Antidepressants	7,439 (12,728)	61.61%
Narcotic analgesics	8,246 (13,396)	54.19%
Glucocorticoids	5,833 (7,305)	51.21%
Antihypertensives	3,253 (4,409)	61.79%
Anticonvulsants	3,733 (5,883)	63.59%

Table 1 shows the results of the calculations. The mean adherence for each segment is presented in the far right column, with the sample size in the second column and the segment in the first. Adherence results are calculated on segments by sex, race, age, and a few of the top drug classes as ranked by quantity of prescriptions. Overall, mean adherence is ~56% for the filtered data set. In the second column, under drug class segments, the number in parentheses indicates the total number of patient/drug combinations for each drug class. Each patient/drug combination will have at least two prescriptions, as this is required to calculate medication adherence using smPDC.

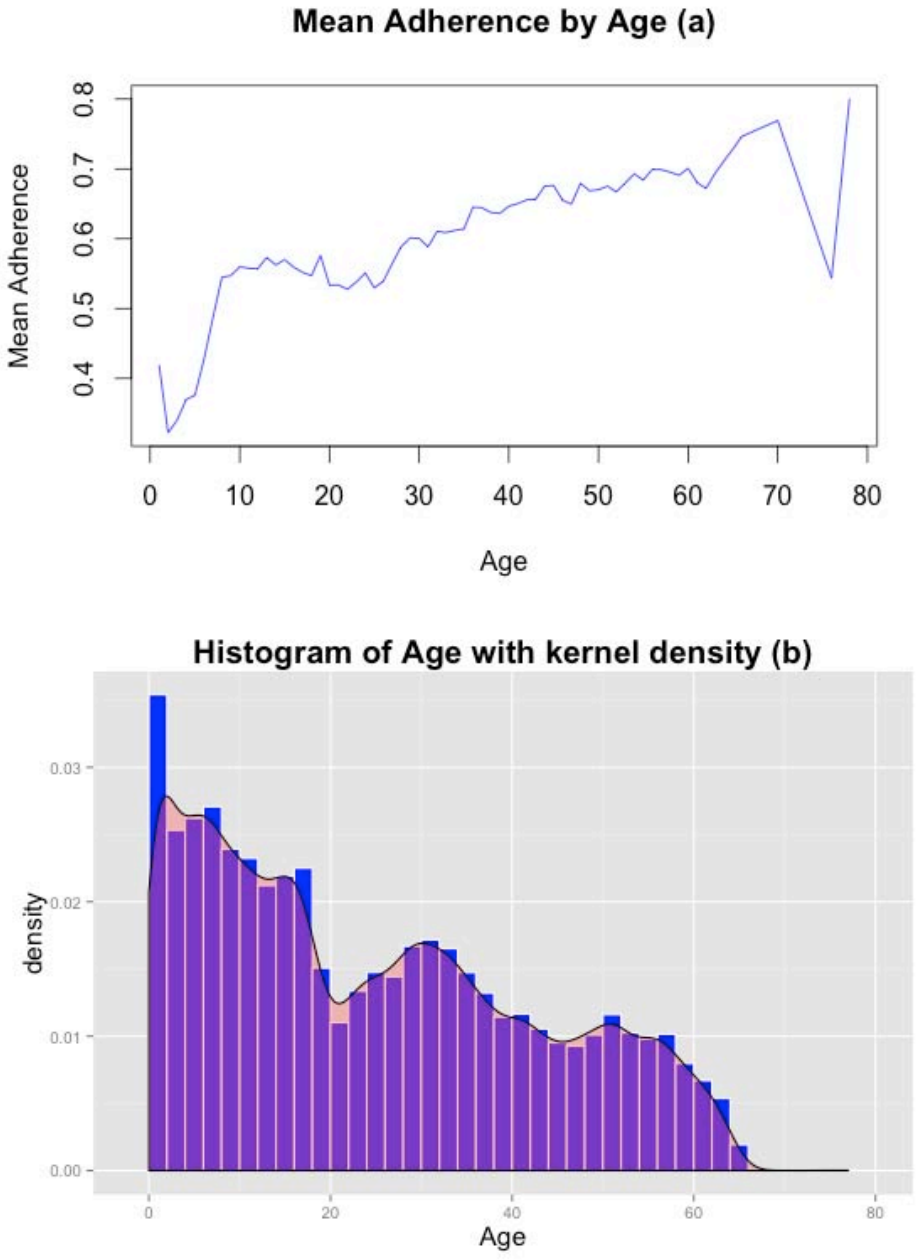


Figure 3. Mean adherence by age (a) and histogram and kernel density curve of age (b)

To capture the trend of adherence rates for various ages, the mean adherences are calculated for each discrete age in the filtered data set. Figure 3a displays a graph of mean adherence by age. Mean adherence generally increases with age, with a slight dip in the early- to mid-twenties. A local minimum is reached at age 75, with a sharp

increase to the global maximum of 80% adherence at age 77. Figure 3b shows a histogram of the population by age, using bins of 2 years. The largest segment of the population by age is clearly 0-2 years, with a local maximum at 6-8 years. After age 18, a sharp decline is seen in the number of patients, followed by additional local maxima at 30-32 years and 50-52 years.

Discussion

With an overall mean medication adherence of ~56%, the results illustrate that adherence rates can be calculated using these methods, and that there may be opportunity for improvement in adherence rates across all populations. Across genders, the adherence rates were essentially the same for males and females (~.7% greater adherence for females). As seen in both Table 1 and Figure 3, medication adherence increases with age across the population. The youngest age groups 0-2, 2-4, exhibit the lowest adherence rates out of any of the segments measured in this study (42% or lower). Among the drug classes, patients with high blood pressure and epileptic seizures display the highest adherence (61.79% and 63.59% for Antihypertensives and Anticonvulsants, respectively). The lowest adherence among the most frequently prescribed drug classes is seen in the Glucocorticoids (51.21%), used in inflammatory disorders such as asthma, and not generally intended for long term use.

There are a few ways to account for patient age in this study. Date of birth is the actual value present in the original data, thus the age is a calculated value. Originally age was calculated based on the age of the patient on the prescription fill date. This presented a dilemma, as patients with multiple prescriptions over the two-year study period would naturally have a birthday (and perhaps two) during the measurement period. Patients with multiple ages confound the calculation of mean adherence across age-based segments. Thus, a decision was made to set the value of the age of a patient to individual's age on 1/1/2013, or the midpoint of the study period.

Another challenge present in the data occurred with the labeling of sex for individuals. While greater than 99.99% of the patients had a consistent value set for each prescription record, a small number (7) of patients had sex codes for both male and female. It is unclear if this represents legitimate cases or is a clerical error. Nonetheless, these individuals were placed into the Unknown category shown in Table 1.

In this study, all drug classes were taken into consideration for the analysis. Future research will further characterize medication adherence for chronic illnesses, such as diabetes and cardiovascular disease. Additionally, some medications are prescribed as PRN, or take as needed. There are no indicators within our data set of whether a medication is prescribed as PRN. However, medications from certain drug classes are often prescribed as PRN, which could be used as a proxy for PRN medications. It is also possible that the rate of chronic drugs vs. PRN medications increases with age as well (not explored in the current analysis). While some of the code and queries used in this research are specific to the HAN Medicaid data set, the tools are generalizable and may be used to calculate medication adherence in any similarly structured pharmacy claims data set.

While the present results are focused on the two-year measurement period, medication adherence is calculated for all years and accessible to authorized OUSCM clinical staff and researchers via the Pentaho web-based portal. In the Medication Adherence environment, users can explore adherence by individual patient, provider (e.g. a clinician can view his or her patient's adherence results), drug class, and a number of additional dimensions. Aggregate information is also available for researchers interested in population health statistics.

Conclusion

We have presented the results of the medication adherence rates for a subset of Oklahoma Medicaid patients enrolled in the Sooner Health Access Network. Prior to this work, our community medication adherence rates remained elusive. Armed with this information, health care managers can now begin the task of working with clinicians to improve these rates across a broad range of population segments, beginning with the most critically non-adherent patients. Future initiatives of our research include selecting specific chronic conditions or disease categories for study and working closely with clinicians and clinical teams to generate periodic reports for patient teams. Ultimately, the goal is for medication adherence rates to appear in each patient profile, which will require integration of adherence calculations in the existing EMR user interface. This will allow providers to determine which patients exhibit low medication adherence during patient encounters and thus have a positive impact on patient health outcomes.

References

1. Welcome To The Oklahoma Health Care Authority. http://www.okhca.org. Accessed August 16, 2013.
2. Sooner Health Access Network: The University of Oklahoma. http: //soonerhan.ouhsc.edu. Accessed September 4, 2013.
3. Patient-centered medical home. http://www.ncqa.org/Programs/Recognition/PatientCenteredMedicalHomePCMH.aspx. Accessed November 6, 2013.
4. First DataBank. http://www.firstdatabank.com. Accessed February 20, 2014.
5. Pathak, J., Murphy, S. P., Willaert, B. N., Kremers, H. M., Yawn, B. P., Rocca, W. A., & Chute, C. G. (2011). Using RxNorm and NDF-RT to classify medication data extracted from electronic health records: experiences from the Rochester Epidemiology Project. *AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium*, *2011*, 1089–98.
6. Drug Approvals and Databases > National Drug Code Directory. http://www.fda.gov/drugs/informationondrugs/ucm142438.htm. Accessed March 4, 2014.
7. Ho, P. M., Bryson, C. L., & Rumsfeld, J. S. (2009). Medication adherence: its importance in cardiovascular outcomes. *Circulation*, *119*(23), 3028–35. doi:10.1161/CIRCULATIONAHA.108.768986
8. Heidenreich PA. Patient adherence: the next frontier in quality improvement. Am J Med. 2004; 117: 130–132.
9. Zhong, W., Maradit-Kremers, H., St. Sauver, J. L., Yawn, B. P., Ebbert, J. O., Roger, V. L., ... Rocca, W. A. (2013, July 1). Age and sex patterns of drug prescribing in a defined American population. *Mayo Clinic Proceedings. Mayo Clinic*. Mayo Foundation for Medical Education and Research.
10. Bramley TJ, Gerbino PP, Nightengale BS, Frech-Tamas F. Relationship of blood pressure control to adherence with antihypertensive monotherapy in 13 managed care organizations. J Manag Care Pharm. 2006; 12: 239–245.
11. Rasmussen JN, Chong A, Alter DA. Relationship between adherence to evidence-based pharmacotherapy and long-term mortality after acute myocardial infarction. JAMA. 2007; 297: 177–186.
12. Hope CJ, Wu J, Tu W, Young J, Murray MD. Association of medication adherence, knowledge, and skills with emergency department visits by adults 50 years or older with congestive heart failure. Am J Health Syst Pharm. 2004; 61: 2043–2049.
13. Fischer, M. A., Stedman, M. R., Lii, J., Vogeli, C., Shrank, W. H., Brookhart, M. A., & Weissman, J. S. (2010). Primary medication non-adherence: analysis of 195,930 electronic prescriptions. *Journal of General Internal Medicine*, *25*(4), 284–90. doi:10.1007/s11606-010-1253-9
14. Ho, P., JS, R., FA, M., & al, et. (2006). Effect of medication nonadherence on hospitalization and mortality among patients with diabetes mellitus. *Archives of Internal Medicine*, *166*(17), 1836–1841
15. Osterberg, L., & Blaschke, T. (2005). Adherence to medication. *The New England Journal of Medicine*, *353*(5), 487–97. doi:10.1056/NEJMra050100
16. Sokol MC, McGuigan KA, Verbrugge RR, Epstein RS. Impact of medication adherence on hospitalization risk and healthcare cost. Med Care. 2005;43:521–30. doi: 10.1097/01.mlr.0000163641.86870.af.
17. Nau, D. P. Proportion of Days Covered (PDC) as a Preferred Method of Measuring Medication Adherence. http://ep.yimg.com/ty/cdn/epill/pdcmpr.pdf. Accessed September 03, 2013.
18. Nau, D. P. The Importance of Measuring Adherence. https://www.urac.org/wp-content/uploads/Importance-of-Measuring-Adherence.pdf. Accessed January 09, 2014.
19. RxNorm Technical Documentation. https://www.nlm.nih.gov/research/umls/rxnorm/docs/2014/rxnorm_doco_full_2014-1.html. Accessed January 27, 2014.

Sophia: A Expedient UMLS Concept Extraction Annotator

Guy Divita, MS, Qing T Zeng, PhD, Adi V. Gundlapalli, MD, PhD, MS
Scott Duvall, PhD, Jonathan Nebeker, MD, Matthew H. Samore, MD, PhD
VA Salt Lake City Health Care System and University of Utah School of Medicine,
Salt Lake City, UT

Abstract

An opportunity exists for meaningful concept extraction and indexing from large corpora of clinical notes in the Veterans Affairs (VA) electronic medical record. Currently available tools such as MetaMap, cTAKES and HITex do not scale up to address this big data need. Sophia, a rapid UMLS concept extraction annotator was developed to fulfill a mandate and address extraction where high throughput is needed while preserving performance. We report on the development, testing and benchmarking of Sophia against MetaMap and cTAKEs. Sophia demonstrated improved performance on recall as compared to cTAKES and MetaMap (0.71 vs 0.66 and 0.38). The overall f-score was similar to cTAKES and an improvement over MetaMap (0.53 vs 0.57 and 0.43). With regard to speed of processing records, we noted Sophia to be several fold faster than cTAKES and the scaled-out MetaMap service. Sophia offers a viable alternative for high-throughput information extraction tasks.

Introduction

There is a pressing need for clinical concept extraction and concept indexing to unlock currently obscured information from large corpora holding clinical narratives. Natural language processing (NLP) tools such as MetaMap[1], cTAKES[2] and HITex[3] have traditionally been used for concept extraction and have performed well in the clinical domain. However, these have not been scaled-up to handle big data while preserving processing speeds. Large health care systems such as Kaiser Permanente, Mayo, Vanderbilt and the US Department of Veterans Affairs (VA) would have need for scaling up their concept (information) extraction tasks. As an example, the VA maintains a fast growing corpora of 2.6 billion clinical notes through a secure research environment (Veterans Informatics and Computing Infrastructure, VINCI[4]). Using currently available tools running on several multi-core servers, we estimated that it would take multiple years to create concept indexes for the notes available in VINCI notes to facilitate further information extraction and retrieval.

For clinical, health services and genomic research, there is a critical and ongoing need for NLP tools to mine the free text of medical records to supplement structured data queries to identify patient cohorts and phenotypes. In developing these tools, researchers consider several criteria: usability, maintenance, efficacy (in terms of recall/precision/f-score), ability to incorporate and use local lexica (or terminology), high throughput performance and adoption within the NLP community. While no currently available tools satisfy all criteria, we set out to develop a tool which would be useful for high throughput while maintaining efficacy.

We report the development of Sophia which is a UIMA-AS[5] based UMLS[6] concept extraction annotator. Sophia is now a key component of the v3NLP Framework used by VINCI for information extraction tasks. Sophia shares some methodologies found in MetaMap and cTAKES, but includes some attributes that cTAKES does not, and also excludes some functionality that MetaMap has. More importantly, Sophia is designed for fast processing, while most prior efforts emphasize extraction accuracy.

State of the Art in Extracting UMLS Concepts

There are a number of open source NLP tools and techniques specifically developed to extract UMLS concepts from clinical text. Among them, cTAKES and HITex are well represented in the field. MetaMap and SAPHIRE[7] were tools initially designed for UMLS concept extraction within the bio-literature domain that have been adapted for use with clinical text by several organizations. There are a number of non-open source successful efforts to extract UMLS concepts within clinical text include MedLEE[8], MedKAT[9] and KnowledgeMap[10].

Many of these efforts are built upon two frameworks adopted or developed for use within the NLP field: GATE[11] and Apache-UIMA. A relevant component common to these two frameworks is the notion of a pipeline

composed of a sequence, or end-to-end chaining, of atomic modules often referred to as annotators. An annotator adds stand-off highlights, mark-ups, labels or annotations related to the original text. The annotations from one module are used as input to a downstream annotator. A phrase chunker annotator, for example, depends upon part-of-speech annotations added via an upstream annotator in a pipeline. Efforts built upon the UIMA platform have the potential for being scaled out through the replication of pipeline instances via a related framework: Apache UIMA's Asynchronous Scale-out[12] (UIMA-AS). The UIMA-AS framework provides for pipeline component replication in addition to the full pipeline instance replication to address bottleneck annotators.

Methods

Sophia Annotator Defined

The Sophia annotator identifies UMLS Concepts using a lookup algorithm to match longest spanning matches to an index of known UMLS concepts. A conscious decision was made to find longest spanning matches rather than shortest spanning or by including all possible matches. Longest spanning matches reduce the ambiguity issue by finding the most specific match, for instance finding *chest pain* rather than *chest* and *pain*. While including the constituent components such as pain and chest might be useful for building google-like search indexes to aid retrieval techniques, the first iteration of Sophia does not include this capability because such a capability was not part of the motivating use cases.

The lookup algorithm relies on exact match retrieval to keys in the index, rather than uninflected or stemmed key retrieval. An exact match retrieval looks up the words as they appear in the sentence to find keys in an index. Within the sentence the *patients were transferred*, exact match retrieval would look up the words patients, were, and transferred within a dictionary. Within an uninflected lookup algorithm, each of the words within the sentence would be transformed into the uninflected keys: patient is transfer. These uninflected keys are what would be looked up within an index that holds the uninflected forms. The index includes all possible fruitful variants[13] for a given UMLS concept to insure that valid matches will be found. Fruitful variants for a given term includes spelling variants, inflections, synonyms, acronyms and abbreviations, acronym and abbreviation expansions, derivations and combinations of these transformations such as the spelling variants of synonyms. The burden of computation to make a match is shifted from the cost of normalizing words in the text to be looked up, to having a larger index where the variant expansion cost was taken up at index creation time. An early MetaMap paper[14] showed that this technique increases match precision or accuracy over stemming normalization techniques.

The lookup algorithm works on a window that initially includes all the tokens of a sentence as the longest span to find. Subsequent lookups drop successive tokens from the beginning of the sentence until a match is made. The algorithm does not rely on phrasal boundaries on the grounds that there are important UMLS terms that include multiple phrases, particularly those that include multiple prepositional phrases (*of*, *with*, *without*, *with/without*). Techniques that rely on phrasal barriers to determine the window size sometimes miss the longer, less ambiguous, more specific matches. Both MetaMap and cTAKES have post phrase identification to re-join specific kinds of prepositional phrases to the adjoining noun phrases to partially ameliorate this condition.

No phrasal boundaries are necessary for Sophia's lookup technique. As a consequence, no part-of-speech tagger is necessary to identify phrasal markers, eliminating two common up-stream annotators commonly found in other concept extraction systems.

The Sophia lookup algorithm evolved from the SPECIALIST Text Tools[15]. The SPECIALIST Text Tools lookup dropped tokens from the beginning side of the sentence where-as the Sophia algorithm drops tokens from the ending side of the sentence. While neither version is perfect, dropping tokens from the ending side of the sentence favors having the head of a term as the last token matched. For example, the prior version would have matched *heavy chain* and *smoking* from the sequence *heavy chain smoking* whereas the current version would match *heavy* and *chain smoking*, given the situation where the index includes *heavy chain*, *chain smoking*, and *heavy* as keys (and not *heavy chain smoking*).

The index entry key creation is important to the overall Sophia scheme. Each UMLS string has a set of lexical variants generated to create keys in the Sophia index. These lexical variants include spelling variants, inflections, un-inflections, synonymy, derivations, acronym or abbreviation expansions and acronyms and abbreviations. Fruitful combinations of each of the above mentioned variants are also generated including derivations of spelling variants, derivations of synonyms, and derivations of derivations. These variants are generated from a configuration of the LVG tool[16] distributed by the National Library of Medicine called the

fruitful variants flow. This tool over-generates variants for Sophia's purposes. Variants that are generated from terms that, themselves are acronym/abbreviation are most likely fallacious. For instance, generating spelling variants to the acronym *A.I.D.S.* generates the term *AIDS*, applying (un) inflections to that term creates *AID* (already fallacious), and applying either inflectional or derivational suffixes such as *ing* will lead to additional fallacious terms including *aiding*.

A post-processing filter is applied to prune out any variant generation combination that includes acronym/abbreviation or acronym/abbreviation expansion plus any additional mutation. Long sequences of synonyms or derivations are likewise pruned out.

The Sophia Pipeline

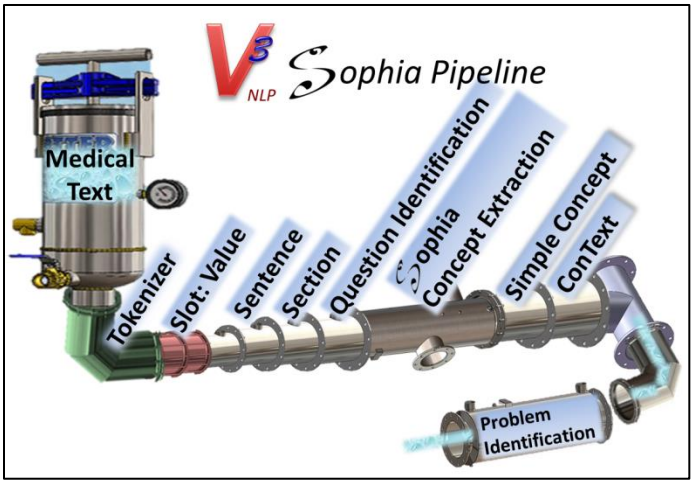


Figure 1. v3NLP Sophia pipeline

To evaluate Sophia against MetaMap and cTAKES, we created a pipeline that includes the prerequisite upstream annotators, the Sophia annotator, and needed downstream annotators (the Sophia Pipeline). The pipeline is built upon the UIMA-AS framework and is expected to be configured to utilize the resources of well-endowed production hardware to provide big-data concept extraction across the Veteran Administration's 2.6 billion clinical records.

The Sophia annotator works on the tokens within sentence or utterance boundaries. As such, the Sophia annotator requires upstream token and sentence annotators to identify words and sentences. Sophia was developed with the v3NLP framework, which includes token and sentence boundary annotators. V3NLP's sentence boundary annotator takes advantage of an annotator that identifies *slot:value* structures (they have a special kind of sentence grammar), and an annotator that identifies section headings. Since many clinical notes include question and checkbox boiler-plated sections, an annotator was added to recognize questions and their related checkbox structures to correctly handle concept assertions within these entities. The Sophia Annotator will blindly find UMLS concept mentions. The conText[17] assertion annotator is run downstream of the Sophia annotator to provide assertion attributes to the concepts found. Figure 1 shows a skeuomorphic representation of the Sophia pipeline with a medical problems identification annotator at the tail end of the pipeline. The problems identification component was added here to extract medical problems from clinical text as an extrinsic evaluation.

Evaluation

The Sophia Pipeline, MetaMap Pipeline and cTAKES were evaluated in an extrinsic task to identify medically significant problems mentioned in clinical text. The evaluation includes a span comparison compared to a human reference set and the throughput performance, i.e., how many records per second were processed. This paper provides the basis for baseline efficacy and performance metrics of the software devoid of the deployment environment.

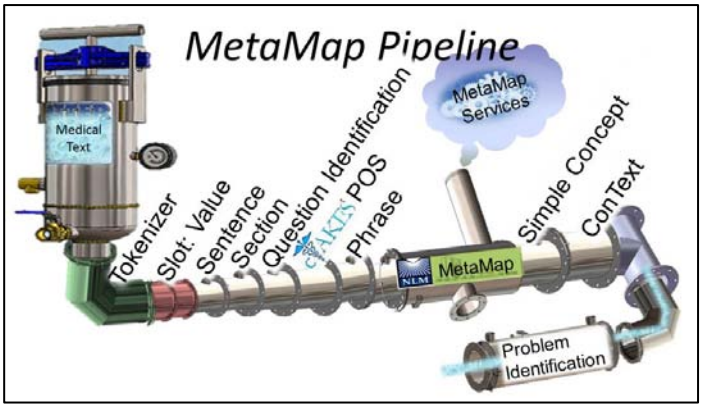


Figure 2. MetaMap pipeline

Span Evaluation

A gold standard reference corpus with span level clinically significant utterances was developed through a Consortium for Healthcare Informatics Research[18] (CHIR) Information Extraction and Modeling task from the Veterans Administration. Human annotators highlighted problems (as defined by the i2b2/VA 2010 Challenge[19]) in 145 clinical records in a corpus chosen at random from the 100 most frequent VA document types by the annotation team that annotated the i2b2/VA 2010 Challenge. Assertion attributes of negation, conditional, and subject were also annotated for these problems.

A MetaMap pipeline and a Sophia pipeline were created for this effort. Each of the pipelines included the problem identification annotator after the concept extraction annotation. The problem annotator filtered to concepts that were of the proscribed problem semantic type. The conText annotator was also applied to add assertion attributions. The conText annotator marked concepts with asserted, negated, conditional, applies-to the-patient, and historical attributions.

Figure 2 shows the MetaMap pipeline used for this evaluation. The cTAKES part-of-speech annotator and a phrase annotator were added and the Sophia annotator is replaced with the MetaMap annotator. The MetaMap annotator is a wrapper around a client that goes out to a MetaMap service running on external machines. This annotator gathers and uniques the phrases for a given record, then makes one request out to the MetaMap service. The MetaMap service runs MetaMap in the *term processing mode*. The MetaMap service is a restful service that includes 60 instances of MetaMap. The service treats each incoming term as a new request to the next available MetaMap process. Even with this environment, 86% of the processing time taken within the MetaMap pipeline is taken within this one annotator when analyzed via the UIMA CPE tool.

The cTAKES application was run separately on the 145 records. The output was fed into a post processing v3NLP pipeline that converted the cTAKES UMLS concept annotations to the CHIR model's *CodedEntries*. The same problem annotator was applied. cTAKES includes assertion attributions.

The span level comparison was done with overlapping matching spans compared to the reference standard on asserted problems associated with the patient.

Both the MetaMap server and Sophia indexed using the 2011AA Level0+9 configuration. cTAKES uses the 2011AA SNOMED concepts. The evaluation used an f-score computed as

$$F\ Score = \frac{2(precision * recall)}{(precision + recall)}$$

where

$$precision = \frac{true\ positives}{(true\ positives + false\ positives)}$$

$$recall = true\frac{positives}{(true\ positives + false\ negatives)}$$

Efficacy Results:

Table 1 shows MetaMap, cTAKES, and Sophia compared to the 145 record reference standard. The evaluation was a span-only evaluation, where credit was given for partial matches. cTAKES has the overall better F-Score at 0.568, followed closely by Sophia at 0.531. MetaMap had an overall f-score of 0.431. Sophia performed better at recall with a metric of 0.71 followed by cTAKES at 0.66. MetaMap had a recall metric of 0.38. cTAKES and MetaMap had a precision metric of 0.5 vs Sophia's .422. Sophia's precision was noticeably lower because of a plethora of false positives for this task. Those false positives from Sophia that were reviewed indicated that many could have been considered medical problems associated with the patient but the annotators chose not to mark them as such.

Table 1. Problem span comparison

	TP	FP	FN	Recall	Precision	F-Score
MetaMap	436	436	717	0.380	0.500	0.431
cTAKES	757	760	391	0.660	0.500	0.568
Sophia	823	1125	325	0.717	0.422	0.531

Differences between Sophia, MetaMap and cTAKES

There was a large overlap between the systems. It cannot be construed that those concepts unique to one system or another are fallacious. Of interest is how well each system identified multi-word spans. The more tokens involved in the match, the less ambiguity is left for downstream processors to deal with. Table 2 shows the Sophia pipeline compared to MetaMap and cTAKES, using MetaMap and cTAKES as reference standards.

Table 2. Problem span comparison using MetaMap and using cTAKES as the reference standard

	TP	FP	FN	Recall	Precision	F-Score
Sophia compared to MetaMap reference standard	1000	1238	169	0.855	0.45	0.587
Sophia compared to cTAKES reference standard	1496	1117	562	0.727	0.57	0.641

Many of the differences found included how each system chunked phrases, with no clear indication of whether either system did better or not. Table 3 shows instances where MetaMap picked up multi-word concepts but Sophia chunked them into separate concepts and instances of where Sophia picked up multi-word terms that include phrasal barrier markers.

Table 3. Multiword matching differences between MetaMap, Sophia and cTAKES

MetaMap	Sophia	cTAKES
right sided facial weakness	facial weakness	facial weakness
multiple old infarcts	infarcts	Infarcts
	hard of hearing	
	change in bowel habits	
	lives with family	
	unable to sit	
	Sensitive to touch	

The largest category of differences between Sophia and cTAKES was that CTAKES annotated terms found in section headings that Sophia did not. A concept mention within a section heading would not indicate that the mention is related to the patient. For instance, a section heading Pain Management would not automatically indicate the patient has or does not have pain; only that there is a section in the document that includes a section with pain in the name. The reference standard did not include annotations from within section headings. The Sophia pipeline includes a sectionizer that marks section headings to be ignored. Table 3 shows multi-word terms that Sophia suggested that were missed by cTAKES.

Time Performance

The Sophia pipeline, the MetaMap pipeline and the cTAKES assertion aggregate annotator were run against two corpora on a development virtual machine (VM) provided by the VA to securely process clinical records. The CHIR reference standard has a shorter average character length than other available corpora. The i2b2 2010/VA Corpus[i] provided an additional benchmark to a corpus with known attributes within the NLP community. The fastest of 3 runs are reported here (Table 4). The throughput numbers are meant to be interpreted as a means to rank the relative performance between the three systems. The performance time of these systems on well-

[i] Parts of the Sophia pipeline were used within an entry in the i2b2 2010 VA Challenge. The whole corpus was used for additional training to improve pipeline components after the challenge. This training invalidates any efficacy evaluation to this corpus.

endowed production servers are vastly different than the VM's provided for development purposes or current desktop machines. cTAKES and Sophia performance time on the i2b2 corpus on the a core i7 desktop was 4 times faster than the development virtual machine, and the MetaMap performance time was 2 times faster on the same corpus on the core-i7 using a less endowed MetaMap server.

Sophia has a significant initialization cost to load all the keys into an in-memory hash. This initialization is the same whether kicking off one instance or 100 due to the way the hash is shared across server threads. The impact of this initialization becomes less as more records are processed. Table 5 shows the initialization cost and the average per-record cost with and without taking into account the initialization cost. The initialization cost with MetaMap is hidden behind the running MetaMap service that was employed. CTAKES does have an initialization time of 36 seconds vs Sophia's 40 seconds observed on a desktop core i7 with solid state drives.

Table 4. Time performance in milliseconds to run Sophia, MetaMap and cTAKES on two corpora of records

	# of Records	Sophia	MetaMap	cTAKES
i2b2 2010 VA Corpus	349	1,395,271 (23.24 min)	4,804,951 (80.08 min)	24,524,827 (408 min)
Problem Reference Standard Corpus	145	343,384 (5.7 min)	478,824 (8 min)	3,060,000 (51 min)

Table 5. Time performance to run Sophia on two corpora of records, reported in milliseconds

	# of records	Initialization in milliseconds	Average milliseconds per Record	Average milliseconds per record w/out initialization	Total milliseconds
i2b2 2010 VA Corpus	349	187,013	70,271	69,735	24,524,827
Problem Reference Standard Corpus	145	185,664	2,351	1,079	343,384

This time evaluation is not perfect. The number of external CPU's and threads employed by the MetaMap services makes it difficult to replicate the same MetaMap pipeline performance if moved to an environment that does not employ the VA's MetaMap services. Even with these constraints, the single threaded Sophia annotator out-performs the MetaMap annotator by a factor of 7 and out-performs cTAKES by a factor of 18.

Pipeline Performance Analysis

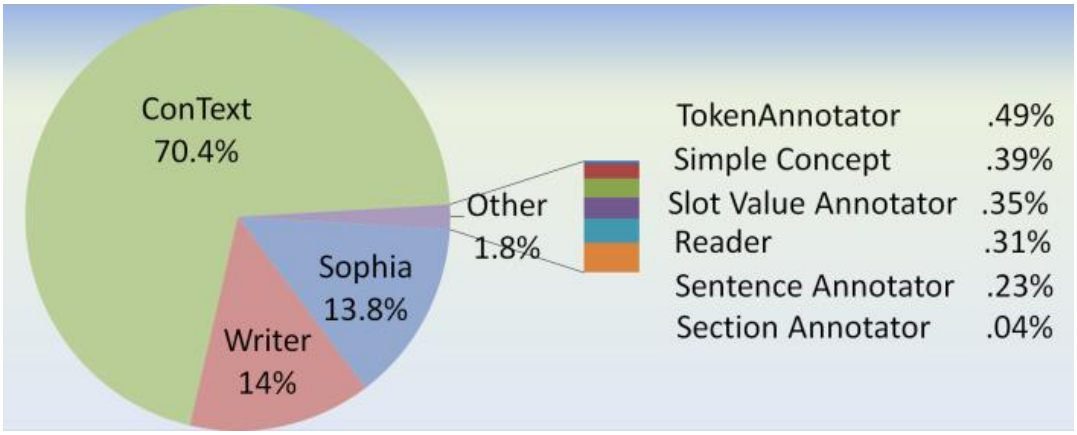


Figure 3. Sophia pipeline proportion of each annotator's processing

A pipeline performance analysis was performed to analyze the time contribution for each of the components within the Sophia, MetaMap, and cTAKES pipelines. The UIMA Component Processing Engine (CPE) was employed to break down each of the component times. See the pie charts in figures 3-5 of the relative amount of time each component consumed. It is the assertion component that takes up the most time (70%) within the Sophia pipeline, and the second highest amount of time (27%) in the cTAKES pipeline, yet it is a mere 9% within the MetaMap pipeline. Within Metamap and cTAKES, other components consume much more processing relative to the assertion module. Efficiencies to conText should be explored before improving performance elsewhere for the Sophia pipeline.

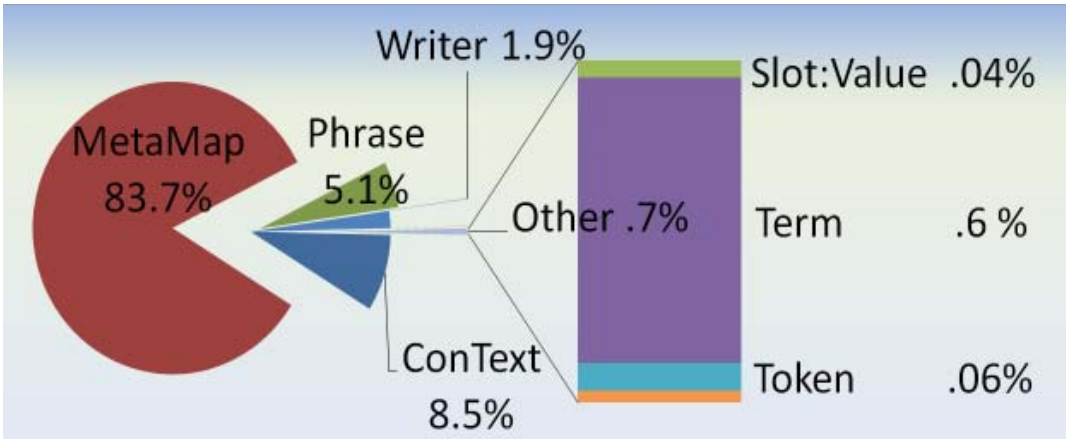


Figure 4. MetaMap pipeline proportion of each annotator's processing

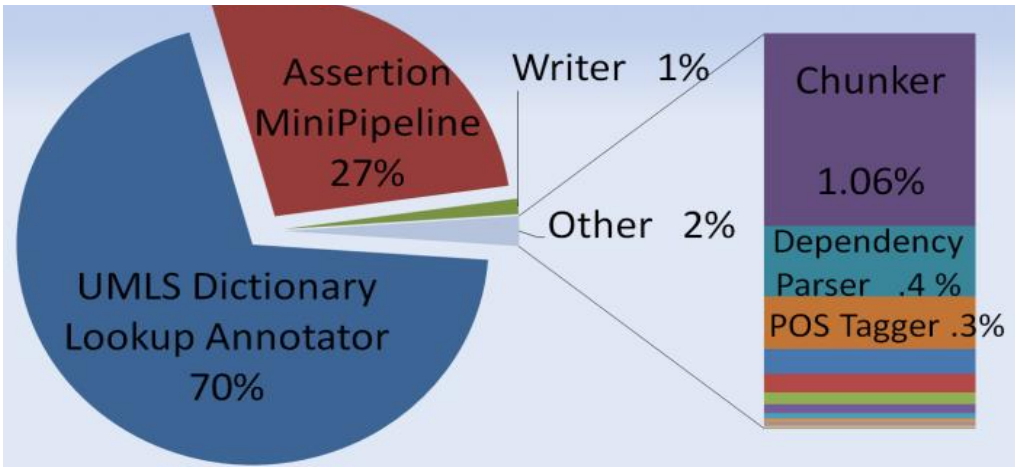


Figure 5. cTAKES pipeline proportion of each annotator's processing

Some efficiency had been built into Sophia's conTEXT wrapper, by spawning off a pool of threads to handle the conTEXT processing. This change contributed a 30% performance improvement compared to using no additional threads.

Discussion

Inspired by currently available tools and with the objective of improving total throughput performance in NLP tasks, we developed Sophia as an expedient UMLS concept extraction annotator. The Sophia pipeline, as configured as a single end-to-end UIMA application for evaluation purposes significantly out performs both MetaMap and cTAKES in throughput. Components of each of the pipelines were examined to further elucidate the bottleneck components. For the Sophia Pipeline, assertion is the most time consuming component, even with some efficiency built around the ConText methods. Evaluation using the extrinsic task of finding clinical problems showed that Sophia has a similar over-all f-score to cTAKES, and out performs MetaMap. Furthermore, Sophia had a better recall than both cTAKES and MetaMap on this task.

The techniques within Sophia are an evolution of the techniques embedded within MetaMap. Sophia borrowed heavily from NLM's SPECIALIST Text Tools™, which were included in the first Java implementation of MetaMap Technology Transfer (MMTx). The Text Tools included the lexical lookup, the part-of-speech tagger and the phrase identification components of MetaMap. Sophia's lexical lookup is a direct descendant to the Text Tools™. Sophia differs from MetaMap in that it does not do the brute force mapping that was included in MMTx; it keeps only the longest spanning matches from the variant table, that is, it does not compute partial matches; it does not do part-of-speech tagging or phrase identification; and it combines the concept information within the variant and lookup table, rather than relying on tables to do the lexical lookup, then lookup in tables to find the variants, followed by lookup in tables to find the concept information for each match.

MetaMap's strength is in the evaluation and ranking it achieves once candidate concepts are pulled from the index. It is in this evaluation where MetaMap churns away. It is the most computationally expensive part of the algorithm, by far. Neither Sophia nor cTAKES includes such an evaluation component. This evaluation component allows MetaMap to retrieve and rank quality near matches that don't quite cover, or cover too much (partial matches, concept gaps and over-matches) from the corpus text. Neither cTAKES nor Sophia retrieves partial matches, concept gaps or over-matches. This increases coverage for information retrieval tasks. If one limits to exact matches (those that have 1000 as the final mapping score within MetaMap), results, in theory, should be equivalent. MetaMap's ranking takes into account the cognitive distance it took between seen text and a UMLS Concept. Sophia retains the cognitive distance but does not use it. Even with this ranking mechanism, MetaMap still returns ambiguous concepts when the ambiguity is at the lexical level. Embedded within MetaMap are techniques to limit ambiguity where it can without having to call upon the services of Word Sense Disambiguation (WSD). Such techniques include stop word filters, the ability to filter by semantic type, truncating by frequency hit cut offs and the like. MetaMap has an add-on WSD service to help ameliorate this facet as well. The Sophia pipeline considers all its ambiguous retrieval results to be a WSD issue that should be addressed properly in a downstream process or annotator, where both local and global context can be utilized.

In comparing Sophia, MetaMap and cTAKES methodology, the Sophia annotator shares many attributes with the cTAKES dictionary lookup annotator, and the *Dictionary Lookup Annotator UMLS* aggregate engine. Both are UIMA based, both include similar windowed lookup techniques. Whereas cTAKES uses LVG's normalization to a normalized index of UMLS strings, Sophia looks up unadulterated tokens in Sophia indexes that are generated via LVG's fruitful variants flow using UMLS Strings as its input. This algorithm was developed as part of MetaMap[1], and had become an LVG function in the early 2000's. Sophia relies on a post filtering of this flow to prune off unnecessarily aggressive or likely to be fallacious variants.

CTAKES matches to the SNOMED vocabulary subset of the UMLS. Sophia indexes to the level 0 + 9 UMLS terminologies which include MeSH and SNOMED. Both the cTAKES and Sophia pipelines were designed for use within the clinical setting, and as such, utilize tokenizer, sentence and section annotators and downstream annotators to add negation, conditional, hypothetical, or not-relating-to-the-patient context.

Sophia relies on sentence annotations created from upstream annotators within a UIMA pipeline. In this way, Sophia is similar to the cTAKES Lookup annotator functionality. Sophia adds *Clinical Statements* filled with *CodedEntries* to each annotated document. Clinical Statements are roughly equivalent to cTAKES *EntityMentions* and *EventMentions*, and even more roughly equivalent to MetaMap's final mappings. A *CodedEntry* is equivalent to cTAKES' *UMLSConcept*, and roughly equivalent to MetaMap's *Candidate Concept*.

Whereas MetaMap and cTAKES formulate candidate phrases for lookup using similar techniques, Sophia does not. MetaMap and cTAKES break text into phrases before concept lookup by first tokenizing into sentences, then doing part-of-speech annotation, followed by phrase detection prior to phrase-to-candidate concept lookup. Sophia, in contrast, relies on upstream annotators to label sentences. Sophia looks up longest matching terms within the sentence, similar to MetaMap's lexical lookup algorithm. Like MetaMap's lexical lookup, Sophia's term lookup uses a longest spanning match, which is an evolution of the algorithm embedded in the SPECIALIST Text Tools, which was embedded in MMTx, the java implementation of MetaMap. It should be noted that MetaMap has the ability to do both longest and shortest spanning matches. Sophia's lookup mechanism has two new attributes not found in the SPECIALIST Text Tools. First, it starts its matches from left to right, using an index where the token keys are reversed. This is done to favor picking up right headed noun phrases. Second, UMLS Concept information is embedded within the indexes, so further lookup is not needed.

Whereas MetaMap first looks up terms within the SPECIALIST Lexicon, then uses those terms for phrase barrier determination, then looks up the phrase tokens to find UMLS concepts from an index of UMLS Concept variants, Sophia looks up terms in the UMLS Concept variant table directly without need for part-of-speech or phrasal boundaries.

An early UMLS principle was to keep knowledge resources like the SPECIALIST Lexicon and the UMLS Metathesaurus separate to keep semantic components out of the SPECIALIST Lexicon and to keep syntactic components out of the Metathesaurus. This allowed maintenance cycles for these resources to be de-coupled. This principle carried forth to continue to decouple the syntactic processing from the semantic processing via first finding terms, , then phrases, then concepts within those phrases, as MetaMap, and to some extent, cTAKES does. Finding multi-word terms, particularly if they come from the SPECIALIST Lexicon, and particularly if they decrease ambiguity, greatly helps phrasal boundary detection from part-of-speech taggers that tag at the single token level of granularity. MetaMap uses the MedPost part-of-speech tagger, which was trained using a corpus that had sparse coverage of the majority of multi-words found in the SPECIALIST lexicon, and from the UMLS Metathesaurus. Term lookup followed by part-of-speech tagging on those words and terms within MMTx is used to make phrasal barrier decisions.

Sophia does away with the need for phrases and consequently, parts-of-speech. The term indexes and the UMLS Concept information are folded into one, indexed off the same key. That's not to say that other annotators shouldn't be run to keep around both part-of-speech and phrasal information. A consequence of ignoring phrasal boundaries within Sophia, longest matching terms that span across phrasal boundaries are retrievable within Sophia, but would be missed via MetaMap and cTAKES.

Although not incorporated here, the Sophia Pipeline, in practice, is often augmented with a local concept annotator combined with a file of local terms and their categories to address tasks where the UMLS lacks coverage. This is a capability included within the v3NLP framework that is not easily replicated within MetaMap or cTAKES.

Future Work

Future versions of Sophia will be integrated into v3NLP's scaled-out architecture, where the slower annotators are replicated as multiple instances behind services, and called via wrappers around clients to these services. The next version of Sophia will be updated to the latest version of the UMLS.

The next version of the Sophia pipeline will include annotators to filter out non-salient false positives including the units of measure, dates, and the like that MetaMap effectively filters out. Further analysis will be spent to understand those multi-word instances that Sophia missed, and vice versa.

Assertion attribution will be looked at further to choose what assertion modules perform the best in respect to time and efficacy.

There is on-going interoperability work to enable v3NLP annotators, Sophia being one, with cTAKES to enable the use of cTAKES annotators within v3NLP and vise-versa.

Availability

Sophia is available via an Apache license, and is distributed from the http://v3nlp.utah.edu/sophia. End users are required to validate their own UMLS license via an application that validates UMLS licenses through the National Library of Medicine's UMLS Terminology Services (UTS) before unlocking the content of indexes that contain UMLS derivative content within this distribution.

Conclusions

Sophia has been developed as an expedient UMLS concept annotator. The Sophia pipeline out performs both cTAKES and MetaMap in recall and has an f-score that is only 0.04 different than cTAKES. The pipeline runs 18 times faster than cTAKES and 7 times faster than the scaled-out MetaMap services. For those information extraction applications where fast throughput is needed and/or recall is favored over precision, the Sophia pipeline is an acceptable solution.

Acknowledgements

This work is funded by US Department of Veterans Affairs, Office of Research and Development, Health Services Research and Development grants VINCI HIR-08-204, CHIR HIR 08-374, ProWATCH grants HIR-10-001 and HIR 10-002. We would like to express our gratitude to the administration and staff of the VA Informatics and Computing Infrastructure (VINCI) for their support of our project. A special thanks to Shuying Shen for her efforts in providing the reference annotations. We also acknowledge the staff, resources and facilities of the VA Salt Lake City IDEAS Center for providing a rich and stimulating environment for NLP research.

References

1. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proceedings / AMIA Annual Symposium AMIA Symposium. 2001:17-21. Epub 2002/02/05.
2. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association : JAMIA. 2010;17(5):507-13. Epub 2010/09/08.
3. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC medical informatics and decision making. 2006;6:30. Epub 2006/07/29.
4. VA Informatics and Computing Infrastructure (VINCI). 2012 [cited 2013]; Available from: http://www.hsrd.research.va.gov/for_researchers/vinci/.
5. Ferrucci D, Lally A. UIMA: an architectural approach to unstructured information processing in the corporate research environment. Nat Lang Eng. 2004;10(3-4):327-48.
6. Lindberg C. The Unified Medical Language System (UMLS) of the National Library of Medicine. J Am Med Rec Assoc. 1990;61(5):40-2. Epub 1990/04/09.
7. Hersh W, Hickam D. Information retrieval in medicine: the SAPHIRE experience. Medinfo MEDINFO. 1995;8 Pt 2:1433-7.
8. Friedman C. A broad-coverage natural language processing system. Proceedings / AMIA Annual Symposium AMIA Symposium. 2000:270-4.
9. MedKAT/p: http://ohnlp.org/index.php/MedKAT/p
10. Denny JC, Peterson JF, Choma NN, Xu H, Miller RA, Bastarache L, et al. Development of a natural language processing system to identify timing and status of colonoscopy testing in electronic medical records. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2009;2009:141.
11. Cunningham H. GATE, a general architecture for text engineering. Computers and the Humanities. 2002;36(2):223-54.
12. Apache UIMA-AS: http://uima.apache.org/doc-uimaas-what.html
13. Lexical Variant Generation Documentation: Fruitful Variants : http://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/lvg/current/docs/designDoc/UDF/flow/fG.html
14. Aronson AR. The effect of textual variation on concept based information retrieval. Proceedings : a conference of the American Medical Informatics Association / AMIA Annual Fall Symposium AMIA Fall Symposium. 1996:373-7.
15. Brown AC, Divita G. The SPECIALIST Text Tools: http://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/textTools/
16. McCray AT, Srinivasan S, Browne AC. Lexical methods for managing variation in biomedical terminologies. Proceedings / the Annual Symposium on Computer Application [sic] in Medical Care Symposium on Computer Applications in Medical Care. 1994:235-9.
17. Chapman WW, Chu D, Dowling JN. ConText: an algorithm for identifying contextual features from clinical text. Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing; Prague, Czech Republic. 1572408: Association for Computational Linguistics; 2007. p. 81-8.
18. Collaboration between VINCI and CHIR. 2012 [cited 2013]; Available from: http://www.hsrd.research.va.gov/for_researchers/vinci/chir.cfm.
19. Uzuner O, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association : JAMIA. 2011;18(5):552-6.

Development of iBsafe: A Collaborative, Theory-based Approach to Creating a Mobile Game Application for Child Safety

Cinnamon A. Dixon, DO, MPH[1], Robert T. Ammerman, PhD[1], Judith W. Dexheimer, PhD[1], Benjamin Meyer, MFA[2], Heekyoung Jung, PhD[2], Boyd L. Johnson, RA[3], Jennifer Elliott, PhD[4], Tom Jacobs, BA[5], Wendy J. Pomerantz, MD, MS[1], and E. Melinda Mahabee-Gittens, MD, MS[1]

[1]Cincinnati Children's Hospital Medical Center, Cincinnati, OH; [2]Univeristy of Cincinnati, Cincinnati, OH; [3]BIMMIB, Cincinnati, OH; [4]Voorstellen, Cincinnati, OH; [5]InterVision Media, Eugene, OR

Abstract

Unintentional injury is a leading cause of death worldwide, and the number one cause of child death in the United States. The American Academy of Pediatrics promotes safety recommendations to decrease child injury risk, however the majority of educational programs delivering these strategies are school-based or in community campaigns. Mobile technology provides an opportune platform to deliver pediatric injury prevention programs given its massive global reach and underrepresentation within the current mobile health market. This paper describes the development of iBsafe, a novel mobile safety game application designed to prevent injury in 5- to 6- year old children. Our multidisciplinary team utilized a step-wise approach to create an innovative child game application which is based in behavioral theory and promotes evidence-based safety recommendations. Results and future directions for iBsafe aim to interactively educate children on how to be safe and ultimately improve their safety behaviors.

Introduction

Unintentional injury is a leading cause of death worldwide[1] and in the United States (US), the number one cause of child death and nonfatal injury.[2] Each year, nearly 10,000 children under 18 years of age die from injuries, and over 8 million suffer non-fatal injuries.[2,3]

Haddon's Matrix – a conceptual framework of etiologic factors – has been used to guide unintentional injury control.[4] Using this framework, injury experts have identified factors that contribute to the risk of unintentional injury and aid in developing evidence-based and expert-guided strategies to decrease this risk. The American Academy of Pediatrics (AAP) and other safety organizations espouse such prevention recommendations, which are often aimed at parents and children, and are categorized by the child's age or the cause of injury.[5] Child safety programs designed to promote these recommendations are typically delivered in classroom curricula or community campaigns which can be time consuming, costly and often have narrow scope and limited penetration.

More recently, there has been growing support and evidence for the use of behavioral science theories and models to help understand and prevent injury risk.[6-8] Computers or games that utilize some or all of these tenets within their design and development have the potential to educate and change the behavior of the user.[9-11] Within the field of health care, a variety of game interventions have been effective in teaching medical education,[12] and interactive games have increased child understanding and management of the heath conditions.[13]

The rapidly expanding field of mobile technology is a potential platform to deliver injury prevention programs with extensive reach. Currently, there are over 6 billion mobile device subscriptions worldwide, representing nearly three-quarters of the world's inhabitants.[14] In the US, approximately 240 million individuals use mobile devices and nearly 160 million people own smartphones.[15] Half of all children under 9 years of age have access to mobile devices in their home and nearly 30% of US parents download applications for their children to use. Additionally, the majority of child time using these devices is spent playing games,[16] presenting an unprecedented opportunity to seed injury prevention messages in the form of game play.

Mobile health (mHealth) is increasing in popularity with health organizations, providers, and consumers. In 2013, there were more than 40,000 health-related applications in the Apple iTunes App Store representing an estimated 660 million downloads. Of these, 16,000 were consumer/patient-targeted healthcare applications, ranging in services from prevention and wellness, diagnosis, finding health care providers, health care education, prescription refills and compliance. Furthermore, 50% of consumer oriented healthcare applications targeting specific demographic groups were for children's health. [17]

Despite this rapid growth of mHealth and the burden of child unintentional injury worldwide, mobile applications related to injury prevention remain underrepresented and few interactively address child injury prevention. Thus the goal of this project was to systematically develop a mHealth tool for child unintentional injury prevention. Given that most children access mobile devices to play games, we believe there is a significant need to develop a safety game that gives children vital prevention information and offers an alternative to other non-educational mobile game play. We aimed to develop this mobile device game application, founded in behavioral theory, using a multidisciplinary team and evidence-based AAP recommendations.

Application Development

Development Team

A multidisciplinary team of experts developed this intervention. Our team included: three pediatricians (two child injury prevention experts [CD and WP] and one child prevention intervention expert [MMG]); a child behavior specialist with expertise in prevention programs for children [RA]; a professional graphic designer and former industry child game programmer [BM]; a human-computer interaction design professional [HY]; two informatics and design experts [JD and BJ]; two masters-level design students; an educational professional specializing in learning with gaming education and multimedia applications [JE]; and a media design/ software development company specializing in interactive technologies for health care, behavioral research and education [TJ].

Behavioral Theory

Social cognitive theory (SCT) is a behavioral theory which has been used within the field of injury to help guide intervention designs and measures.[6,7] SCT theorizes that to understand, predict and change human behavior, one must look at the interaction between the person, environment and behavior. It posits that individuals learn vicariously through observing and modeling other's behaviors and through social reinforcements. For this modeling to be effective the behaviors must keep the observer's attention and be memorable; the ability of the observer to adopt the behaviors relates to his/her self-efficacy and motivation. [6-8,18,19]

Based on these premises, we developed a modified conceptual model of SCT (Figure 1) to link child unintentional injury prevention and gaming. We proposed that central to the prevention of child unintentional injury are the child's safety knowledge and applied behaviors, both of which could be achieved within the our mobile safety game experience and further translated to the environment by the reciprocal nature of the constructs. Specifically, the interactive game centers around three main determinants: Personal, such as knowledge and expectations; Environmental, such as the influence of others and the environment; and Behavioral, such as skills and practice. The interactive game creates an engaging Environment in which the child learns vicariously through their character's behaviors (modeling) and the subsequent rewards and consequences (reinforcements) of these behaviors. Self-efficacy is achieved within the simulated Environment as the child becomes more knowledgeable about safety, which translates to the child's real-life Personal knowledge and expectations. Motivation to adopt safe Behaviors in the real world is reinforced by the child's experiences within game play itself.

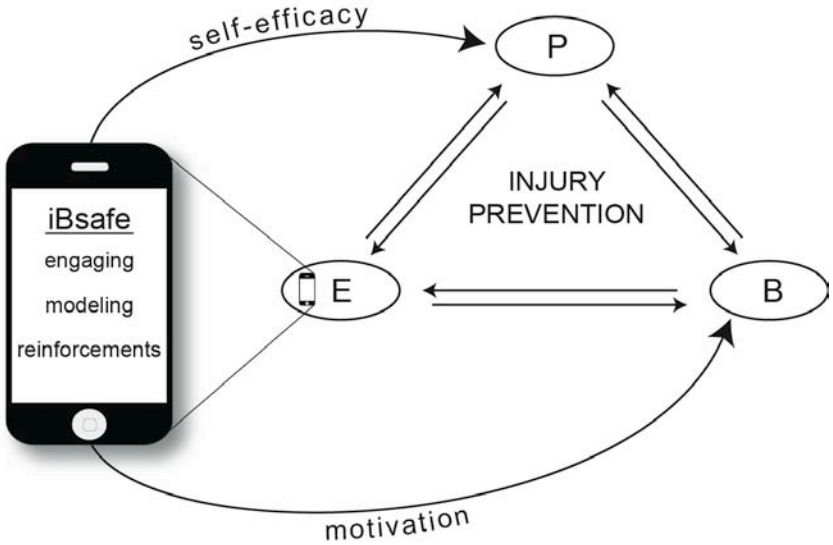


Figure 1. iBsafe Conceptual Model

User Group and Injury Mechanism Selection

There are many different causes of child unintentional injury and associated risks vary by child age and developmental capability. Ideal games should be developed with these factors in mind. Our first priority was to identify our user group and specific injury causes we aimed to address.

Knowing that younger children bear much injury burden, we chose to develop a game for kindergarten-aged children. This 5- to 6-year-old population represents approximately 500 deaths and over 700,000 nonfatal unintentional injuries in the US each year.[2,3] Of these, about 2,000 are seen in the nation's emergency departments (ED) every day and their associated annual direct and indirect costs are estimated to be nearly $3.5 billion.[20]

Two of the top ten leading causes of non-fatal injury in 5- to 6-year-old children are bicycle injuries and dog bites.[3] Compared to the rate of these injuries among the general population (17.3

per 1,000 and 11.5 per 1,000, respectively), bicycle injury and dog bite rates among 5- to 6-year-old children are nearly double (36.7 per 1,000 and 22.6 per 1,000, respectively).[3] Together these two injury mechanisms account for over 50,000 ED visits and 1,000 hospitalizations of 5- to 6 year-old children each year, with annual costs approximating $250 million.[20]
Therefore, our team elected to address the topics of bicycle safety and dog bite prevention in our child safety application. We created and named our application iBsafe for "Interactive Bike and Bite Safety".

Injury Content
Many AAP bicycle injury and dog bite prevention recommendations exist for 5- to 6-year-old children.[21-25] These recommendations are expert-guided or supported by scientific evidence. Using these recommendations and the existing literature as a guide, we selected 10 specific safety strategies to be the learning objectives of the iBsafe game (see Table 1). These objectives are concepts or actions that a 5- to 6-year-old child is capable of independently understanding and performing.

Table 1. Injury prevention strategies.

Bike	Dog
Always wear a helmet	Never approach or touch a dog that is eating or chewing on bones/toys
Always wear helmet on the top of head	Never approach or touch a dog that is caring for puppies
Always click helmet straps under chin	Never approach a dog that is behind a fence
Never ride on the street	Never approach a dog that is tied
Always ride with an adult	Never run from an unknown dog

Mobile Device Output
In addition to delivering injury prevention strategies, we aimed to design and program iBsafe to facilitate the capture and recording of game scores and frequency of use. Such data is stored within the Apple iPod touch® device's "sandbox" which can be downloaded to a host computer or uploaded to a server.[25] This step would allow us to better understand and analyze player game utilization during iBsafe testing.

Game Design and Application Programming
Design and programming of the game application occurred from November 2011 – October 2013. The principal developer (CD) oversaw all phases of iBsafe development. Collaborators and consultants in their respective areas of expertise were involved throughout the entire process. The following stepwise framework was used:

1) *Surveying current game applications and vetting potential game designs (~ 1 month).* In this step, our group reviewed multiple game applications on the market for children aged 3-9 years. We assessed applications for playability of our ultimate user group, analyzing games for the typical 5- to 6-year old child's development and comprehensibility. From this step, we selected a game design that would allow the player to use screen-touch and device motion play. This review also revealed that no game applications addressing child bicycle or dog bite

safety in our user age group were publically available on the Apple iTunes App Store.

2) *Determining the schematic design on paper (~2 months).* This step entailed the creation and review of our game's schematic design. During this iterative process, we identified and selected character and player graphics; design of game play using streets, sidewalks, houses, yards and dogs; and the visual scoring schema.

3) *Creating the game script (~ 1 month).* Game script creation was a concerted process harmonizing the premise of our conceptual model, teaching the injury prevention strategies and interactively engaging the user group based on their level of cognitive development. The script was designed to emphasize injury prevention strategies multiple times while engaging the user with different scenarios and multiple levels of play.

4) *Programming initial prototype (~8 months).* Programming of the initial prototype was an iterative process. Given that our ultimate user group is a young child, we chose to develop the game for an iPod touch using the iOS platform. Unity game development software was used for game programming. We held regular meetings to assess the current status of the prototype, and ensure consistent progress and maintenance of a common vision and goal.

5) *Prototype completion, incorporation of device output, and internal usability testing (~12 months).* Final programming of the prototype, completion of schema, and incorporation of device output was led by the principal developer and our mHealth developer company. Internal usability testing was performed by mHealth experts, select individuals in the team, and children aged 5-6 years. We asked individuals to "think-aloud" during internal usability testing and asked the children to give general game feedback.

Results and Discussion

We have developed iBsafe – a novel interactive mHealth game application for child safety. The development of this application used a unique multidisciplinary approach involving experts in the fields of child injury prevention, child behavior, intervention development, informatics, game education, design, programming, and usability. Furthermore its development was based on our innovative iBsafe conceptual model, which posits that unintentional child injury prevention requires children to have safety knowledge and apply those behaviors, both of which can be modified in the game experience and translated to real environment. Specifically, the interactive game provides a safe, engaging environment where the child learns by being presented with bicycle and dog bite risks; knowledge and behaviors learned during game play can be further translated to real life.

Figure 2. iBsafe game start screen.

Figure 3. iBsafe character selection screen.

The full iBsafe (Figure 2) game experience entails 67 scenes, 5 levels and 1 bonus level. It commences with the user selecting one of the 6 child characters to represent the player for the remainder of the game (Figure 3). Throughout the game, the player interacts with multiple simulated environments testing and teaching the evidence-based bicycle safety and dog bite prevention strategies. These environments include: a simulated street environment, inside and outside of homes, a dog shelter, and a park.

Specific to bicycle safety, the player is encountered with challenges in which he/she must prepare to and safely ride a bike, recognize characters that are not riding safely, and finally teach others how to be safe while riding bikes. Figures 4 and 5 are example bike safety screen shots.

Figure 4. iBsafe bike safety screen 1.

Figure 5. iBsafe bike riding screen 2.

Specific to dog interactions, the player encounters challenges in which he/she must choose safe interactions when exposed to both familiar and unfamiliar dogs in different scenarios (such as eating, behind a fence or tethered, caring for puppies), as well as situations in which he/she must teach others how to be safe around dogs. Figures 6 and 7 are examples dog safety screen shots.

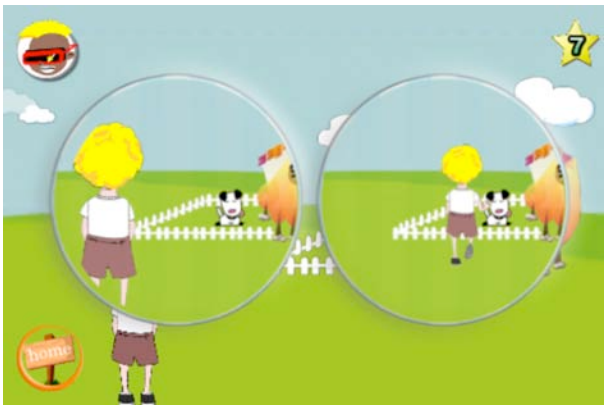


Figure 6. iBsafe dog safety screen 1.

Figure 7. iBsafe dog safety screen 2.

Scoring for game play involves attaining points (represented by stars). There is no maximum score threshold within game play. To gain points and advance to the next level, the player must answer challenges (questions) correctly and perform tasks safely. The player gains one point for every correct answer and loses one point for every incorrect answer; the player gains points for riding their bicycle safely and loses points for riding the bicycle unsafely. The game ends when the player completes all 67 scenes or if the player is hit by a car while riding the bicycle.

Data surrounding every point gained or lost during game play is tracked and can be uploaded as a comma separated value file from the device or onto a server. These data include: a time stamp, character chosen, game and level, current game score and device game high score (see Figure 8.)

Game Number	Level Name	Time stamp	Level	Character	Attempts	Successes	Current Score	High Score
1	Get Ready to Ride	11/19/2013 21:08	0	character3	FALSE	1	0	0
1	Get Ready to Ride	11/19/2013 21:08	0	character3	FALSE	2	0	0
1	Get Ready to Ride	11/19/2013 21:08	0	character3	TRUE	3	1	1
1	Get Ready to Ride	11/19/2013 21:08	1	character3	FALSE	1	0	1
1	Get Ready to Ride	11/19/2013 21:08	1	character3	TRUE	2	1	1
1	Get Ready to Ride	11/19/2013 21:09	2	character3	TRUE	1	2	2
1	Get Ready to Ride	11/19/2013 21:09	3	character3	FALSE	1	1	2
1	Get Ready to Ride	11/19/2013 21:09	3	character3	TRUE	2	2	2
1	Get Ready to Ride	11/19/2013 21:09	4	character3	FALSE	1	1	2
1	Get Ready to Ride	11/19/2013 21:09	4	character3	TRUE	2	2	2
1	Get Ready to Ride	11/19/2013 21:09	5	character3	TRUE	1	3	3
1	Bike game	11/19/2013 21:10	6	character3	FALSE	1	2	3
1	Bike game	11/19/2013 21:10	6	character3	FALSE	2	1	3
1	Bike game	11/19/2013 21:10	6	character3	FALSE	3	0	3
1	Bike game	11/19/2013 21:10	6	character3	FALSE	4	0	3
1	Bike game	11/19/2013 21:10	6	character3	FALSE	5	0	3
2	Get Ready to Ride	11/19/2013 21:12	6	character3	TRUE	1	1	3
2	Get Ready to Ride	11/19/2013 21:12	7	character3	TRUE	1	2	3
2	Get Ready to Ride	11/19/2013 21:12	8	character3	TRUE	1	3	3
2	Get Ready to Ride	11/19/2013 21:12	9	character3	FALSE	1	2	3
2	Get Ready to Ride	11/19/2013 21:12	9	character3	TRUE	2	3	3
2	Get Ready to Ride	11/19/2013 21:13	10	character3	FALSE	1	2	3
2	Get Ready to Ride	11/19/2013 21:13	10	character3	TRUE	2	3	3
2	Get Ready to Ride	11/19/2013 21:13	11	character3	TRUE	1	4	4
2	Bike game	11/19/2013 21:13	12	character3	FALSE	1	3	4
2	Bike game	11/19/2013 21:13	12	character3	TRUE	2	4	4
2	Our friend is riding his bike da	11/19/2013 21:14	13	character3	FALSE	1	3	4
2	Our friend is riding his bike da	11/19/2013 21:14	13	character3	TRUE	2	4	4
2	And how do they make sure t	11/19/2013 21:14	14	character3	FALSE	1	3	4
2	And how do they make sure t	11/19/2013 21:14	14	character3	TRUE	2	4	4
2	Our neighbor got a new dog	11/19/2013 21:15	15	character3	FALSE	1	3	4
2	Our neighbor got a new dog	11/19/2013 21:15	15	character3	TRUE	2	4	4
2	Bike game	11/19/2013 21:15	16	character3	FALSE	1	3	4

Figure 8. iBsafe example of iPod touch output.

Conclusions and Future Directions

We developed iBsafe – an interactive mHealth application – to teach and promote child safety. Development of this application is innovative in four distinct ways. First, it addresses two major child injury mechanisms (bicycle and dog bites injuries), for which there is a significant prevention need. This need is evidenced by the burden of these injuries among children and their associated costs. Second, it utilizes an innovative approach – an interactive game developed and founded in behavioral theory – to educate well-established safety skills, which allows the child user to explore and encounter simulated scenarios of risk in a safe environment. Third, it employs the rapidly expanding technology of mobile devices - specifically applications - to deliver the injury prevention intervention, which can be easily disseminated and applied on a large scale. Lastly, to our knowledge, it is the first interactive mobile device game application aimed at injury prevention for kindergarten-aged children. Given the preponderance of US children using mobile devices to play games, we believe there is an extraordinary market for a safety game application that gives kids vital prevention information and is an engaging alternative to other non-educational mobile game play.

Future directions for iBsafe entail testing it for effectiveness and acceptability in our user population and subsequently expansion into an iBsafe mHealth application series to include other leading causes of child unintentional injury. Our hope is that the iBsafe mobile application game series will interactively educate children on important injury prevention strategies, be supported by sound, scientific evidence, and ultimately, help keep kids safe from harm.

Acknowledgements

This project was supported by the Division of Emergency Medicine at Cincinnati Children's Hospital Medical Center. The authors gratefully acknowledge: University of Cincinnati College of Design, Architecture and Planning (DAAP) graduates Ms. Yingxue (Anne) Zhao and Mr. Da (Todd) Shen, for their hard work and persistence creating initial prototype screen images; DAAP Director of Graduate Program in Design, Mr. Paul Mike Zender, for his coordination of the graduate students and faculty; and Dr. Corey Showalter from Riley Hospital for Children, for providing the game's audio voiceover.

References

1. Peden M, McGee K, Sharma G. The injury chart book: a graphical overview of the global burden of injuries. World Health Organization. 2002.
2. Centers for Disease Control and Prevention (CDC), National Center for Injury Prevention and Control (NCIPC). WISQARS Ten Leading Causes of Fatal Injury Reports. Available at: http://webappa.cdc.gov/sasweb/ncipc/leadcaus10_us.html.
3. Centers for Disease Control and Prevention (CDC), National Center for Injury Prevention and Control (NCIPC). WISQARS Leading Causes of Nonfatal Injury Reports. Available at: http://webappa.cdc.gov/sasweb/ncipc/nfilead2001.html.
4. Runyan C. Introduction: back to the future--revisiting Haddon's conceptualization of injury epidemiology and prevention. Epidemiol Rev. 2003;25:60-64.
5. American Academy of Pediatrics (AAP), Healthy Children Organization. Safety Prevention. Available at: http://www.healthychildren.org/English/safety-prevention/all-around/Pages/default.aspx
6. Gielen A, Sleet D. Application of behavior-change theories and methods to injury prevention. Epidemiol Rev. 2003;25:65-76.
7. Sleet D, Carlson-Gielen A, Diekman S, Ikeda R. Peventing Unintentional Injury: A Review of Behavior Change Theories for Primary Care. Am J Lifestyle Med. 2010;4(1):25-31.

8. Trifiletti L, Gielen A, Sleet D, Hopkins K. Behavioral and social sciences theories and models: are they used in unintentional injury prevention research? Health Educ Res. June 2005;20(3):298-307.
9. Linehan C, Kirman B, Lawson S, Chan G. Practical, appropriate, empirically-validated guidelines for designing educational games. Paper presented at: Association for Computing Machinery (ACM), Computer Human Interaction Conference (CHI) 2011; New York, NY.
10. Lee J, Luchini K, Michael B, Norris C, Soloway E. More than just fun and games: Assessing the value of educational video games in the classroom. Paper presented at the CHI '04 Extended Abstracts on Human Factors in Computing Systems. Vienna, Austria 2004.
11. O'Neil H, Wainess R, Baker E. Classification of learning outcomes: Evidence from the computer games literature. The Curriculum Journal. 2005;16(5):455-474.
12. Blakely G SH, Cooper S, Allum P, Nelmes P. Educational gaming in the health sciences: systematic review. Adv Nurs. Feb 2009;65(2):259-269.
13. Lieberman D. Management of chronic pediatric diseases with interactive health games: theory and research findings. J Ambul Care Manage. January 2001;1:26-38.
14. The World Bank. Mobile Phone Access Reaches Three Quarters of Planet's Population. July, 2012. Available at: http://www.worldbank.org/en/news/2012/07/17/mobile-phone-access-reaches-three-quarters-planets-population
15. comScore Reports January 2014 U.S. Smartphone Subscriber Market Share. Available at: http://www.comscore.com/Insights/Press_Releases/2014/3/comScore_Reports_January_2014_US_Smartphone_Subscriber_Market_Share
16. Zero to Eight: Children's Media Use in America. Common Sense Media. 2011. http://www.commonsensemedia.org/research/zero-eight-childrens-media-use-america
17. IMS Institute for Healthcare Informatics. October 2013 Patient Apps for Improved Healthcare. Available at http://www.imshealth.com/deployedfiles/imshealth/Global/Content/Corporate/IMS%20Health%20Institute/Reports/Patient_Apps/IIHI_Patient_Apps_Report.pdf
18. Bandura A. Self-Efficacy: Toward a Unifying Theory of Behavioral Change Psychology Review. 1977;84:191-215.
19. Bandura A. Social foundations of thought and action: A social cognitive theory New York, NY: Prentice Hall; 1985.
20. Centers for Disease Control and Prevention (CDC), National Center for Injury Prevention and Control (NCIPC). WISQARS Cost of Injury Reports Available at: http://wisqars.cdc.gov:8080/costT/.
21. American Academy of Pediatrics (AAP), Healthy Children Organization. What you should know about dog bite prevention. Available at: http://www.healthychildren.org/English/news/Pages/Dog-Bite-Prevention.aspx.
22. American Academy of Pediatrics (AAP), Healthy Children Organization. Prevent Bite wounds. http://www.healthychildren.org/English/health-issues/conditions/prevention/Pages/Prevent-Bite-Wounds.aspx. Accessed Jan 11, 2012.
23. American Academy of Pediatrics (AAP). The Injury Prevention Program. http://www2.aap.org/family/tippmain.htm.
24. American Academy of Pediatrics (AAP), Healthy Children Organization. What kids should know when bike riding. http://www.healthychildren.org/English/safety-prevention/at-play/Pages/What-Kids-Should-Know-When-Bike-Riding.aspx.
25. App Sandbox Design Guide [computer program]. Version September 27, 2011: Apple Mac OS X Developer Library.

Participatory Design and Development of a Patient-centered Toolkit to Engage Hospitalized Patients and Care Partners in their Plan of Care

Patricia C. Dykes PhD, RN[1,2], Diana Stade[1], Frank Chang MSE[3], Anuj Dalal MD[1,2], George Getty[3], Ravali Kandala[3], Jaeho Lee MD, PhD[1], Lisa Lehman MD, PhD[1,2], Kathleen Leone MBA, RN[1], Anthony F. Massaro MD[1, 2], Marsha Milone RN, MSN[1], Kelly McNally[1], Kumiko Ohashi PhD, RN[1], Katherine Robbins, RN[1], David W. Bates MD, MSc[1,2,3], Sarah Collins, PhD, RN[2,3]

[1]Brigham and Women's Hospital, Boston, MA, [2]Harvard Medical School, Boston, MA, [3]Partners HealthCare, Boston, MA,

Abstract

Patient engagement has been identified as a key strategy for improving patient outcomes. In this paper, we describe the development and pilot testing of a web-based patient centered toolkit (PCTK) prototype to improve access to health information and to engage hospitalized patients and caregivers in the plan of care. Individual and group interviews were used to identify plan of care functional and workflow requirements and user interface design enhancements. Qualitative methods within a participatory design approach supported the development of a PCTK prototype that will be implemented on intensive care and oncology units to engage patients and professional care team members developing their plan of care during an acute hospitalization.

Keywords: participatory design, plan of care, communication, health information technology, patient engagement, nursing informatics.

Background and Significance

Acute care hospitals have been long recognized as complex, dynamic, and fast-paced environments characterized by suboptimal communication.[1, 2] Providing tools to support communication, patient activation and engagement can improve patient outcomes and lower healthcare costs.[3] Engaging patients in their recovery plan and providing them with information they need to be informed participants in that plan are key strategies for adverse event and error prevention.[4] The Meaningful Use legislation includes incentives for providing patients with timely information including the use of secure electronic messaging for providers to communicate with patients about their health status.[5] However, there is limited literature describing tools to activate patients and engage them in their plan of care during an acute hospitalization.

In our previous work we found that by providing all care team members including patients and family with information about a patient's risk for falls and a personalized prevention plan, the rate of patient falls was reduced by 22%.[6] Our team has also found that effective patient education can reduce the incidence of preventable adverse drug events.[7] Others have found that interactive educational tools can improve patient adherence with their treatment plan and improve patient outcomes.[8] Based on these findings, in prior work our team conducted focus groups and interviews with patients, family caregivers, and professional providers to identify the core set of information needed to engage in their recovery during an acute hospitalization[9] and determined the following core information requirements: 1) plan of care, 2) tailored patient education, 3) communication of safety alerts, 4) diet restrictions and 5) medications. We built an electronic Bedside Communication Center (eBCC) prototype to activate patients and to bridge the communication gap between the healthcare team and patients during an acute care hospitalization. We then conducted initial usability testing with hospitalized patients on medical units and found that overall patients and family found the eBCC both useful and easy to use.[10] However, they identified some gaps in the functionality including the ability to message their care team and the ability to directly participate in developing the plan of care. The purpose of this paper is to describe the participatory design process that our team is using to engage patients, family, and health care professional stakeholders in the development of a patient-centered toolkit (PCTK) that will be implemented on intensive care and oncology units to engage patients and family in their plan of care during an acute hospitalization. The research questions are:

1) What are the workflow requirements needed to support patient involvement in developing the plan of care and direct communication with providers?

2) What are the content/user interface requirements needed to support patient involvement in developing the plan of care and direct communication with providers?

AHRQ Conceptual Framework of Patient and Family Engagement

This project is informed by the AHRQ conceptual framework of Patient and Family Engagement[11] that defines the organizational and individual characteristics and behaviors that are antecedents to patient engagement. The framework identifies anticipated outcomes of patient engagement needed to ultimately achieve the triple aim of improving the patient experience, improving the health of populations, and reducing the cost of healthcare.[12]

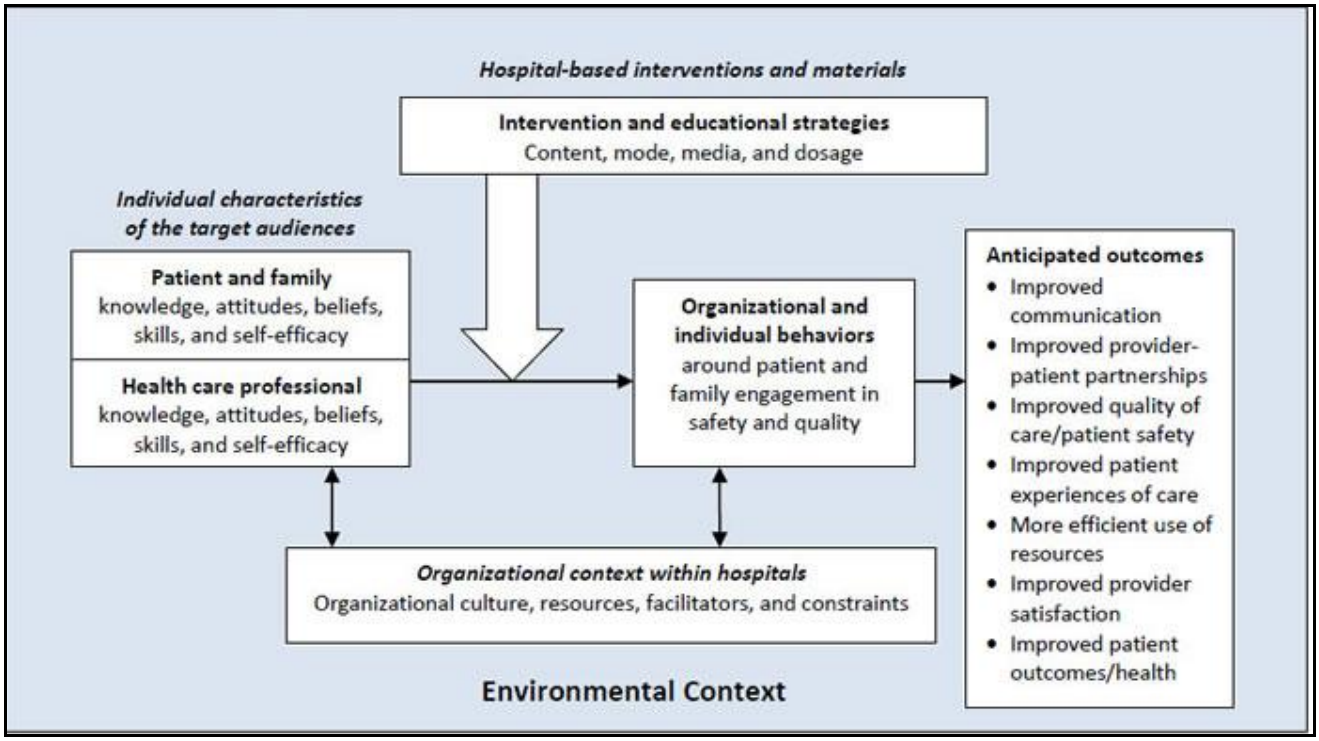


Figure 1: AHRQ Conceptual Framework of Patient and Family Engagement

According to this framework, the PCTK intervention will be effective if it facilitates patient, family, and health care professional engagement in the specific behaviors needed to support safety and quality. Therefore the PCTK should focus on organizational and individual behaviors (patient, family and health care professional) that are acceptable to each stakeholder and that are feasible within the hospital setting. Engagement with stakeholders is needed to identify the specific behaviors that are acceptable and feasible and to develop a PCTK that can be used by all stakeholders in the context of acute care workflows.

Research Methods

Our specific aim was to identify the workflow and user interface requirements needed to support patient involvement in developing the plan of care and communicating with providers related to that plan during an acute hospitalization. We used an iterative participatory design approach to develop the plan of care and provider communication features of the PCTK. The study was reviewed and approved by the Partners HealthCare Human Research Committee and was conducted at Brigham and Women's Hospital (BWH), an academic medical center in the Northeastern United States. We used a combination of individual interviews and group interviews to engage patients, family, and professional health care team members in development and refinement of a series of prototypes. Participants included former patients and their family members, and nurse and physician providers who cared for patients on the intensive care and oncology units at BWH. Interviews were conducted at the bedside with hospitalized inpatients and family members. Group interviews were conducted with discharged patients and family and with professional care team members in conference rooms at BWH. Inclusion criteria for patients included current or past inpatient admission (patient or family member) in an intensive care or oncology unit or a professional care team member working on those units, over 18 years of age, awake and alert, able to understand and speak English, able to provide feedback on their experience and on the PCTK prototype.

Iterative Participatory Software Development. We started with a PCTK prototype that we developed previously for patients hospitalized with an acute medical condition.[9, 10] An iterative participatory software development process (see **Figure 1**) was used to engage stakeholders in developing a series of prototypes; each incorporating what we learned related to the user's (patient/family and professional care team member) role and behaviors in plan of care workflow and associated communication. Through this process we aimed to identify workflow and system usability constraints that were then addressed in subsequent prototypes to produce a PCTK

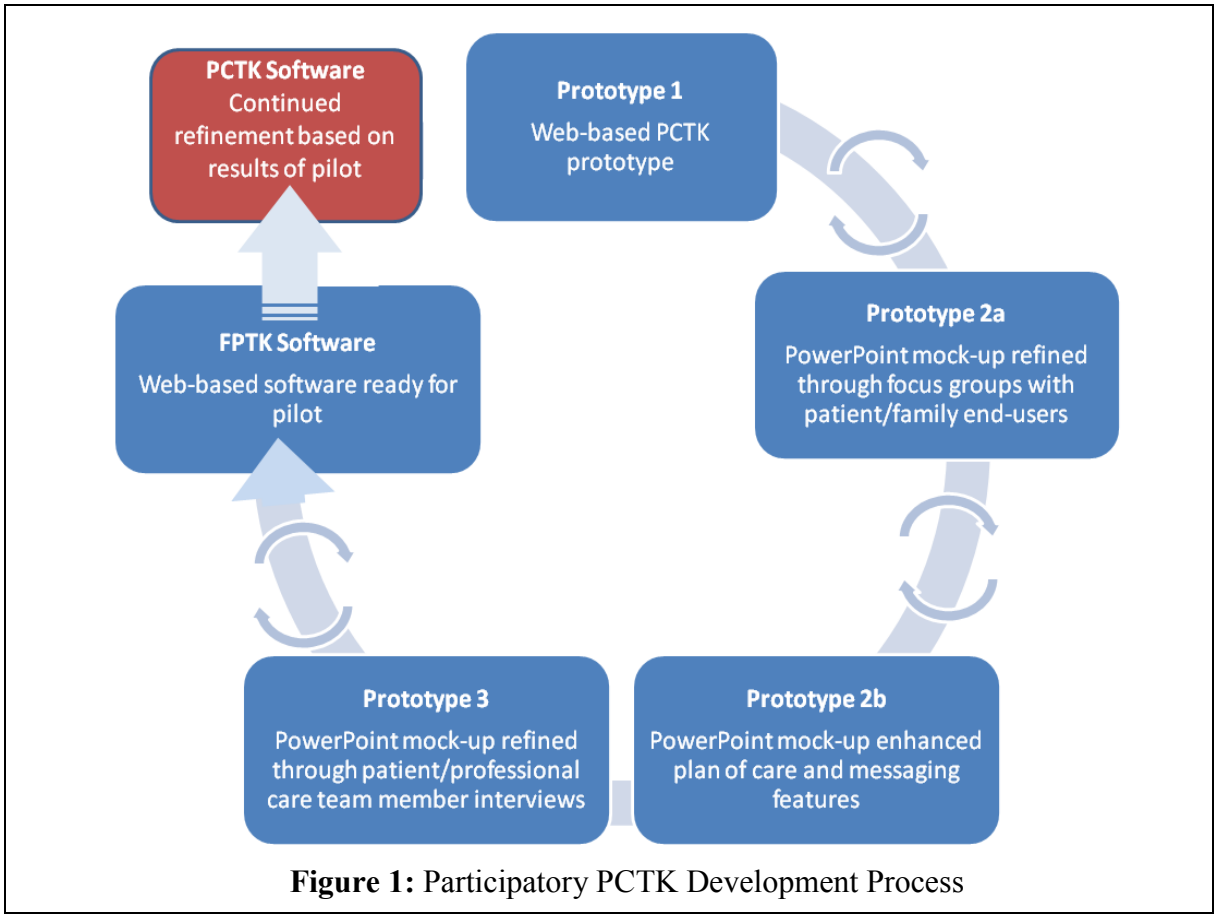


Figure 1: Participatory PCTK Development Process

that would enable patients, family, and professional care team members to collaborate and communicate in regards to the plan of care.[13, 14]

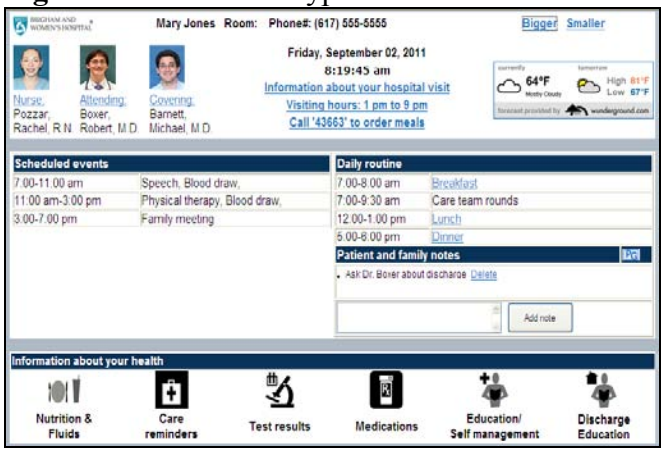
The types of interviews conducted to engage end users in development and refinement of the PCTK are as follows:

1. *Bedside interviews with patients and family (Round 1):* We started with the prototype developed in our previous research (*Prototype 1* see **Figure 2**).[9, 10] Prototype 1 was accessed via the internet on a secure mobile tablet and tested for usability at the bedside with hospitalized medical patients and family members.

2. *Patient/Family Advisory Council group interview:* Group interview with former oncology and intensive care patients and their family members using PowerPoint mock-ups (*Prototypes 2a, 2b*) followed by individual interviews to verify requirements and user interface enhancements.

3. *Professional care team group interviews:* Series of interviews conducted over 6 months with professional care team members to identify existing workflows and opportunities for integrating the toolkit into workflow and to identify opportunities for secondary use of data to populate the PCTK. PowerPoint mock-ups (*Prototypes 2a, 2b*) and the web-based software (*Prototype 3*) were informed by these interviews.

Figure 2: PCTK Prototype 1



4. *Bedside interviews with patients/family (Round 2)*: Individual interviews were conducted on oncology units using an interview guide and Prototype 3 to identify patient/family preferences related to participation in development of the plan of care and to vet the features and workflows defined by professional providers. Prototype 3 was accessed via the internet on a secure mobile tablet.

5. *Meetings with clinical and administrative leaders and stakeholders:* In addition to engaging, patients, family, and professional providers, we met with clinical and administrative leaders and stakeholders to ensure that our project fit in with the BWH organizational informatics strategy.

Interview guides were used to structure each group and individual interview to ensure that outstanding questions related to plan of care workflows, communication and the user interface were addressed. Investigators took detailed notes during the interviews and debriefed at weekly research team meetings where notes were reviewed, emergent themes identified, and requirements refined. Changes were incorporated into PowerPoint mock-ups and iteratively revised by the research team. Mock-ups were then shared with interview participants to validate new requirements and to vet the refinements made to the user interface.

Results

1. *Bedside interviews with patients and family (Round 1):* Eight patients and three family members participated in the first round of bedside interviews. Sessions were conducted on general medical units. While detailed results have been reported elsewhere[10], two related findings were that patients and family requested additional functionality to support their participation in the patient plan of care and for communicating directly with providers on issues related to their plan.

Resulting Mock-up: Distinct plan of care and communication features added to the prototype.

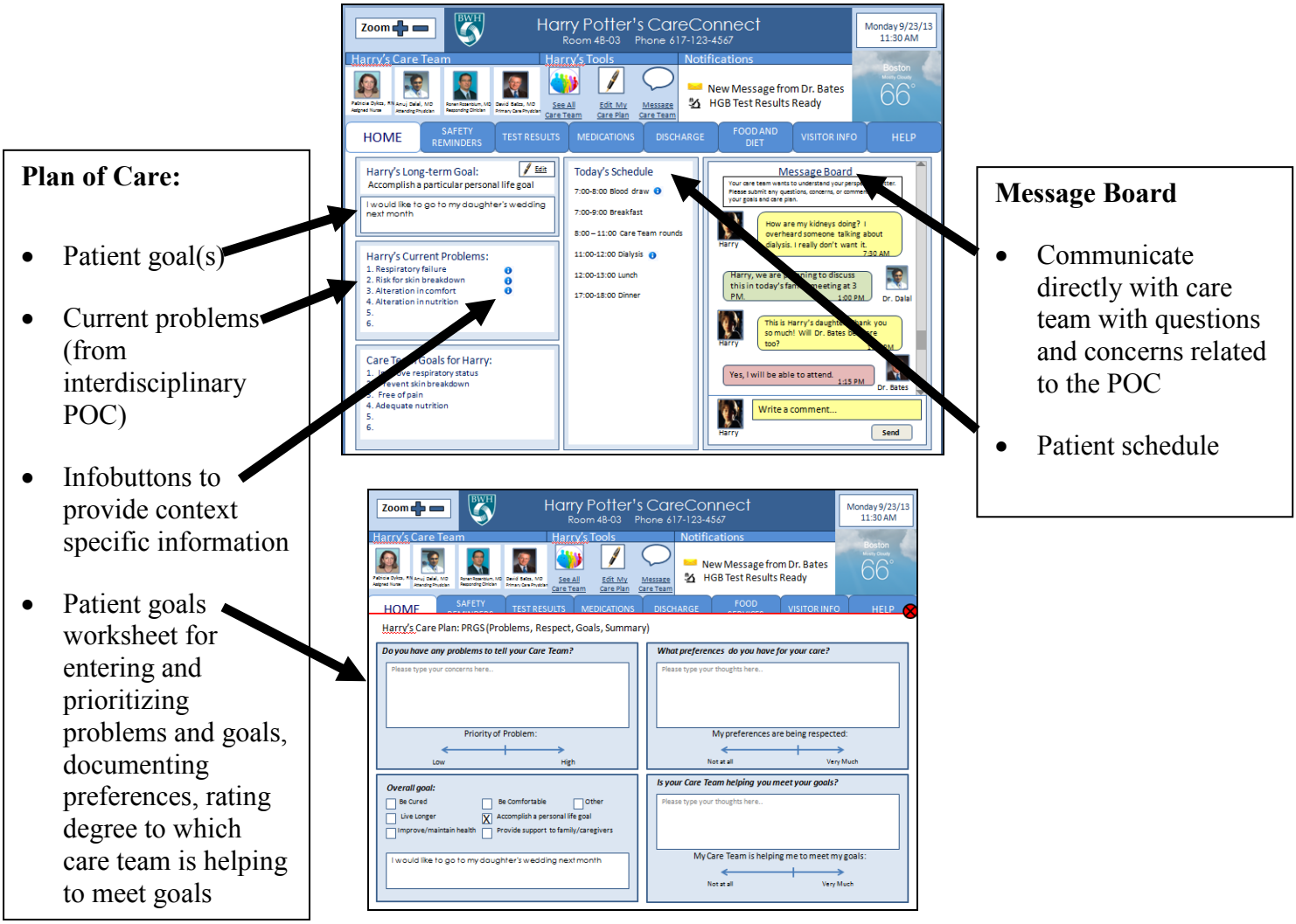


Figure 3: PCTK Prototype 2

2. *Patient/Family Advisory Council group interview:* Discharged patients and family members participated in a group interview and follow-up individual interviews. A total of 12 patients/family members participated in the group interview. Individual follow-up was completed with one patient and one family caregiver to further explore and validate the themes identified in the group interview. While the majority of participants said that they liked the idea of providing patients and family access to information to help them participate in their POC, many said that the user interface was too busy and could be confusing for patients who are not technically savvy.

3. *Professional care team group interviews:* A total of 18 Nurses and 10 physicians participated in a series of group interviews. Overall participants said that they thought that the PCTK would be useful for some patients but they were concerned that the user interface would be difficult to navigate for older patients.

Resulting Mock-up: User interface simplified.

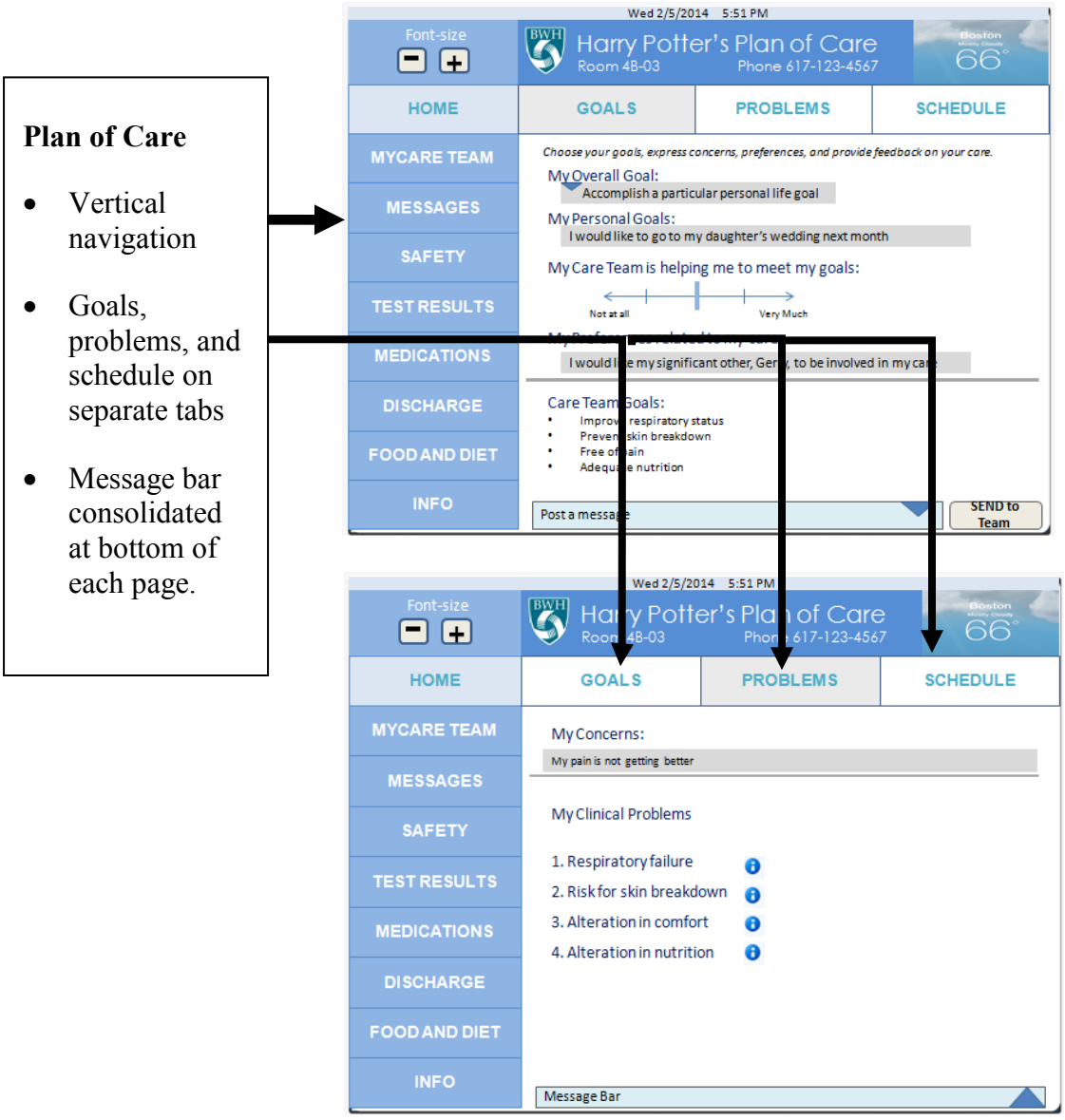


Figure 4a: PCTK Prototype 3 (Patient Goals and Problems)

Patient Schedule

- Consults, tests, meals, rounds
- Infobuttons to provide context specific information

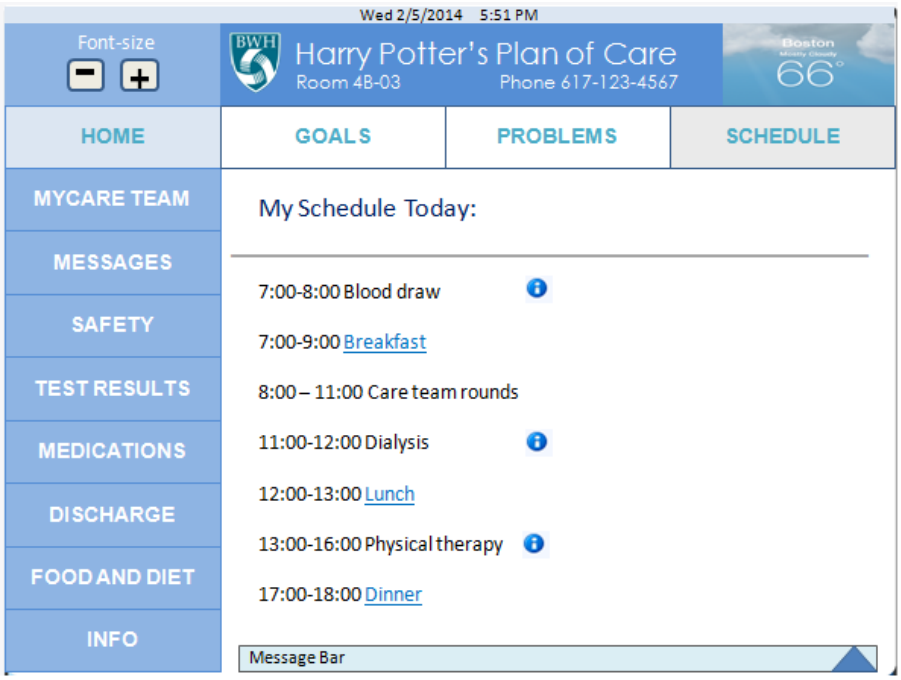


Figure 4b: PCTK Prototype 3 (Patient Patient Schedule)

4. *Bedside interviews with patients/family (Round 2)*: A total of five patients and two family members participated in the second round of bedside interviews. **Table 1** includes a summary of patient/family preferences related to participation in development of the plan of care and the features and workflows defined in group interviews by professional providers.

Plan of Care Elements and Workflow	Personally would use (n=7)	Thought others would use (n=7)	Favorite Element
Personal goals for hospitalization	86%	100%	
Feedback to care team on how well they are helping to accomplish goals	71%	100%	
Preferences for care	57%	100%	
Care team clinical goals	86%	100%	
Concerns related to care	71%	100%	
Medical problem list/Infobuttons	100%	100%	
Outpatient providers access to hospitalization plan of care	71%	100%	
Schedule	100%	100%	
Care team names/Pictures/Roles	86%	100%	14%
Messaging capability re: plan of care	86%	100%	
Safety dashboard/Infobuttons	71%	100%	
Test results/Infobuttons	100%	100%	28%
Medications/Infobuttons	100%	100%	28%
Food and Diet	100%	100%	
Discharge check list	86%	100%	
Did not answer	0%	0%	30%

Table 1: Patient/Family feedback related to Plan of care elements and workflow

5. *Meetings with clinical and administrative leaders and stakeholders:* Based on our meetings with clinical and administrator leaders stakeholders we learned that the 18 inch touch screen devices that we planned to install at the bedside were not feasible given the organizations plan to install bedside devices for providers as part of a larger electronic medical record implementation. We made a decision to go with mobile devices and to implement the PCTK on the iPad Air. As a result, we had to re-configure the user interface to optimize use on a mobile tablet device with a much smaller screen size. This new set of requirements led us to rework the user interface to accommodate the smaller screen size and to address the usability issues identified by patients, family, and professional stakeholders.

Discussion

We used an iterative participatory design approach to develop the plan of care and provider communication features of the PCTK for use with intensive care and oncology patients and family members. Participatory design is based on the classic principles of user centered design including: 1) Early focus on users and tasks; 2) Empirical measurement; and 3) Iterative design and development.[15] In our earlier work, we involved end users (patients and family members) and other stakeholders (professional care team members) in identifying the core set of information needed to engage patients in their plan of care.[9, 10] In this project, we continue to work with patients, family, and professional care team members to broaden our understanding of the information, tools and tasks associated with involving patients in the plan of care development process. Interestingly, through this approach, a set of information needs identified from our prior work was confirmed as useful by *all* patients and family members that participated in interviews: medical problem list, schedule, test results, medications, and food and diet. In addition to information needs identification, our approach enabled the identification of communication needs. We found that patients and family desire tools to help them communicate their goals, problems, concerns, and care preferences. They also would like to communicate directly with their professional care team members when questions and concerns about their plan of care arise. The rapid prototyping method was an effective way to involve patients, family and stakeholders in identifying usability problems and gaps in functionality that could be a barrier to use if not corrected. Our team has usability evaluation studies with *clinicians* to identify explicit information and communication needs in the context of a documentation workflow for goals of care[16]but to our knowledge this is the first study to define *patient and family* information and communication needs in the context of a documentation workflow for patient-centered care planning.

The initial prototype lacked specific tools to allow patients and family to document their personal goals, concerns, problems and preferences for care. It did include a space for patients and family to write questions that they could later share with the care team, but patients and family told us that this was insufficient; they wanted to directly communicate with their care team members. We added the care planning tools and a message board to the next prototype but soon learned from patients, family, and care team members that the user interface was too busy and that it would be difficult for older patients or patients who are not computer savvy to use. We simplified the user interface and the care planning tools in the next version of prototypes and have received positive feedback about patient and family intention to use the toolkit.

In addition to the feedback from stakeholders, our choice of hardware had an impact on the user interface. Specifically, changing from a mounted 18-inch touch screen device to an iPad Air forced us to reconfigure the interface from a horizontal to vertical tab, rearrange/remove content from the home screen, to rethink our navigation, and to leverage mobile optimized web-content. This was a difficult change to make but due to our ongoing communication with stakeholders, the issue was recognized early in the project when these types of changes could be made at a minimal cost.

In our most recent bedside interviews there was a limited sample of seven participants. Patients told us that they would use the care planning tools if they were available today and that they thought other patients would use it as well. Patients and family also reaffirmed the value of the additional information available from the PCTK such as pictures and roles of care team members, medications, test results, a discharge checklist, food and diet information, tailored information about their condition and their safety risks. Patients identified having access to their medications, schedule and laboratory results with Infobutton links to Medline Plus and other valid and reliable consumer literature as favorite features. One patient remarked, “I would like to have access to my medication

schedule when the nurses are giving me my medication because all of the medications have such strange names--- I like to keep track of what they are giving me". Another family member remarked, "This is useful because you know the first thing we do when we get home is Google what these things mean." We did have one patient who told us that although she could see how the PCTK could be useful for others that she would not use it. She said that she has the "right to ask for her paper chart" and that if she wanted information, she would request it. Given the emerging and mixed evidence of associations between engaged patients and a decrease in medication errors, these findings could inform our approach for measuring the impact of the PCTK on levels of patient engagement and adverse harm events.[17]

As we refine the PCTK software, we will continue to engage patients and family in using the software and in providing feedback on the user interface. As we have done in previous projects[18], we plan to implement the PCTK with our clinical nurse and physician champions and a subset of patients one month before go-live to ensure that we address any additional workflow or software issues before the tools are available to all patients on the oncology and intensive care units.

There are several limitations associated with this work. Our project is being conducted on a limited number of units in a single hospital. Our qualitative work has involved a limited number of patients and family and our findings may not reflect the views of other patients and family who have had similar or different experiences at BWH or at other hospitals. We consider ongoing patient, family and stakeholder involvement in the development of software as essential to designing software that is usable, useful and that will promote patient engagement in their plan of care. We will continue to engage patients and other stakeholders in this process as we further refine the PCTK software. In addition, we recognize that not all patients or family want to use computers or other devices to access information while in the hospital. Furthermore, some patients do not want to participate in developing their plan of care. The PCTK is meant to enhance current resources. A range of tools are likely needed to support hospitalized patients with different preferences and technical abilities.

Conclusion

While it is recognized that using technology to facilitate patient and family involvement in the plan of care could improve quality, safety and cost outcomes, their involvement to date is variable and evidence in this area is limited. Tools and processes are needed to provide hospitalized patients access to information about their condition and a set of tools to engage them in development of their plan of care. We found that many patients and family want to be knowledgeable about their condition and that they see value in tools that will help them to communicate their goals, problems, concerns, and care preferences. We also found that many patients, family members, and professional care team members are willing to participate in developing tools that will ultimately improve patient care and that they welcome the opportunity to make this contribution.

Acknowledgements

The authors would like to thank the BWH patients, family members, nurses, and physicians who provide ongoing feedback to inform development of the PCTK software. The BWH PROSPECT Project is part of the Libretto Consortium supported by the Gordon and Betty Moore Foundation.

References

1. Kripalani, S., et al., *Deficits in communication and information transfer between hospital-based and primary care physicians: implications for patient safety and continuity of care.* JAMA, 2007. **297**(8): p. 831-41.
2. Quint, J.C., *Communications problems affecting patient care in hospitals.* JAMA, 1966. **195**(1): p. 36-7.
3. Hibbard, J.H. and J. Greene, *What the evidence shows about patient activation: better health outcomes and care experiences; fewer data on costs.* Health Aff (Millwood), 2013. **32**(2): p. 207-14.
4. AHRQ. *20 Tips to Help Prevent Medical Errors: Patient Fact Sheet* 2011 [cited 2014 March 1]; Available from: http://www.ahrq.gov/consumer/20tips.htm.
5. HealthIT.gov. *Achieve Meaningful Use Stage 2*. [cited 2014 March 7]; Available from: http://www.healthit.gov/providers-professionals/achieve-meaningful-use/core-measures-2/use-secure-electronic-messaging.
6. Dykes, P.C., et al., *Fall prevention in acute care hospitals: a randomized trial.* JAMA, 2010. **304**(17): p. 1912-8.
7. Schnipper, J.L., et al., *Role of pharmacist counseling in preventing adverse drug events after hospitalization.* Arch Intern Med, 2006. **166**(5): p. 565-71.
8. Houston, T.K., et al., *Culturally appropriate storytelling to improve blood pressure: a randomized trial.* Ann Intern Med, 2011. **154**(2): p. 77-84.
9. Caligtan, C.A., et al., *Bedside information technology to support patient-centered care.* Int J Med Inform, 2012. **81**(7): p. 442-51.
10. Dykes, P.C., et al., *Building and testing a patient-centric electronic bedside communication center.* Journal of gerontological nursing, 2013. **39**(1): p. 15-9.
11. AHRQ. *Guide to Patient and Family Engagement: Environmental Scan Report*. 2012 [cited 2014 March 1]; Available from: http://www.ahrq.gov/research/findings/final-reports/ptfamilyscan/ptfamily1.html.
12. Berwick, D.M., T.W. Nolan, and J. Whittington, *The triple aim: care, health, and cost.* Health Aff (Millwood), 2008. **27**(3): p. 759-69.
13. Rogers, M.L., et al., *Usability Testing and the Relation of Clinical Information Systems to Patient Safety*, in *Advances in Patient Safety: From Research to Implementation: Concepts and Methodology*, K. Henriksen, et al., Editors. 2005: Rockville, MD.
14. Dabbs, A., B. Myers, and M. Dew, *User-Centered Design and Interactive Health Technologies for Patients.* Comput Inform Nurs, 2009. **27**(3): p. 175.
15. Gould, J. and C. Lewis, *Designing for usability: Key principles and what designers think.* Communications of the ACM, 1985. **28**(3): p. 300-11.
16. Collins, S.A., et al., *Model development for EHR interdisciplinary information exchange of ICU common goals.* Int J Med Inform, 2011. **80**(8): p. e141-9.
17. Berger, Z., et al., *Promoting engagement by patients and families to reduce adverse events in acute care settings: a systematic review.* BMJ Qual Saf, 2014.
18. Dykes, P., et al., *Fall TIPS: strategies to promote adoption and use of a fall prevention toolkit.* AMIA Annu Symp Proc, 2009. **2009**: p. 153-7.

Evaluation of need for ontologies to manage domain content for the Reportable Conditions Knowledge Management System

Karen L. Eilbeck, PhD[1], Julie Lipstein, BA[2], Sunanda McGarvey, BS, CBAP[3], Catherine J. Staes, BSN, MPH, PhD[1]
[1] University of Utah, Salt Lake City, UT; [2] L-3 STRATIS, Atlanta, GA; [3] Northrup Grumman Information Technology, Atlanta, GA.

Abstract

The Reportable Condition Knowledge Management System (RCKMS) is envisioned to be a single, comprehensive, authoritative, real-time portal to author, view and access computable information about reportable conditions. The system is designed for use by hospitals, laboratories, health information exchanges, and providers to meet public health reporting requirements. The RCKMS Knowledge Representation Workgroup was tasked to explore the need for ontologies to support RCKMS functionality. The workgroup reviewed relevant projects and defined criteria to evaluate candidate knowledge domain areas for ontology development. The use of ontologies is justified for this project to unify the semantics used to describe similar reportable events and concepts between different jurisdictions and over time, to aid data integration, and to manage large, unwieldy datasets that evolve, and are sometimes externally managed.

Introduction

Public health surveillance, investigation and intervention are essential for the prevention and control of communicable and non-communicable diseases. Toward that end, every jurisdiction in the U.S. publishes a list of "reportable conditions" that function as a communication tool between public health entities and reporting entities. When a reportable condition is identified, reporters (e.g., hospitals, laboratories) are required to report the event to public health authorities [(1-5)]. Timely and accurate reporting is critical to identify, investigate, and control public health threats.

Accurate, timely and complete reporting of reportable conditions depends on reporters having correct and current information about a) the list of conditions that are reportable for the jurisdiction where they or their patients or clients are located, b) criteria that make a condition reportable, c) how quickly a report should be sent, d) what information should be included in a report, e) the preferred method to send a report, and f) where the report should be sent. This is referred to as the "who, what, when, where, and how" of reporting. Automated detection and electronic reporting depends on having this information available in machine-processable form to be stored as rules within a clinical decision support system, and used by an EHR or LIMS. Currently, this information is scattered across many different websites and documents[(2)]. It is difficult for human users to navigate, and the information is not usually presented in a manner suitable for machine-processing. The Reportable Condition Knowledge Management System (RCKMS) is positioned to provide a solution for this challenge by providing a single entry point to manage and access reportable condition specifications.

The RCKMS project is managed using a phased approach and is supported by the Centers for Disease Control and Prevention (CDC). The current phase (2014-2015) targets an authoring framework and administration screens that will capture reporting specifications, transform them to executable rules, and store them in an open source clinical decision support tool. Prior work included a view/query interface, accompanying printable reports, a user login/profile management interface, and subscription management capability. Jurisdictional reporting specifications were collected from six jurisdictions for three conditions, and default (or base content) was defined for 13 conditions, along with recommendations for resolving underlying issues in the Council of State and Territorial Epidemiologist (CSTE) Position Statements[(6)] that limit their use in defining computable criteria.

The RCKMS project promotes adherence to standards and reuse of existing authoritative resources. In keeping with this philosophy, the team participated in the Office of the National Coordinator's (ONC) Standards and Interoperability (S&I) Health eDecisions (HeD) pilot as an artifact provider. For the HeD pilot, an output file of reporting specifications was produced for pertussis reporting requirements that was both human-readable and machine-processable. Project members participate in other S&I initiatives related to the reporting lifecycle (e.g., Public health reporting Initiative, Data Access Framework, and Structured Data Capture), and the project will

continue to assess adoption options as standards evolve. Finally, RCKMS will use, not replace, existing value set repositories managed by the CDC (e.g., PHINVADS[7]) or the National Library of Medicine (e.g., VSAC[8]).

Objective

Given the breadth and complexity of the data needed to automate the reporting process on a national level, the RCKMS Knowledge Representation Workgroup was tasked to a) explore the need for ontologies to support RCKMS functionality, and b) develop recommendations for ontology development if indicated.

Methods

The RCKMS Knowledge Representation workgroup convened weekly to bi-monthly by phone and web conference from April through September 2013. The workgroup included members from various backgrounds representing those a) who may author content or query for and use the system output; b) who have developed knowledge management systems or components that may address parts of the overall problem; and finally, c) ontology developers from outside of the reportable condition domain.

The workgroup represented a diverse set of members from within and beyond the reportable condition community, enabling a wide variety of viewpoints to be expressed. The processes and resources currently involved with public health reporting, available tools, other exploratory projects, and the prototype RCKMS were presented to the workgroup. Meetings were recorded for those who missed them, and consensus was gathered iteratively through discussion and email. The resulting artifacts provide a comprehensive survey into the kinds of knowledge management needed to automate much of the currently manual processes.

Specifically, the workgroup participated in the following tasks:

1. *Evaluate Content:* The workgroup reviewed a variety of topics to understand the scope of content required and available for RCKMS, including the existing pilot application, other ontology and knowledge representation projects, related national initiatives, existing terminologies applications, and research that may contribute toward solving the defined problems. The workgroup surveyed the domain of public health reporting and described and evaluated the status of resources necessary for the RCKMS project.
2. *Define user stories and identify required data resource linkages:* The workgroup reviewed user stories submitted for the following stakeholders groups: Public Health epidemiologists at all levels of public health, knowledge curators or terminologists, clinicians, researchers, electronic health record vendors, laboratorians, and public health officials. The user stories were grouped according to the area of RCKMS affected. Next, the workgroup identified the need to represent linkages between resources necessary to meet the requirements observed in the user stories.
3. *Assess candidate knowledge domains for ontology development:* The following questions were developed by the workgroup and used to evaluate whether ontologies were needed to manage one or more of the sub-domains of information in RCKMS. Selected sub-domains were identified as candidates for ontological representation when the questions below were affirmative.

    - How complex are the relationships between the data?
        - How many different types of relationships exist in the area under consideration?
        - Is there a standard discernible structure that exists between the data?
            - Is there an inheritance of qualities between pieces of data?
            - Are there conflating ideas in the data that should be separated?
    - Is there a logical structure to the data?
    - Are users interested in questions that can more easily be answered if the data relationships are ontology-based?
    - Does the data change over time?
        - Are new concepts added that need to be classified?
        - Is the structure of the data rearranged over time?
    - Do we need to standardize the way more than one group describes or uses the data to share a common understanding?
    - Are there concerns about the quantity and maintenance of the data?
        - Is there sufficient data to warrant the time, effort and cost to implement an ontology?
        - Is maintenance of the data a manual burden/bottle neck?

    - Is the speed of response to change an issue?
- Does reasoning need to be done against the data structure?
    - Is there a need to enable tooling to validate rules?

4. *Develop recommendations:* Recommendations to advance the integration of ontologies into the RCKMS project were drafted based on input from the group, and refined iteratively over a course of meetings and comments submitted by workgroup members.

Results

Evaluate content

The workgroup identified domains of knowledge necessary for the RCKMS project and existing external coded resources for coded concepts represented in RCKMS. For each domain of knowledge, we described existing data sources, problems and recommendations for use and incorporation, and the relevance for RCKMS.

The workgroup identified the following three knowledge resources for RCKMS, representing data that needs to be managed for the project:

- *State reporting rules:* Reporting rules are available in HTML or PDF documents on state/county/city/tribal websites. While this information is authoritative, it is not computable. In addition, under the current paradigm, reporters are not notified of changes so it is time consuming and inefficient to remain current. Within RCKMS, there will be an authoring interface that will allow jurisdictions to manage their reporting specifications and make this information available in both human-readable and machine-processable formats. RCKMS will also support notification of changes to be sent to reporters who subscribe to receive changes for conditions or by jurisdictions.

- *Reporting logic for national surveillance:* The logic is available in Position Statements published as PDF documents on the CSTE website[(6)]. While this information can be used as a default to encourage uniformity, the reporting logic in the Position Statements is not uniformly adopted across jurisdictions. A detailed CTSE report concerning the content is available from the Public Health Data Standards Consortium[(9)]. Within RCKMS, criteria described in the Position Statements can be used as base content that can be copied and directly used by a jurisdiction, or modified to address jurisdiction-specific requirements.

- *Nationally notifiable conditions*: Once conditions are reported to a state or other jurisdictional public health agency, case definitions are used to determine which reports should be forwarded from the 'local' public health agency to the CDC for national surveillance. The logic for defining a confirmed or probable case (i.e. case definitions) for each condition tracked on a national level is available in HTML on the CDC website[(10)]. In addition, the coded concepts for these nationally notifiable events and other selected relevant value sets of coded concepts are available on the PHIN VADS website managed by the CDC[(7)]. While the nationally notifiable conditions are similar to reportable events, they do not include all conditions and criteria for reporting in a given jurisdiction. In addition, the conditions change over time, but the reasons for change and relationship between current and previous conditions are not captured.

The workgroup identified concepts in RCKMS that can be represented using coded values from existing external resources. For example, the following major domains of concepts, resources for coded values and key requirements were identified by the workgroup:

- *Clinical condition* (the name of a disease or health condition under surveillance)
    - Coded values: Subsets from SNOMED-CT® /ICD-9 CM/ ICD-10 CM
    - Requirement: Allow jurisdictional reportable conditions to link to clinical condition.
- *Lab test names* (described as test names and often test methods)
    - Coded values: LOINC®
    - Requirement: Support default value sets for lab test names, and allow jurisdictions to update these to meet jurisdiction-specific needs.
- *Lab test results* (specifically for results that are organisms or positive or negative)
    - Coded values: SNOMED-CT® for organisms
    - Coded values: SNOMED-CT® or HL7/PHIN VADS codes for positive/negative value sets

    - Requirement: Support default value sets for lab test results, and allow jurisdictions to update these to meet jurisdiction-specific needs.
- *Clinical observations*
    - Coded values: LOINC®
    - Requirement: Align with the new agreement between LOINC® and SNOMED®(11)
- *Clinical values/findings*
    - Coded values: SNOMED-CT® concepts for nominal and ordinal values
    - Requirement: Support default value sets for clinical findings, and allow jurisdictions to update these to meet jurisdiction-specific needs.
- *Jurisdiction context* (Name of jurisdiction and designation of geographic coverage)
    - Coded values: Federal Information Processing Standard Code (FIPS)
    - Requirement: Flow of reporting within jurisdictions must be represented to assist reporters in identifying to whom a report should be sent, for example, if it should be sent to a local health department, the reporter, must be able to determine which local health department and if it is based on the residence of the patient, the site of care delivery, or the location of the servicing laboratory.

Define User Stories and identify required data resource linkages

The workgroup reviewed user stories provided by the following stakeholders groups: epidemiologist working at a state or local health department, CDC epidemiologist, knowledge curator/terminologist, reporter, researcher, EHR vendor, laboratorian, and public health official. We grouped user stories according to the area of RCKMS that may be affected by the user story. Each set of stories includes a variety of stakeholders but similar data structure needs.

1. *User stories addressing relationships between reportable events by jurisdiction:* Story example: "I am the epidemiologist for the communicable diseases branch of my health department and I would like to see the reporting specifications for all conditions in our neighboring jurisdictions to compare against ours. In order to compare the criteria for a given reportable topic, there needs to be a way to link similar reportable events with different names or different criteria for the topic." Resource example: Show the criteria for reporting pertussis-related events from a lab or clinical setting for the following spatial contexts: Utah and Colorado. The system would need to know that pertussis and whooping cough were the same.

2. *User stories that require understanding about the relationship between semantic changes in reportable events.* Story example: "I am an epidemiologist defining a new or updating a reportable event and need to decide which subset of a condition should be made reportable. To improve consistency across jurisdictions, I want to search through the RCKMS to see related conditions from other jurisdictions and be able to view hierarchical relationships across related conditions." Resource example: In Utah, the condition 'Hantavirus infection and pulmonary syndrome' was retired and replaced by 'Hantavirus pulmonary syndrome'. In Colorado however, the change has yet to be made. What is the relationship between the two terms and the nationally notifiable condition used for national surveillance?

3. *User stories related to linking national surveillance to reportable events.* Story example: "I am the CDC epidemiologist with the National Notifiable Disease Surveillance System (NNDSS) and I need to know if states are collecting information, in other words does the state include relevant reportable events in their jurisdiction to assert that a given state is reporting Nationally Notifiable Conditions (NNCs) for their jurisdiction. Currently, this knowledge is manually derived from CSTE's State Reportable Condition Assessment (SRCA), which is retrospective and difficult to interpret." Resource example: Each year, CDC's NNDSS needs to know if each nationally notifiable condition is or is not reportable in each jurisdiction. This requires knowledge of each reportable condition in each jurisdiction and to understand whether or not it should map to a specific nationally notifiable condition for the surveillance year.

4. *User story that combines semantic changes over time and national surveillance.* Story example: I am a researcher reviewing notifiable condition reports for Shiga toxin-producing Escherichia coli over the last 10 years. I want to take into account changes in naming and classification at the national level as well as differences in how jurisdictions report the disease so my data is inclusive of all the reports that are applicable. Resource example: Three notifiable conditions (*Enterohemorrhagic Escherichia coli (EHEC) shiga toxin+ (serogroup non-O157), Enterohemorrhagic Escherichia coli (EHEC) O157:H7, and Enterohemorrhagic Escherichia coli (EHEC) shiga toxin+ (not serogrouped))* were retired in 2006 and replaced by the supertype *Shiga toxin-producing Escherichia coli (STEC)*.

5. *User stories related to criteria and mappings to code systems.* Story example: "I am a terminologist and I need to select the LOINC® codes that meet the epidemiologist's description of criteria using the organism, method and specimen source. How do I find all the relevant codes, and how do I reselect the codes as new versions of LOINC® are released." Resource example: What lab test codes should be included in the value sets corresponding to the selection logic for tuberculosis where the epidemiologist has specified a target organism, laboratory method and specimen types to be used in the logic?

6. *User story that extends #5 and combines with earlier user stories.* Story example: "I am a terminologist, and a reportable condition is being updated into two reportable conditions. I need a report to show all vocabulary that was referenced by the condition so I can determine how the references may need to be updated based on the split (or combination, or reclassification)." Resource example: Ehrlichiosis, human monocytic (HME) is being replaced by two new events: *Erlichia ewingii* and *Ehrlichia chaffeensis*.

The workgroup then reviewed an existing concept map previously created by one of the authors (CJS) that shows many of the concepts represented in reporting requirements (Figure 1). The concepts were extracted from the websites and PDF's described above, and were organized to be illustrative. The major concepts of interest are shaded. The concept map was useful for illustrating subdomains of content and some of the linkages required between concepts. The workgroup articulated several relationships between concepts, some of which are shown in the additional blue boxes (Figure 1).

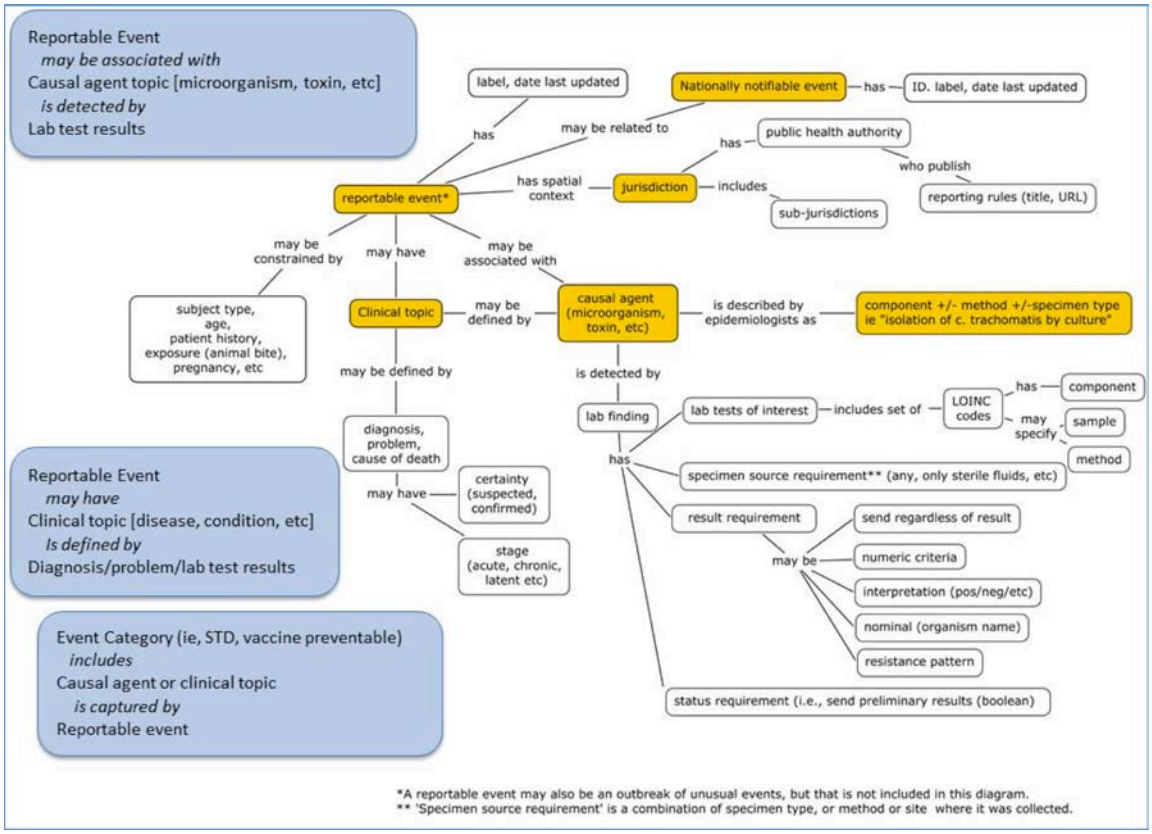


Figure 1. Map of selected concepts needed to represent reporting requirements and identify candidate areas for use of ontology to manage RCKMS knowledge (Key concepts shaded. Newly-identified relationships in boxes).

There are existing resources such as SNOMED-CT that link diseases and causative agents, but most of the linkages required for clinical and laboratory reporters to know 'what", "when", "where" and "how" to report to jurisdictional public health authorities do not exist in computable format sufficient to meet the user stories. For example, there are no existing resources to support queries based on a user's need to see reportable events related to a diagnosis ("show me all the influenza reportable events, such as the criteria for pediatric deaths or influenza hospitalizations"), or jurisdiction ("show me what is required to be reported in New York City and New Jersey"). In addition, while a

Reportable Condition Mapping Table was published in 2011 that lists LOINC® and SNOMED-CT concepts for Electronic Laboratory Reporting, the mappings were created through a manual, labor intensive process that requires manual updates as source data is updated, and the Mapping Table does not represent state reporting requirements[12].

Assess candidate knowledge domains

Using the criteria defined in the methods, the workgroup evaluated three subdomains of information within RCKMS. Ontologies were considered as tools to manage the following knowledge: reportable events, jurisdiction relationships, and reporting logic. The ontologies would be used for both standardizing the authoring of content and for querying RCKMS data. The findings are summarized in Table 2. The assessment illustrates a strong need for the use of ontologies to manage information about reporting logic and reportable events. There is not strong evidence for a need to manage jurisdiction information in an ontology. The hierarchical structure of jurisdictional information can be managed using mapping tables.

Table 2. Evaluation of candidate domain areas for ontology development based on criteria for ontology use.

Sub-domains of RCKMS knowledge		
Reportable event	**Jurisdiction**	**Reporting logic (particularly laboratory-based logic)**
1. How complex are the relationships in the data?		
Complex relationships between reportable events, conditions and reporting criteria. Also, need to track semantic changes over time.	Simple relationships	Complex relationships between selection criteria and value sets.
1(b). Is there an inheritance of qualities between the data?		
Within a single jurisdiction, events tend to be mutually exclusive. When attempting to aggregate across jurisdictions or over time, inheritance becomes an issue	Cities/Counties inherit state-based rules and sometimes include additional reporting rules of their own.	Lab tests have inherent hierarchical structure.
2. Is there a logical structure to the data?		
Yes, events are about conditions, and have defining criteria. Events evolve over time and have relations to previous events. Jurisdictional events can be related to national events.	Yes. Spatial structure and reporting flow structure.	Yes
3. Are users interested in questions that can more easily be answered if data relationships are ontology-based?		
Yes - see user stories 1,2,3,4,6.	No user story described a need.	Yes - see user story 5.
4. Does the data change over time?		
Yes, jurisdictional and national reportable events change yearly (or more frequently).	Stable	Yes, updates to reporting criteria occur yearly or more frequently. LOINC® files are updated every 6 months
5. Are there multiple user groups, and need for a common understanding?		
Yes. Authoring and reporting users. Also, national and jurisdictional reportable event developers.	Yes. Authoring and reporting users.	Yes. Authoring and reporting users. Multiple other uses for value set creation.
6. Is maintenance of the data a manual burden and creating a bottle neck?		
Yes, Large manual burden. Currently, the State Reportable Condition Assessment[13] provides some capacity, but only retroactively with several year delay. No logical relationships used to manage tracking of event development, or mapping between national and jurisdictional events	No. Small burden	Yes. Large manual burden. Updates to selection criteria or underlying data sources (i.e., LOINC®) result in large manual undertaking to produce new value sets.
7. Does reasoning need to be performed to utilize the knowledge?		
Yes, traversal of terms in a well-developed ontology could aid in surveillance questions.	Reasoning can be performed over existing spatial relations	Yes, tests can be described logically and reasoned into

Reasoning may be valuable.	(e.g.,"What events should be reported for X city in Y county in Z state?")	hierarchical structure for management of value sets.

Recommendations and Discussion

The workgroup evaluated the knowledge needed and the resources currently available to effectively address the user stories and necessary linkages between resources. Two areas of the reportable condition domain were agreed upon by the workgroup as candidates for ontology development. The undertaking of ontology development for a large project like RCKMS is not to be taken lightly and the workgroup made recommendations regarding implementation and ontology management.

Recommendations for domains to be managed using ontologies

The following areas of RCKMS knowledge were recommended to be managed using ontologies.

- *Reportable Events:* An ontology of reportable events would allow RCKMS to meet the surveillance and querying requirements. The concepts to be managed by this ontology are the events, the relationships between active and retired events, the relationship between national and state-defined events and the criteria associated with given events. A reportable event ontology is a complex undertaking that involves management of concepts from the national and jurisdictional level. This is of high priority. The surveillance use cases require RCKMS to be able to track conditions over time, but the events that capture the conditions change. The querying use cases need to be able to evaluate what is currently reported, with what was previously reported, and to do so across changes at the state/local jurisdiction level, as well as at the national level. Therefore, the ontology would need to relate the reportable events to their criteria, and how they relate to each other. A key component of this reportable event ontology will be defining the relationships that exist between terms such as *replaced by*, that allow us to understand the evolution of the usage of a condition.
- *Reporting Logic:* An ontology to manage reporting logic, particularly to select value sets using LOINC®, would provide consistency when authoring RCKMS content. The base content for this development may come from information already gathered by CSTE, including Position Statements and their associated technical implementation guides, and information gathered from the State Reportable Condition Assessment (SRCA). The technical implementation guides were a onetime effort and have not been updated since 2008 to reflect updates that may have occurred to CSTE Position Statements for Notifiable Conditions. After the initial population of terms, there will be a phase of jurisdictional curation. The content will need periodic updating as the referenced datasets (e.g., LOINC® are updated, and reporting logic sets are updated by each jurisdiction. These updates will be jurisdiction specific. Previous work[14-16] successfully demonstrated the use of ontologies to hierarchically query the LOINC® database, and highlighted issues that must be resolved: 1. LOINC® cannot effectively be queried hierarchically in its present form. 2. The component axis of LOINC® is not suited for epidemiological queries. 3. There are data restriction issues relating to public access of the hierarchical structure in LOINC®

Recommendations for ontology management

- *Allocate adequate resources to support the stages of ontology development***.** The first stage; exploration of necessity, has been completed by this workgroup. The use of ontologies is justified for this project to unify the semantics used to describe similar concepts between different groups, to aid data integration, to allow logical inference over data and to manage large, unwieldy datasets. Ontology development needs a team approach: It is a specialist task and requires both domain knowledge and understanding of ontology technology and the existing data infrastructure and the semantics of the data. Creation of ontologies will be undertaken by knowledge curators with the assistance of technologists. The knowledge curators will be domain experts in the field of public health, who have some training in the ontology development. The technologists will be familiar with ontology development and application, and be trained in the use of an ontology management system. They will also be familiar with databases and the existing data. They will utilize input from a variety of sources to drive development (e.g., spreadsheets, textual information). To ensure that the ontologies accurately represent the semantics of the domain, a process of oversight and curation will be utilized. Testing and deployment tasks will be undertaken by the technologists and IT team. Maintenance will be vital to continue to support the community after development and deployment. This will involve updates coming from external data sources such as LOINC® and updates from public health data (national and jurisdictional). Maintenance should be less labor intensive, but longer running than the initial development.

- *Use a hybrid approach to ontology development for the RCKMS.* There are three high level strategies for ontology development: top-down, bottom-up and hybrid; each approach has pros and cons and is suited to a different entry point in the process of elucidating and codifying ontologies. *Top-down* ontology development starts with defining and then extending top level core concepts, from very general terms, to very specific. This approach forces the developer to identify the key foundational concepts and their relationships from the beginning, and often leads to a comprehensive ontological model of the data. Existing foundational ontologies may be used (for example, Basic Formal Ontology, BFO[17]) to form the basis of the domain ontology. This method is good for consensus building among multiple groups. *Bottom-up* approaches to ontology building use pre-existing data sources as the starting point of development and take the reverse approach of specific to general term development. This is well suited for data integration projects, but often produces an incomplete model of the domain. The *hybrid approach* takes a pragmatic approach to the modeling problem – in that it uses both of the previous approaches to come to a consensus view. It is suitable for projects where there is need to harmonize the semantics used by the community and help provide conversion to heterogeneous datasets. The problems being addressed by RCKMS are wide, and there are elements that are best suited to both top-down and bottom-up development. There are many existing resources used in this project, such as each jurisdiction's reportable event criteria, that may best be handled with a bottom up approach. It would however, be of great benefit to this community to develop a comprehensive domain ontology that is rooted by a foundational ontology. There is also evidence that suggests that using a foundational ontology does not slow development and improves interoperability and quality [18].

  The ontologies developed for this project have the potential to provide a valuable resource to other investigators in the wider biomedical community. To enable and promote reuse, it is recommended that an upper ontology be used to place the terms developed into a wider context. An upper ontology such as BFO provides very high level terms upon which more specific ontologies can be developed. For example, there is a division between continuants and occurants, and between dependent and independent continuants. This upper level specification of the world forces the specific ontology developer to frame the new terms with the organizing principles in mind. The agreement imposed by the upper level ontology promotes reuse and interoperability of ontology terms. Providing consistency in the definitions of relationships used in the ontologies enables inference and reasoning over the ontologies. The Open Biomedical Ontology group's Relations Ontology [19] provides a starting place to obtain and then extend relations for the ontologies built for RCKMS. Another option to promote the reuse of the ontologies developed here would be to host them on the National Center for Biomedical Ontology's BioPortal website [20].

Recommendations for ontology development

The workgroup had specific recommendations regarding the management of ontology development: It is recommended that the ontology development be transparent, open source and versioned, preferably with a versioning system such as SVN and hosted via a collaborative site such as Google Code.

The project should provide a system to capture change requests to the ontologies. It is envisioned that there will be two mechanisms for development and change to the ontologies. The first is batch sets of changes relating to the updates of other datasets (new terms in LOINC®, yearly updates to selection logic by a jurisdiction). The second mechanism for change is via a developer working in a given knowledge area discovering the need for a new concept. These term requests should be tracked and managed in a way that is transparent and traceable. Many ontology projects use a term request tracker to document this process.

Where applicable, existing ontologies and knowledge representations should be used. The necessary terms can be imported into the RCKMS ontologies and strategies need to be in place for updates of existing terms. Some tools such as Ontofox [21] exist for Web Ontology Language (OWL) ontologies that allow this process to be managed.

The knowledge captured by the RCKMS project has the potential of being vast and complex. Large ontologies are not uncommon in the biomedical domain. Bada et al. evaluated the Gene Ontology for elements of success and concluded that **clear goals**, **simple intuitive structure**, **early use**, **community engagement**, and **continuous evolution** were the key factors [22]. The following questions have been addressed to find solutions that will work for the particular requirements of RCKMS:

- *Clear Goals***:** One large all-encompassing ontology or several small ontologies? A practical solution to the knowledge management issues presented to this workgroup has led to the delineation of two areas for focused ontology development within the scope of RCKMS. They include:

a) An ontology about reportable events
b) An ontology to query LOINC® and develop value sets for selection logic

These two areas are sufficiently distinct from each other to warrant separate ontologies, developed by expert content leaders.

- *Intuitive structure: Separate ontology and data.* There are several examples in biomedicine of large ontology projects where the data is captured separately from the ontology; the Gene Ontology Consortium [23] being a good example. The ontology development is managed independently from the annotation of the ontology terms to the data. In this way, a few developers maintain control of the content of the ontology while annotators manage the data. This is a good model for the RCKMS where a small number of terminologists or content managers will be responsible for the terminology, but a larger group of interested parties will use the ontology – for example with the authoring use cases. This proposed work must capture the knowledge necessary to automate the processing of public health data. This knowledge falls into several domain areas. While it is possible to incorporate ontologies into systems, and reason over and query them on the fly, it is also equally feasible to manage the knowledge separately in an ontology, and export the reasoned statements from the ontology to the system. Separation of the ontology from the system has the benefit of easier maintenance, less software complexity, quicker response time when queried, and would offer a quicker path to operationalize within the current architecture.
- *Community engagement.* Key stakeholders (in the downstream use of the knowledge) must be involved in the development of the ontologies.
- *Early use.* RCKMS should provide ontologies and annotations to public health community for use as early as possible. The ontologies and systems developed should be promoted and disseminated to the wider community, while providing a forum for feedback. There must be a visible, accessible website for RCKMS documentation, including links to the ontology resources and documentation to allow community use and feedback. The ontology should be released in formats digestible by current tools.

Conclusion

Currently the management of knowledge surrounding reportable conditions is a manual task, for both those providing the rules and those interpreting the rules. The tasks of authoring are manually driven, as base content for conditions is not maintained electronically. Surveillance of conditions over multiple jurisdictions is an onerous manual task, as there is no connection between the national reportable event, and what is reported locally. Reporting has a large manual component as the jurisdictional rules must be interpreted and implemented locally – requiring each reporter, for example to decide which positive LOINC® tests to report for a given condition. Given that the content of external resources and reporting rules change periodically, the amount of manual mapping and updates is considerable. The RCKMS aims to provide efficient, automated solutions to these problems.

Complex information is required to support RCKMS and provide computable, viewable and usable information for reporters to know what, where, when and how to report to public health. Ontologies have been applied successfully to manage the knowledge of other large domains within the biomedical informatics community, and the workgroup concluded would be an appropriate tool for to manage reportable condition knowledge. Ontologies are indicated for two specific domains of content: selection logic criteria and reportable events. The undertaking of ontology development for a large project like RCKMS is not to be taken lightly, and ontology development should be transparent, open source, and versioned using strategies that leverage existing ontologies and allow reuse of the knowledge developed for the RCKMS effort.

Acknowledgements

The RCKMS project is funded by the CDC, Office of Public Health Scientific Services (OPHSS). We would like to thank members of the RCKMS Knowledge Representation Working group for voluntary participation in weekly discussion during 2013. Members include Sundak Ganesan, Anna Orlova, Austin Kreisler, Nikolay Lipskiy, Arun Srinivasan, Jeff Kriseman, Cecil Lynch, Scott Keller, Jerry Sable, Sheila Abner, Ted Klein, Ruth Ann Jajosky, Mary Hamilton, and Heather Patrick. In memoriam of Cynthia Vinion who contributed use cases to our analysis.

References

1. Chorba TL, Berkelman RL, Safford SK, Gibbs NP, Hull, HF. Mandatory Reporting of Infectious Diseases by Clinicians. JAMA 1989;262(21):3018-26.
2. M'ikanatha NM, Welliver D P, Rohn DD et al. Use of the Web by State and Territorial Health Departments to Promote Reporting of Infectious Disease. JAMA. 2004;291(9):1069-71.
3. Freund E, Seligman PJ, Chorba TL, Safford SK, Drachman JG, Hull HF. Mandatory Reporting of Occupational Diseases by Clinicians. JAMA. 1989;262(21):3041-4.
4. Roush S, Birkhead G, Koo D, Cobb A, Fleming D. Mandatory Reporting of Diseases and Conditions by Healthcare Professionals and Laboratories. JAMA. 1999;282(2):164-70.
5. Jajosky R, Rey A, Park M, Aranas A, Macdonald S, Ferland L. Findings from the Council of State and Territorial Epidemiologists' 2008 Assessment of State Reportable and Nationally Notifiable Conditions in the United States and Considerations for the Future. J Public Health Manag Pract. 2011;17(3):255-64.
6. Position Statements [Internet]. 2014. Available from: http://www.cste.org/?page=PositionStatements.
7. CDC. PHIN Vocabulary Access and Distribution System Atlanta: CDC; 2011 [cited 2014 June 7]. Available from: https://phinvads.cdc.gov/vads/SearchVocab.action.
8. NLM. NLM Value Set Authority Center (VSAC) 2014 [cited 2014 July 27]. Available from: https://vsac.nlm.nih.gov.
9. Lipskiy L, Orlova A, Klein T, Huang M, Huang G, Minami M, et al. Assure Health IT Standards for Public Health: Enable Electronic Detection of Reportable Conditions Through Improved Codification of Public Health Reporting Criteria. Public Health Data Standards Consortium, 2014.
10. CDC. National Notifiable Diseases Surveillance System (NNDSS) 2014 [cited 2014]. Available from: http://wwwn.cdc.gov/nndss/.
11. IHTSDO. SNOMED CT and LOINC to be linked by cooperative work 2013. Available from: http://www.ihtsdo.org/about-ihtsdo/governance-and-advisory/harmonization/loinc.
12. CDC. Reportable Condition Mapping Table (RCMT) Another step toward standardizing electronic laboratory reporting (ELR) 2014 [cited 2014 July 27]. Available from: http://www.cdc.gov/EHRmeaningfuluse/rcmt.html.
13. Council of State and Territorial Epidemiologists. State Reportable Conditions Website 2011 [cited 2014 June 5]. Available from: http://www.cste2.org/izenda/entrypage.aspx.
14. Eilbeck K, Jacobs J, McGarvey S, Vinion C, Staes C, editors. Exploring the use of ontologies and automated reasoning to manage selection of reportable condition lab tests from LOINC®. International Conference for Biomedical Ontology; 2013; Montreal.
15. Eilbeck K, Jacobs J, Staes CJ, editors. Optimize Querying of LOINC with an Ontology: Give Me the Chlamydia Tests the Epidemiologists Want Me to Use! 46th Hawaii International Conference on System Sciences (HICSS); 2013 7-10 Jan. 2013; Hawaii.
16. Adamusiak T, Bodenreider O. Quality assurance in LOINC using Description Logic. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2012;2012:1099-108.
17. Grenon P, Smith B, Goldberg L. Biodynamic Ontology: Applying BFO in the Biomedical Domain. In: Pisanelli DM, editor. Ontologies in Medicine; Amsterdam: IOS Press; 2004. p. 20–38.
18. Keet M. The Use of Foundational Ontologies in Ontology Development: An Empirical Assessment. Lecture Notes in Computer Science. 2011;6643:321-35.
19. Smith B, Ceusters W, Klagges B, Kohler J, Kumar A, Lomax J, et al. Relations in biomedical ontologies. Genome biology. 2005;6(5):R46.
20. Whetzel PL. NCBO Technology: Powering semantically aware applications. Journal of biomedical semantics. 2013;4 Suppl 1:S8.
21. Xiang Z, Courtot M, Brinkman RR, Ruttenberg A, He Y. OntoFox: web-based support for ontology reuse. BMC research notes. 2010;3:175.
22. Bada M, Stevens R, Goble C, Gil Y, Ashburner M, Blake JA, et al. A short study on the success of the Gene Ontology. Journal of Web Semantics. 2004;1(2).
23. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics. 2000;25(1):25-9.

The Number Needed to Remind: a Measure for Assessing CDS Effectiveness

Jonathan Einbinder MD, MPH[1,2,3], Esteban Hebel, MD, MMSc[1,4], Adam Wright, PhD[1,2,3], Morgan Panzenhagen[3], RN, Blackford Middleton MD, MPH, MSc[1,2,3]

[1]Clinical Informatics Research and Development, Partners Healthcare Systems, Wellesley, MA; [2]Harvard Medical School, Boston, MA; [3]Massachusetts General Hospital, Boston, MA; [4]German Clinic, Santiago, Chile.

Abstract

Background: Clinical decision support (CDS) is associated with improvement in quality and efficiency in healthcare delivery. The appropriate way to evaluate its effectiveness remains uncertain.

Methods: We analyzed data from our electronic health record (EHR) measuring the display frequency of eight reminders for Coronary Artery disease and Type 2 Diabetes and their associated performance according to a predefined methodology. We propose two key performance indicators to measure their impact on a target population: the reminder performance (RP), and the number needed to remind (NNR), to evaluate the impact that Clinical decision support reminders have on the adherence to guideline derived CDS interventions on the entire patient population, and individual providers receiving the interventions.

Results: Data were available for 116,027 patients and a total of 1,982,735 reminders were displayed to a subset of 65,516 patients during the study period from January 1 to December 31, 2010. The evaluation framework assessed provider acknowledgement of the CDS intervention, and the presence of the expected performance event while accounting for patients' exposure to the CDS reminders. The total RP was 2.7% while the average NNR was 3.1 for all the reminders under study.

Conclusions: The proposed framework to asses of CDS performance provides a novel approach to improve the design and evaluation of CDS interventions. The application of this methodology represents an indicator to understand the impact of CDS interventions and subsequent patient outcomes. Further research is required to evaluate the impact of these systems on the quality of care.

1 Introduction

Meaningful use of health information technology is viewed as essential for effecting change in healthcare delivery.[1] Effective use of clinical decision support (CDS) is one of the components that help physicians and other health care providers treat patients according to evidence-based guidelines for care.[2–5]

Reminders are one common type of CDS usually triggered by patient data or information entered by the user. They are intended to prompt the healthcare provider about the appropriate interventions or to avoid certain actions to improve the patient individual care.[6] In the literature, reminders have been found to improve some preventive practices and compliance with clinical guidelines in regard to medication selection and diagnostic testing.[7]

However, some research has found that reminders achieve modest or no improvement in care.[6,8–10] In addition, current research has suggested that an overabundance of reminders may counteract their effectiveness and lead to user dissatisfaction.[11,12] Our approach focuses on the expected actions in preventive care that can be attributable to the reminder being displayed and acted upon from a population perspective. Existing evidence on the topic is conflicting and understudied.[13]

In research conducted by the CDS Consortium[14,15], we developed a CDS Dashboard to inform end users as to their use of decision support and compliance with CDS recommendations, and implemented it at the Partners Healthcare System. The CDS Dashboard also provides feedback to the research team about CDS performance characteristics, including usage and compliance rates, user performance for key metrics, compared to other users or reference benchmarks.

The purpose of this paper is to provide a better understanding of population based CDS performance measurement, to identify best practices for designing and implementing CDS, and to introduce two new quality measures, titled Reminder Performance (RP) and the Number Needed to Remind (NNR) for evaluating the effectiveness of clinical reminders in the context of the CDS Dashboards.

2 Methods

2.1 Study Setting

The CDS Consortium was funded by the Agency for Healthcare Research and Quality (AHRQ) to address the challenge of documenting, generalizing, and translating the CDS adoption experience at advanced sites to broader community settings.[14] The Consortium was created when investigators from Partners HealthCare (PHS) Information Systems (IS) formed an alliance with several other institutions intimately involved in creating and providing CDS tools and services in EHR's. Furthermore, a CDS rules service has been implemented to enable sharing of CDS on an advanced rules engine platform among consortium members.[16]

Within the AHRQ CDSC project, eight reminders were studied (Table 1) from January 1, 2010 to December 31, 2010. All were synchronous passive reminders triggered by opening the patients' electronic chart. For each reminder, we obtained data from the Longitudinal Medical Record including: patients who were eligible for the measure (the denominator), patients that had already or subsequently received the recommended action (the numerator), the reminders

displayed – when and to whom- as well as the provider acknowledgement to the reminder, and a coded response. This data was loaded into the Partners Quality Data Warehouse (QDW) and dashboards were constructed in Report Central[17] using Crystal Reports™.

Table 1 Reminders

Condition	Reminder	Measure	Reminder
CAD	CAD and no Aspirin	Patient has CAD and aspirin is on the medication list	Patient has CAD-equivalent on problem list and aspirin is not on the medication list. Recommend aspirin.
Diabetes	Diabetic overdue for HbA1c	Diabetes, HbA1c completed in the past 6 months	Patient with Diabetes Mellitus overdue for HbA1C
Diabetes	Diabetic almost due for HbA1C	Diabetes, HbA1c completed in the past 6 months	Patient with DM is due for HbA1C by <mm/dd/yyyy>
Diabetes	Diabetes overdue for Microalbumin/creatinine ratio	Diabetes, Microalbumin completed in the past year	Patient with diabetes overdue for urine Microalbumin/creatinine ratio
Diabetes	Diabetes almost due for Microalbumin/creatinine ratio	Diabetes, Microalbumin completed in the past year	Patient with diabetes almost due for urine Microalbumin/creatinine ratio by <mm/dd/yyyy>
Diabetes	Diabetic overdue for ophthalmology exam	Diabetes, Ophthalmology exam completed in past year	Patient with diabetes mellitus overdue for ophthalmology exam
Diabetes	Diabetic almost due for ophthalmology exam	Diabetes, Ophthalmology exam completed in past year	Patient with diabetes mellitus almost due for ophthalmology exam
Diabetes with Renal Disease	Diabetes Mellitus and Microalbumin/creatinine ratio >30	Diabetes, Microalbumin/creatinine ratio>30, and on ACE-inhibitor, ARB	Patient with diabetes mellitus, Microalbumin/creatinine ratio >30 and not on ACE-inhibitor, ARB. Recommend ACE-inhibitor or ARB

Detailed list of CDS reminders by condition.

2.2 Dashboard Design and Development

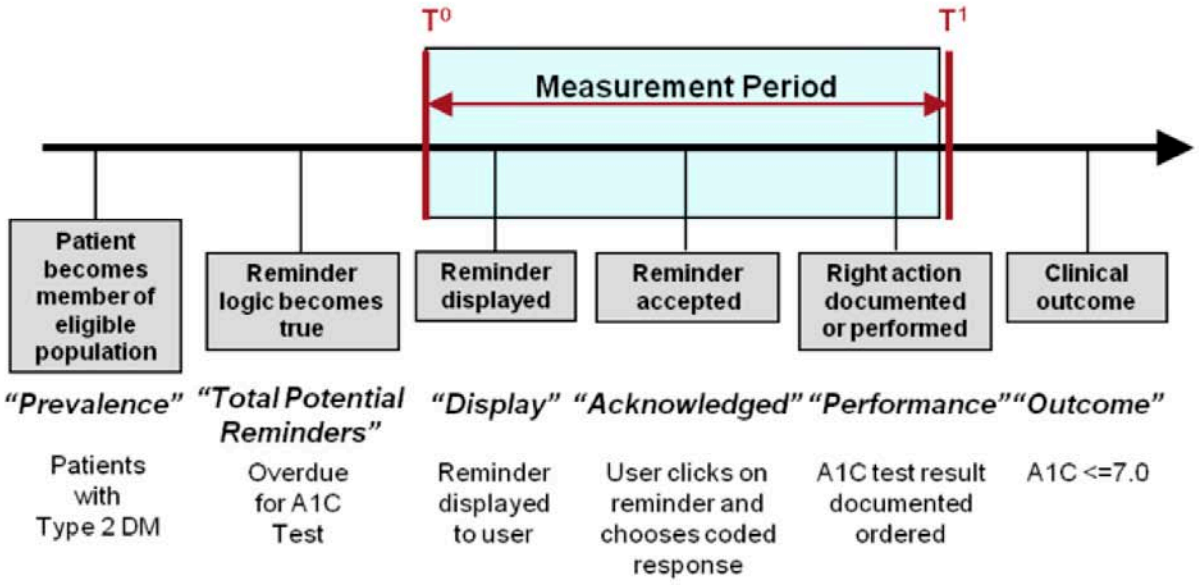
Two separate dashboards were built to target clinicians and CDS implementers (knowledge engineers). To accomplish this, the thoughts and requests of each type of user were considered and incorporated in the design process. Iterative designs were prototyped incorporating feedback from the CDS Dashboard Team, which was comprised of senior medical informaticians, senior knowledge engineers, and the principal investigator. It was determined that the clinician dashboard would present information regarding clinical performance for patients determined by the reminder logic (Table 2), which would then be organized by condition and compared to their peers. In comparison, the designer view was created to address reminder performance.

Table 2 CDS Reminder Performance logic per measure

Measure	Performance measurement
Patient has CAD and Aspirin is on the medication list	Patient has CAD and aspirin is on the medication list before the period start date and there is no stop date on the med entry that is before the period end date.
Diabetes, HbA1c completed in the past 6 months	Patient has an HbA1C entry < than the period end date and >6 months before the period start date.
Diabetes, Microalbumin completed in the past year	Patient has a MALCR entry entered < than period end date and >than 12 months before the period start date.
Diabetes, Ophthalmology exam completed in past year	Patient has an Ophthalmologic exam entry entered < than the period end date and > than 12 months before the period start date.
Diabetes, microalbumin/creatinine ratio>30, and on ACEI/ARB	Patient has an ACEI/ARB on their medication list before the period start date and there is no stop date on the medication entry that is before the period start date.

The designer view displays the level of acknowledgement of reminders, reminder performance by patient group, and specific data related to reminder effectiveness, including: reminder prevalence, reminder presentation, reminder acknowledgement defined as “the formal declaration of a reminder being received and acted upon” (Figure 1), CDS performance, and the NNR. Within the QDW, there are two ways of looking at each reminder – the first looks at the number of displays, defining a display event as the reminder being shown on the screen to an LMR user for a particular patient, where display events are counted by provider-patient-month. The second approach defines a reminder event at the patient level for the entire study period, i.e. the reminder being shown on a screen to any LMR user for a particular patient, treating each patient-year as one display event, while considering the patient compliant when the performance action was recorded within the following month from the reminder being displayed.

Figure 1 Timeline detailing the reminder lifecycle.



2.3 Proposed Measurement Framework

A measurement framework was developed to consider the lifecycle of an ambulatory EHR reminder (Figure 1). This lifecycle suggests the kind of events, actions and outcomes that can be used to define rates and measures to assess CDS effectiveness. Each stage of the lifecycle is associated with a particular measure: prevalence, logic, display, acknowledged, performance, or outcome. For each reminder, we defined the numerator and denominator for clinical performance. This refers to whether or not a patient who is part of the eligible population received the recommended action, independent of whether a reminder was displayed or not. Ultimately, the key to reminder effectiveness is the contribution the reminder makes to overall

clinical performance, i.e. when the reminder is displayed, how often the recommended action is subsequently taken. This can be expressed in a measure we call "CDS Reminder Performance" (RP) (Table 3).

2.4 Number Needed to Remind (NNR)

In 1988, Laupacis *et al.* proposed a measure of clinical benefit intended to capture value of certain clinical interventions – the number needed to treat (NNT)[18–20], calculating the inverse of the absolute risk reduction. Similarly, the Number Needed to Harm (NNH) reflects the number of patients that have to be exposed in order to harm one patient that otherwise would not have been harmed. We applied the same approach in an analogous fashion to reminders and proposed a new measure of reminder effectiveness called the "Number Needed to Remind" (NNR). Consequent with the NNT, the ideal NNR is 1, reflecting that every reminded patient will benefit from the reminder being displayed to one or more of their care providers.

The NNR corresponds to the number of patients reached by the reminder to result in one recommended action being taken or:

$$NNR_t = \frac{\text{total patients with reminders displayed over time t}}{\text{total patients with reminders and perfomance over time t}}$$

The denominator represents the number of patients to whom a reminder was followed by the appropriate performance event within 30 days of the end of the reporting period. The numerator corresponds to the cumulative count of patients with reminders displayed in each measurement period.

We used the performance data, including the NNR, to assess the effectiveness of each reminder and were able to gain some potential insights related to how well they function in the clinical setting. The logic to calculate the performance per each of the measures under study is explained in detail in table 2.

3 Results

During the study period, 1,982,735 reminders were triggered for the rules described in Table 1, being displayed to 12,327different providers. From the cohort of 116,027 patients included in the study, 65,516 of them were exposed to the one or more of selected reminders during the study period.

As a validation through example, we evaluated one reminder for diabetic patients who were overdue for their HbA1c after 6 months. During the study period the reminder was displayed 355,361 times (Table 3) to a total of 40,745 different patients (Table 4). Out of those patients, 16,549 (40.6%) of them had a performance event, defined as an HbA1c present in laboratory tests results database (Table 2) during the 30 days following the reminder being displayed. The RP for this rule would be the ratio of patients with performance over the total number of reminders displayed; for this particular example the RP for the Overdue HbA1c would equal 4.7 patients in performance per 100 reminders displayed. This indicator reflects the effectiveness of each reminder displayed on the expected performance. In contrast, when we calculate the inverse

of the patient performance we obtain a NNR of 2.5, reflecting the number of patients receiving the reminders in order to get one patient to comply with the defined performance measure (Figure 2).

Table 3, shows the number of reminded patients for whom the relevant performance action was found, and the total number of patients displayed with a reminder. The RP is the percentage of patients with reminder and performance over the total number of reminders displayed. Table 4 shows the reminders with performance for displayed patient reminders per month: NNR (the Number of patients needed to remind).

Table 3 CDS Reminder Performance for all patients' total reminders per month

Rule Name	***Reminder with Performance***	***Total Reminders***	***CDS Reminder Performance (%)***
CAD and no Aspirin on medication list	2,589	131,721	***2.0%***
Diabetes Mellitus and microalbumin/creatinine ratio >30 and no ACEI/ARB on medication list	1,668	57,678	***2.9%***
Diabetes overdue for Microalbumin/creatinine ratio	2,927	215,626	***1.4%***
Diabetes almost due for Microalbumin/creatinine ratio	11,864	82,385	***14.4%***
Diabetic overdue for HbA1c	16,549	355,361	***4.7%***
Diabetic almost due for HbA1c	13,751	70,024	***19.6%***
Diabetic overdue for ophthalmology exam	2,863	1,048,478	***0.3%***
Diabetic almost due for ophthalmology exam	1,947	21,462	***9.1%***
Total	***54,15***	***1,982,875***	***2.7%***

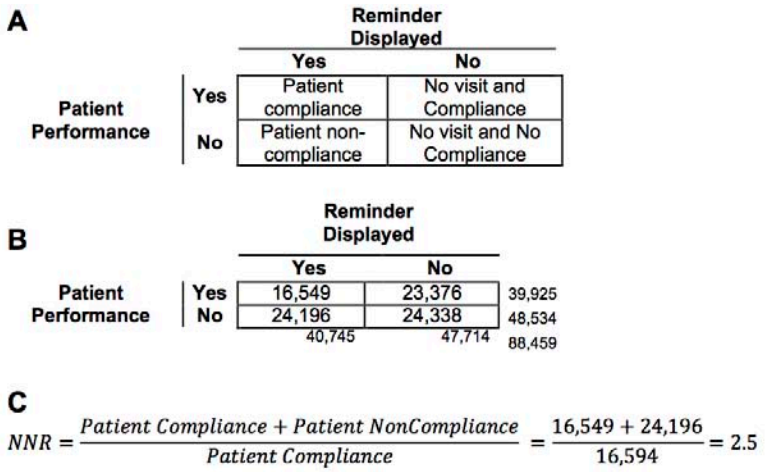
Table 4 Reminders with Performance for displayed patient reminders per month: NNR (the Number of patients needed to remind).

Rule	***Reminder with Performance***	***Patients with Reminders Displayed***	***Number Needed To Remind(NNR***
CAD and no Aspirin on medication list	2,589	*7,051*	2.7
Diabetes Mellitus and microalbumin/creatinine ratio >30 and no ACEI/ARB on medication list	1,668	2,949	1.8
Diabetes overdue for Microalbumin/creatinine ratio	2,927	*21,466*	7.3
Diabetes almost due for Microalbumin/creatinine ratio	11,864	*15,227*	1.3
Diabetic overdue for HbA1c	16,549	*40,745*	2.5
Diabetic almost due for HbA1c	13,751	*18,403*	1.3
Diabetic overdue for ophthalmology exam	2,863	*57,889*	20.2
Diabetic almost due for ophthalmology exam	1,947	*5,020*	2.6
Total	***54,158***	***168,750***	***3,11***

The evaluation of the CDS reminder performance by differentiating four scenarios in decision support performance, as explained in Figure 2, can distinguish patients in compliance with the performance measure from non compliant patients, while accounting for those patients affected by the reminder. From all eligible patients, considered the prevalence for the evaluated condition, only a fraction has a healthcare provider visit during a specified time frame, thus since the patient had no recent performance event (e.g. patient is overdue for HbA1c) and not having a reminder being displayed, produced the effect of patients without reminders and no performance.

The opposite scenario occurs when a patient is in compliance although no reminder was triggered. These patients could be in performance most commonly because of a proactive care team.

Figure 2 CDS performance evaluation framework for passive reminders for HbA1c to be completed every six months to the entire population of diabetic patients found in the Partners Healthcare System. (A) Distribution of all the active diabetic patients depending if they have received a reminder and whether they are in compliance with the performance measure. (B) HbA1c compliance among diabetic patients in Partners population. (C) Computation of the Number Needed to Remind based on CDS performance framework.



4 Discussion

Using routinely collected data in our EHR, we were able to examine the effectiveness of ambulatory care reminders for diabetes mellitus and coronary artery disease. Specifically, we measured the number of eligible patients that were receiving recommended care (clinical performance) and the effectiveness of the reminders displayed toward the healthcare providers (reminder performance). By identifying the patients with reminders displayed and those that received a suggested action, we calculated the Number Needed to Remind, i.e. the number of patients required to be reached by CDS reminders that are associated with one additional patient receiving recommended care. This measure, the NNR, may be useful for monitoring and differentiating the relative effectiveness of CDS rules.

The NNR measure clarifies the difference between the effectiveness of the reminder (RP, Reminder Performance) as an isolated entity from the Reminder effect on the patient level, and furthermore highlights the subtle differences in the reminder logic or documentation requirements that are associated with dissimilar performance. For example, looking at the almost due and overdue reminders for eye exam and HbA1c (Table 4), the “Overdue” reminder had a notably higher NNR than the “Almost Due” reminder. We speculate that this may be attributable to a number of factors that create a bias in the patient population that receives each reminder.

This unintended patient selection bias could be explained by the fact that to trigger an “Almost Due” reminder, a prior performance event needs to be present in the database and, more importantly, needs to be properly documented in a previous episode of care including a computable time stamp. Further, the “Almost Due” temporal logic allows for a higher degree of tolerance meaning that the reminder may fire in advance of a strict due date. In this case, the reminder will work better in those sites where providers document at the point of care and use the features in the EHR appropriately.

Advanced CDS may fail to achieve ideal care if it relies on incomplete or unstructured information that is considered difficult to acquire and maintain for the appropriate display of reminders and other forms of decision support. Here we can differentiate two scenarios. First, a lack of structured information will impede the display of the reminder or it will be displayed in a wrong clinical scenario. The second will occur when the performance event is not documented appropriately and the reminder will not stop triggering. The Ophthalmologic exam to diabetic patients (Table 1) is usually not documented as discrete data in the LMR – it is often free text in patient notes. Consequently, despite the performance action has been taken, the reminder keeps triggering and the corresponding performance evaluation cannot be reliably measured.

To evaluate the performance of CDS knowledge artifacts and compare them cross EHR interventions would enable a comparative effectiveness analysis of reminders implemented in different EHRs, locations or implementations. To evaluate the same rule and assess whether one implementation vs. another had a better NNR, or assess the variations in the NNR across EHR systems using the same reminders, would be revealing and could allow discovery of the appropriate design patterns that EHR and CDS interventions should follow.

In a recent study, a framework to evaluate the appropriateness of clinical decision support was proposed.[21] In contrast, our study examines the analytical component of the quality of care within the institution. Our approach seeks to examine the effectiveness of the knowledge artifacts as isolated entities as well as from the patient perspective and provide the tools that would enable to elucidate the reasons behind the poor performance that clinical decision support interventions may have. We believe that both approaches are complementary since the first event to be scrutinized in the clinical performance of decision support is whether the reminder is being displayed appropriately. We broaden the scope for this study and defined that improvement efforts should be driven by clinical performance.

Evaluation of Clinical Decision Support Systems is becoming increasingly important in those institutions that have chosen to use these systems to improve quality and safety. The development of accurate performance indicators will allow the comparison among different institutions, playing a key role on standardizing care in the near future. By setting a common ground and start comparing the effects of different reminder rules we will start discovering features and functionalities that can solve the issues we confront on a daily basis.

5 Conclusion

Healthcare information technology is changing the way that we treat our patients and CDS can play an important role to improve the quality of care that we deliver. We identified different measures that might be helpful to assess the performance of CDS, and described the potential interpretation in context. More research is required to develop a comprehensive CDS assessment framework, and to further evaluate these measurements. New indicators to accurately reflect the clinical performance of a CDS reminder are required, to concentrate the improvement strategies on specific aspects throughout knowledge artifact lifecycle. This study reinforces the importance of clinical performance evaluation as a key function in the CDS lifecycle that could increase effectiveness and finally improve care.

6 Acknowledgments

This publication is derived from work supported under a contract with the Agency for Healthcare Research and Quality (AHRQ) Contract # HHSA290200810010.

7 Disclosures

The findings and conclusions in this document are those of the author(s), who are responsible for its content, and do not necessarily represent the views of AHRQ. No statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

8 References

1. Blumenthal D, Glaser JP. Information technology comes to medicine. *N. Engl. J. Med.* 2007;356(24):2527-2534. doi:10.1056/NEJMhpr066212.

2. Bates DW, Gawande AA. Improving safety with information technology. *N. Engl. J. Med.* 2003;348(25):2526-2534. doi:10.1056/NEJMsa020847.

3. Osheroff JA, Teich JM, Middleton B, Steen EB, Wright A, Detmer DE. A roadmap for national action on clinical decision support. *J Am Med Inform Assoc* 2007;14(2):141-145.

4. Chaudhry B, Wang J, Wu S, et al. Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. *Ann. Intern. Med* 2006;144(10):742-752.

5. Berner E. AHRQ White Paper: Clinical Decision Support Systems: State of the Art. 2009. Available at: http://healthit.ahrq.gov/images/jun09cdsreview/09_0069_ef.html. Accessed November 7, 2011.

6. Shojania KG, Jennings A, Mayhew A, Ramsay CR, Eccles MP, Grimshaw J. The effects of on-screen, point of care computer reminders on processes and outcomes of care. *Cochrane Database Syst Rev* 2009;(3):CD001096. doi:10.1002/14651858.CD001096.pub2.

7. Bright TJ, Wong A, Dhurjati R, et al. Effect of clinical decision-support systems: a systematic review. *Ann. Intern. Med.* 2012;157(1):29-43.

8. Heselmans A, Van de Velde S, Donceel P, Aertgeerts B, Ramaekers D. Effectiveness of electronic guideline-based implementation systems in ambulatory care settings - a systematic review. *Implement Sci* 2009;4:82. doi:10.1186/1748-5908-4-82.

9. Patterson ES, Nguyen AD, Halloran JP, Asch SM. Human factors barriers to the effective use of ten HIV clinical reminders. *J Am Med Inform Assoc* 2004;11(1):50-59. doi:10.1197/jamia.M1364.

10. Dexheimer JW, Talbot TR, Sanders DL, Rosenbloom ST, Aronsky D. Prompting clinicians about preventive care measures: a systematic review of randomized controlled trials. *J Am Med Inform Assoc* 2008;15(3):311-320. doi:10.1197/jamia.M2555.

11. Sittig DF, Wright A, Osheroff JA, et al. Grand challenges in clinical decision support. *J Biomed Inform* 2008;41(2):387-392. doi:10.1016/j.jbi.2007.09.003.

12. Cash JJ. Alert fatigue. *Am J Health Syst Pharm* 2009;66(23):2098-2101. doi:10.2146/ajhp090181.

13. Schedlbauer A, Prasad V, Mulvaney C, et al. What evidence supports the use of computerized alerts and prompts to improve clinicians' prescribing behavior? *J Am Med Inform Assoc* 2009;16(4):531-538. doi:10.1197/jamia.M2910.

14. Middleton B. The clinical decision support consortium. *Stud Health Technol Inform* 2009;150:26-30.

15. Clinical Decision Support (CDS) Consortium. Available at: http://www.partners.org/cird/cdsc/.

16. Dixon BE, Simonaitis L, Goldberg HS, et al. A pilot study of distributed knowledge management and clinical decision support in the cloud. *Artif Intell Med* 2013. doi:10.1016/j.artmed.2013.03.004.

17. Jung E, Li Q, Mangalampalli A, et al. Report Central: quality reporting tool in an electronic health record. *AMIA Annu Symp Proc* 2006:971.

18. Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. *N. Engl. J. Med* 1988;318(26):1728-1733.

19. Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. *BMJ* 1995;310(6977):452-454.

20. McQuay HJ, Moore RA. Using numerical results from systematic reviews in clinical practice. *Ann. Intern. Med* 1997;126(9):712-720.

21. McCoy AB, Waitman LR, Lewis JB, et al. A framework for evaluating the appropriateness of clinical decision support alerts and responses. *Journal of the American Medical Informatics Association: JAMIA* 2011. doi:10.1136/amiajnl-2011-000185.

Characterizing the Sublanguage of Online Breast Cancer Forums for Medications, Symptoms, and Emotions

Noémie Elhadad, PhD[1], Shaodian Zhang[1], Patricia Driscoll[1], Samuel Brody[2], PhD
[1]Columbia University, New York, NY; [2]Google, Inc., New York, NY

Abstract

Online health communities play an increasingly prevalent role for patients and are the source of a growing body of research. A lexicon that represents the sublanguage of an online community is an important resource to enable analysis and tool development over this data source. This paper investigates a method to generate a lexicon representative of the language of members in a given community with respect to specific semantic types. We experiment with a breast cancer community and detect terms that belong to three semantic types: medications, symptoms and side effects, and emotions. We assess the ability of our automatically generated lexicons to detect new terms, and show that a data-driven approach captures the sublanguage of members in these communities, all the while increasing coverage of general-purpose terminologies. The code and the generated lexicons are made available to the research community.

Introduction

As online health communities like forums, blogs, and mailing lists become increasingly prevalent, patients are turning to these resources for information exchange and interaction with peers [1]. Patients with breast cancer, in particular, rely on cancer-specific online health communities for both informational and emotional support [2–6]. While this type of social networking has become central to the daily lives and decision-making processes of many patients, there are still many research questions open. For many research activities, capturing domain knowledge about topics discussed in a community and organizing terms and concepts discussed into lexicon and terminologies is needed for knowledge discovery and information extraction [5,7]. Designing automated tools to build these lexicons is a challenging task, however, because the language used in online health communities differs drastically from the genres traditionally considered in the field of information processing and from the sublanguages already investigated in the biomedical domain [8,9]. Health community vocabulary is characterized by abbreviations and community-specific jargon [10], and posts are authored in a style-free and unedited manner, with often informal and ungrammatical language. In addition, the content of the posts is both emotionally charged and dense with factual pieces of information, indicating that specific semantic types of information, like emotions, are more prevalent than in traditional biomedical texts.

In the biomedical domain, there are several clinical terminologies available that provide candidate keywords for lexicons, such as names of diseases, procedures, and drugs [11,12]. There exist health-consumer oriented terminologies, but their focus is on health consumers rather than patients in online communities, which have a different level of health literacy and convey at once a larger and more granular vocabulary for health terms than general health consumers [13]. More recently, researchers experimented with crowdsourcing to identify medical terms in patient-authored texts, but show that further processing and supervised learning is still needed to achieve acceptable results [14]. Finally, like for methods trained in other genres, existing terminologies, whether clinical or health consumer, do not cover the many misspellings and abbreviations typical of a given online health community and require manual updating to capture new terms introduced into the sublanguage [15,16].

Automated creation of lexicons has a long history in natural language processing. Unsupervised named entity recognition, and the use of seed terms in particular as a starting point for lexicon building, is a practical and promising method because it does not require a corpus, manually annotated with examples of terms [17]. Instead, seed terms are leveraged by looking for candidate terms with high context similarity to the seed terms. In the biomedical domain, such an approach has been used to identify disease names in medical mailing lists [15], recognize clinical and biological terms [18,19]. The approach is rooted in the Distributional Hypothesis, which states that words with similar contexts tend to have similar meanings [20]. In the biomedical domain, distributional semantics has been used for a wide range of tasks, such as matching MEDLINE abstracts to terms in an ontology [21], automatic generation of synonyms for gene and protein names [22], evaluation of language incoherence in patients with schizophrenia [23], and identifying semantically similar concepts in clinical texts [24,25], to name just a few [26–28].

In this paper, we describe an unsupervised method to generate lexicons representing the sublanguage of an online health community focusing on specific semantic types. Starting from a seed set of terms, all in the same semantic category (like medication names), it computes a typical context in which terms of that category occur. The context representative of a semantic category is then leveraged to identify new terms, which can augment the lexicon. To assess the value of our method and the generated lexicons, we ask the following research questions: (i) Can the method identify new terms to augment a seed set, and if so how accurately? (ii) How well does the method perform on generating lexicons for different semantic categories? And (iii) how stable is the method with respect to the quality of the underlying seed set?

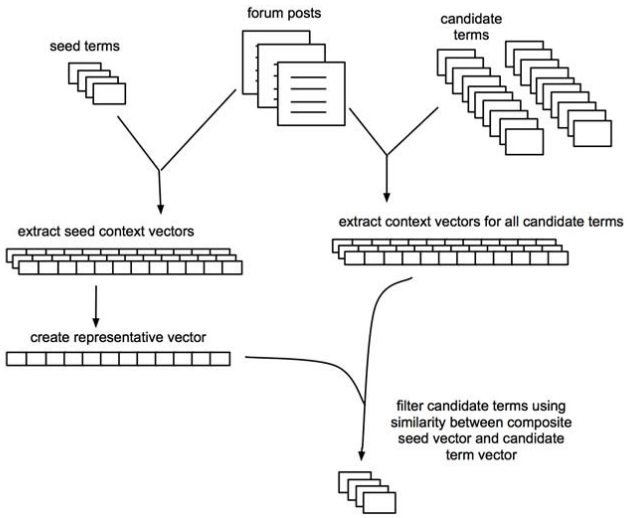


Figure 1. Overall pipeline to identify in an online health forum the terms representative of a specific semantic category.

Methods

The procedure for detecting terms in a forum that are representative of a specific semantic category is outlined in Figure 1. A seed set is gathered (either from an existing lexicon or from a small manually created one) representative of a given semantic category. Seed terms and their context, as defined from their occurrences in the online forum, are aggregated into a representative context vector, which reflect the typical context for terms in the category. As such, the representative vector acts as an implementation of the distributional hypothesis, where a word is defined by the context in which it is conveyed. To identify new terms for the semantic category, candidate terms from the forum are selected and an individual context vector is defined for each. Determining whether a candidate term belongs to a semantic category is achieved by computing the similarity between its individual context vector and the semantic category's representative vector. If a candidate term is used with words and patterns similar to the ones of the semantic category, it is likely the candidate term belongs. In our methods, we focus on three semantic categories of interest: (i) medications, (ii) signs and symptoms, and (iii) emotions and mental states.

Dataset

Following ethical guidelines in processing of online patient data, we focus on a popular, publicly available breast cancer forum with a large number of participants and obtained IRB approval. Posts from the publicly available discussion board breastcancer.org were collected [29,30]. At the time of collection, there were more than 60,000 registered members posting in 60 sub-forums. Our dataset consists of the most popular sub-forums. The extracted corpus for our analysis contains 26,153 threads corresponding to 253,231 posts. Overall, the corpus has 25.8M words for a vocabulary of 145K unique words. When considering only words that appear at least twice in the corpus, the vocabulary consisted of 75K words.

Lexicon Building

Choosing Seeds and Candidates. For each semantic category, we use an existing lexicon or a manually curated list of terms to gather a set of seed terms that are known to belong to the target category (e.g., medications). Using the forum corpus, we also extract a large number of candidate terms that may or may not be members of the target

semantic category. We customize our framework to our three semantic categories: medications, signs and symptoms, and emotions. The creation of these sets is described in Seed and Candidate Sets section below.

Table 1. Feature types used in the vector space model.

preceding word	A cell for each word in the vocabulary, indicating the number of times it appeared directly before the target term
word at -2	A cell for each word in the vocabulary, indicating the number of times it appeared 2 words before the target term
following word	A cell for each word in the vocabulary, indicating the number of times it appeared directly following the target term
word at +2	A cell for each word in the vocabulary, indicating the number of times it appeared 2 words after the target term
word within 3	A cell for each word in the vocabulary, indicating the number of times it appeared within 3 words before or after the target term

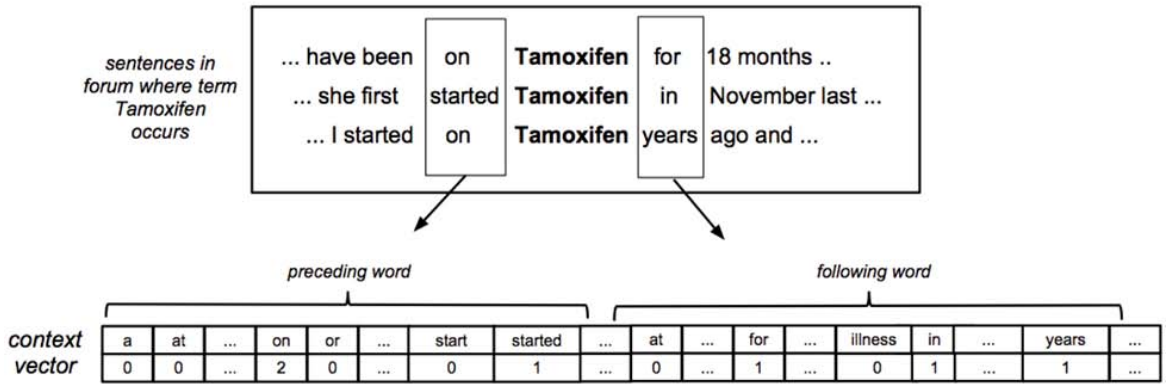


Figure 2. Part of the context vector for the seed term Tamoxifen. Each term context vector has a separate set of counts for preceding word, following word, word at -2, word at +2, and word within 3.

Constructing Context Vectors. Once the sets of seed and candidate terms have been selected, we employ a vector-space distributional similarity method to create context vectors for each term. The context vectors are derived from the vocabulary V found in the forum posts with the constraint that a word appears at least twice in the corpus. Each element in a term's vector contains a count of the number of times a word in V appeared in a certain context, such as directly preceding our term of interest. Because we use 5 contextual feature types, as described in Table 1, each context vector consists of 5|V| elements. We chose a set of local, highly specific contextual features to capture similarity in meaning and usage. For instance, in the example given in Figure 2, we can capture some of the data contained in exact patterns such as "been on X", as well as more general contextual features, such as the presence of the word started somewhere within 3 words of our target. This information helps the method find candidate terms that exhibit similar behavior to our seed terms, and are therefore likely to be in the same semantic category. These context vectors form the underlying representation in our method.

Creating a Representative Vector. In order to create a unified representation of a semantic category, the context vectors of the seed terms are merged into a representative vector for the category. Using vector addition, the individual context vectors are added. The vector is normalized by the number of seeds, producing a vector containing the average value for each of the seed vectors. A smoothing step is then performed, in which any values of the representative vector that are specific to only one seed are set to zero. This is intended to remove any contextual information that is unique to a single seed term and does not represent the semantic category as a whole. For example, assuming Arimidex is a seed term for medications, and it appears in the sentence "I have been on Arimidex (an aromatose inhibitor)", we will want to make use of the feature "preceding word on", since it is an indicator of a medication term, and will likely be shared by other seeds. However, the word aromatose is specific to the seed Arimidex, and we will discard the associated data unless it is shared by at least one other seed.

To further reduce noise and ensure a high-quality representative vector, a pre-filtering step is employed. The initial representative vector as created above is compared with each of the original seed term vectors using a cosine similarity metric [31]. If the similarity is below a certain threshold (.6 in our experiments), the seed term is

considered an outlier, and is removed. The representative vector is then re-created as described above using the filtered group of seed terms.

Calculating Similarity. A candidate term t is more likely to belong to a semantic category if its context vector is similar to the representative vector r for the category. Similarity is computed as the cosine metric between the two vectors. If the vector t is composed of (t_1, t_2, ..., t_n) and r is composed of (r_1, r_2, ..., r_n) then, their cosine is defined as

$$\text{Sim}(t,r)=\frac{t \times r}{||t|| \cdot ||r||} = \frac{\sum_{i=1}^{n} t_i \cdot r_i}{\sum_{i=1}^{n} t_i^2 \cdot \sum_{i=1}^{n} r_i^2}$$

The values of cosine similarity range from zero, indicating no similarity, to 1, indicating maximal similarity. Thus, our procedure scores each candidate term according to the similarity of its vector to the representative vector for the semantic category. The candidates can then be ranked in descending order of their similarity scores.

Seed and Candidate Sets

Seed sets are collected separately for each of the three semantic categories as described below.

Medications. To create a set of seed terms denoting names of medications, we use the comprehensive list of medications provided by RxNorm [32]. The list is then ordered by frequency of occurrence in the corpus, and terms appearing with low frequency in our corpus are removed (less than 50 in our experiments), resulting in a seed set of 137 medication terms.

The set of candidate terms for the medication category is defined initially as all out-of-vocabulary words in a standard English dictionary (dictionary from the Aspell program was used in our experiments), following the assumption that medication names are proper names, and thus not part of the standard English vocabulary. We only considered out of vocabulary terms from our corpus, which were frequent enough (50 times at least). This resulted in a set of 1,131 words as potential candidates for medication names.

Signs and Symptoms. We experiment with two medical lexicons for the construction of a set of seed terms denoting signs & symptoms. The first is the Unified Medical Language System (UMLS), where we use a list of all terms assigned to the 'sign or symptom' semantic type [11]. The second resource is SIDER, a list of terms denoting side effects extracted from FDA drug labels [33]. For each of these lists, we filter out all terms that are more than two words long. We then search for occurrences of the remaining terms in our data and extract all single word terms occurring more than 50 times, and all two-word terms occurring more than 20 times. This procedure provides the four seed sets described in Table 2. Despite the fact that both UMLS and SIDER seed lists share the most frequent term, there is relatively low overlap between them amongst these high-frequency terms (17 single-word terms, and 21 two-word terms).

Table 2. Number and average frequency of the terms in the four seed sets employed for detecting Signs & Symptoms, before and after (in parenthesis) the filtering procedures, along with the most frequent term in each seed set. The rightmost column specifies the coverage (cumulative frequency of all the terms inside the set) of each unfiltered seed set.

Seed Set	Size	Avg. Frequency	Most Frequent	Coverage
UMLS single word	84 (45)	1,205 (1,577)	pain	103,695
UMLS two words	136 (63)	134 (228)	hot flashes	37,702
SIDER single word	88 (51)	918 (1,418)	pain	80,780
SIDER two words	92 (38)	166 (335)	hot flashes	31,926

In the case of signs and symptoms, we cannot restrict candidates to out-of-vocabulary terms, as we did for medications, since signs and symptoms are often conveyed using standard-English words and are often multi-words. Instead, we consider any single-word or two-word term as a potential candidate, provided it appears frequently in our data (more than 50 times for single words, and more than 20 times for two-word terms), and consists of well-formed words (does not include numbers or other non-alphabetic characters).

In addition, for two-word terms, we perform another filtering step to reduce the number of candidates and improve quality. This filter is designed to remove multi-word terms that are very common in the data as a result of the frequency of the component words, rather than the term as a whole. For instance, the two-word term "and I" appears frequently in our data, but has little meaning as a unit, and its frequency is due to it being composed from two very common words. To filter such cases, we compare the probability of the term as a whole to the expected probability of the component words appearing in adjacent positions by chance, according to their individual probabilities, as shown in Equation 1. The ratio r between these probabilities is compared to a manually specified threshold t (in our

experiments, t = 20), and terms with ratios below the threshold are removed from the candidate list. After the selection and filtering procedures, we were left with a candidate list of 10,844 single-word candidates, and 37,015 two-word candidates.

$$\text{Eq. (1)} \quad r(\text{word1 word2}) = \frac{p(\text{word1 word2})}{p(\text{word1}) \cdot p(\text{word2})} \qquad p(x) = \frac{\#\text{ occurrences of x}}{\text{size of data}}$$

Emotions. While there exist terminologies for emotions [34], we experimented with a very small seed set for emotions. Part of our motivation is to test the robustness of our method to discovering new terms when a limited terminology or none is available. Given the most frequent words in the corpus of posts, we manually selected 10 adjectives as a seed set, which conveyed an emotional state randomly: *scared*, *grateful*, *sorry*, *fatigued*, *guilty*, *comfortable*, *nervous*, *confused*, *afraid*, *and happy*. Following the filtering step described above to compute the representative vector, there were six emotion seed terms left: scared (frequency of 5,512 occurrences in the corpus), grateful (frequency 1,445), sorry (frequency 20,768), confused (frequency 1,807), afraid (frequency 3465), and happy (frequency 11,338). For the sake of reproducibility, we replicated the experiments with different seed sets chosen randomly and obtained very similar results to the ones given this instance of seed set, and thus only report on these results.

Experimental Setup

The output of our method for a given semantic category is a ranked list of terms, which can augment a terminology of known lexical variants for the category (ranking is based on the terms' similarity scores to the given semantic category). We asked domain experts (two clinicians and one health psychologist) to review the lists for each of the three categories and tag each ranked term as a true positive (indeed a term that belongs to the semantic category) or a false positive (a term that does not belong to the semantic category). We report on the Precision at K[31], a standard evaluation metric for retrieval tasks in which the overall gold standard is unknown in advance – with different values of K for the top-K returned results, from K=10 to 50. We also report the cumulative coverage of the true-positive terms retrieved at the different K – that is, considering only terms that are not seeds. The coverage is a sanity check that the effort spent on discovering these terms pays off in terms of content that would have been ignored otherwise. For medications, the experts also encountered a number of terms that fell in a gray area. For instance, terms which were general names of treatments, or categories of medications, such as anthracyclines, a class of antibiotics. There were also terms indicating various drug cocktail treatments, as well as names of dietary supplements alternative treatments. Thus, for medication, we report two types of Precisions at K: a strict evaluation, which represents whether the ranked terms were medication names indeed, and one with a less strict definition of medication, which includes medication classes and drug cocktails.

Results

The code and the generated lexicons are available to the research community at *people.dbmi.columbia.edu/noemie*.

Augmenting an Existing Lexicon

Medications. In Table 3 we list the top ten terms according to the similarity with the representative vector for the medication category, along with their similarity score and frequency in the corpus. For the most part, the system correctly identifies terms indicating medications. There are misspellings (e.g., tamoxifin, benedryl, femera) and abbreviations (e.g., tamox) of medication names. The terms bisphosphonates and hormonals indicate classes of medications.

Table 3. List of top 10 retrieved medication terms not included in seed set, along with their similarity score and their frequency.

Retrieved term	Sim. score	Freq.	Retrieved term	Sim. score	Freq.
Tamox	0.888	6,107	bisphosphonates	0.821	549
Hormonals	0.888	1,012	carbs	0.821	326
Tamoxifin	0.880	666	mammos	0.817	704
Benedryl	0.831	402	femera	0.815	452
Fatique	0.827	108	lymphedema	0.815	2,656

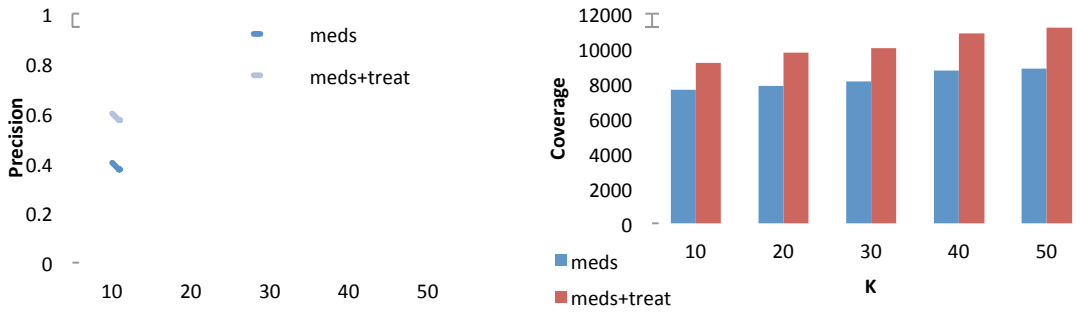


Figure 3. Precision (left) and number of instances covered (right) for the top K={10,20,30,40,50} retrieved terms not in the seed set for medication and treatments (meds+treat) and medication names only (meds).

We can see four classification errors: fatique, carbs, mammos, and lymphedema. The first is a rare misspelling of fatigue in the dataset, with thus little power to be categorized correctly. The term carbs is used in a similar fashion to many medications, since it is an ingested compound and forum users often discuss its effect on their health, much like they discuss medications. In general, we observed that various types of dietary supplements were common in our results for this reason.

Figure 3 shows the precision of the classification as we go down the list of retrieved terms (and following our experimental setup where only words outside of RxNorm were assessed for validity). Coverage ranged from 9,188 for K=10 to 11,191 occurrences for K=50 for medications and treatments, and ranged from 7,627 (K=10) to 8,859 occurrences (K=50) for medication names alone.

Signs and Symptoms. Table 4 shows the top 10 single-word and two-word terms retrieved as Signs and Symptoms retrieved when using SIDER as seed set. Figure 4 shows precision and coverage at K for the Signs and Symptoms category using either UMLS or SIDER as seed set.

Table 4. Top 10 single- and two-word terms retrieved as Signs & Symptoms using SIDER as a seed set.

Retrieved term	Sim. score	Freq.	Retrieved term	Sim. score	Freq.
itching	0.954	807	joint pain	0.985	2,213
caffeine	0.950	342	mouth sores	0.966	604
chemo	0.950	76,737	body aches	0.959	221
depression	0.950	2,575	acid reflux	0.958	205
discomfort	0.945	1,520	nose bleeds	0.954	131
bleeding	0.942	1,376	hair loss	0.952	1,549
bruising	0.942	336	bone aches	0.949	119
soreness	0.935	476	stomach problems	0.948	101
exhaustion	0.935	248	extreme fatigue	0.947	110
surgery	0.934	35,831	mood swings	0.945	309

As mentioned in the Methods section, we made use of two resources to develop two separate seed sets for this semantic category. In the figure, we see that the different characteristics of the seed set (see Table 2), result in differences in performance for our system. The UMLS seed set has better coverage than Sider on single-word terms, for a similar number of words. This means that the single-word terms in the UMLS are more suited to our domain, and this results in higher coverage and precision for the output of our system. For two-word terms the situation is reversed. The SIDER seed set has similar coverage, but is significantly smaller than the UMLS one (see Table 2). This means that the seed terms are more suited to our domain. For two-word terms, we get better coverage and precision when using SIDER as a seed.

There is another important difference worth noting between single-word terms and two-word ones. In the case of single word terms, the coverage of both the lexicons we employ is quite high. This means it is difficult to find new terms not mentioned in the lexicon, and these are found with lower confidence. This is also the reason for relatively low precision for single-word terms in this semantic category (the precision is measured only for the new terms). For two-word terms, on the other hand, initial coverage of the seed sets is quite low. There are many terms in the data that are strong members of this semantic category, but are not mentioned in the lexicons. This means the system can discover high quality new terms, with higher coverage and better precision.

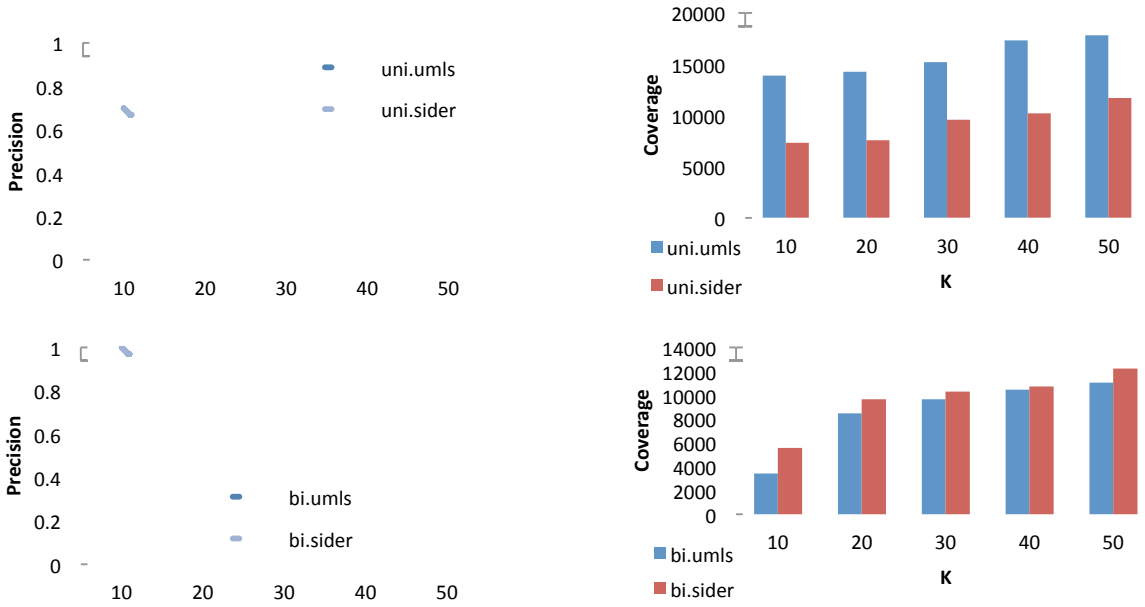


Figure 4. Precision (left) and number of instances covered (right) for the top K={10,20,30,40,50} retrieved single-word signs & symptoms (top) and two-word signs & symptoms (bottom), reported for UMLS and Sider as seed set.

Emotions. Table 5 shows the list of top-10 retrieved emotion terms from the small seed set of six emotion terms. All terms are high-frequency terms in the corpus, except for greatful. Interestingly, the misspelled greatful, despite its low frequency had a high similarity to emotions probably because of its correct spelling grateful was one of the seed term. The precision is much higher with emotions than with the other two semantic categories medications and signs and symptoms, starting at 100% at K=10 and decreasing to 78% at K=50. For this category, we evaluated up to K=100, with a precision of 64%. Moreover, the coverage of the true-positive emotion terms ranged from 20,076 for K=10 to 51,281 for K=50. This indicates two findings: (i) terms relating emotional states are highly frequent in our corpus, confirming that much emotional support is exchanged amongst the forum members; and (ii) our method is particularly good at discovering new terms when provided with a very small seed set (in this case a set of 6 chosen terms).

Table 5. List of top 10 retrieved emotion terms not included in seed set, along with their similarity score and their frequency.

Retrieved term	Sim. score	Freq.	Retrieved term	Sim. score	Freq.
glad	0.878	12,414	thankful	0.741	1,273
relieved	0.847	922	desperate	0.721	252
excited	0.780	1,035	delighted	0.719	152
thrilled	0.769	779	greatful	0.716	80
sad	0.745	2,994	saddened	0.698	175

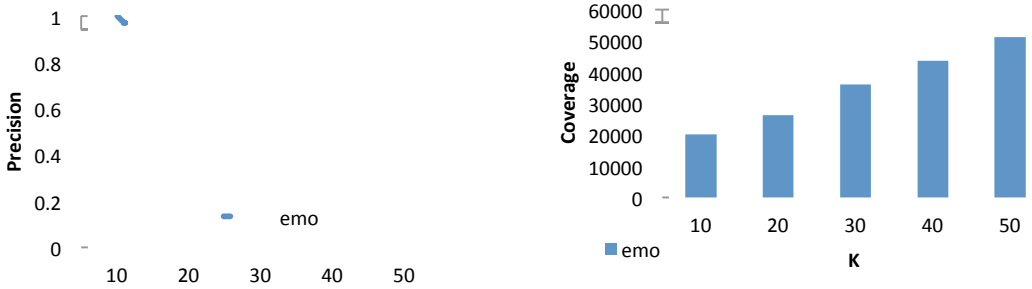


Figure 5. Precision (left) and number of instances covered (right) at K={10,20,30,40,50} for the retrieved emotion terms not in the seed set.

Sensitivity to Choice of Seed Set

A common concern when using statistical methods that rely on seed terms is the sensitivity of the method to the choice of seeds. To investigate this issue in our framework, we compared the results of using a variety of different seeds, and examined the effect on the terms retrieved by the system.

First, we compared the output of the system when using the seed set based on UMLS terms to the output when using seeds from SIDER. Despite low overlap between the two seed sets, the output of the system was similar for both. When comparing the top hundred most highly scored terms, we found an overlap of 91% in the output for two-word terms, and 89% for single word terms. This indicates that the semantic category we are looking for – terms indicating signs and symptoms – is a well-defined one, with specific usage patterns in the data. A practical implication is that any seed set containing good representatives of the semantic category can be used to successfully retrieve other terms in a fairly robust fashion.

We also experimented to discover if single-word terms could be used as seeds to retrieve multi-word terms in the same semantic category. We used the SIDER single-word seed set to rank the two-word candidates. In this case, however, we found much lower correspondence with the output of the two-word seeds (60% when compared to the UMLS two-word seed group, and 57% compared to the Sider seed). These findings indicate that single-word terms describing side effects are used in a different manner than multi-word expressions, in terms of immediate context, and that it is important to use a seed set of the same type as the candidates that are being ranked (single-word seeds for single word candidates, and multi-word terms as seeds for multi-word candidates).

Finally, on the basis of the success of a small, manually selected seed set for the emotion category, we experimented with using a similar strategy for the medication and signs and symptoms categories. We randomly shuffled the posts in our data and manually selected the first ten terms we saw that belonged to each category – i.e., without any reliance on any dictionary. We re-ran our method by filtering these small seed sets and constructing context vectors, and thus the resulting seed sets were at most ten words randomly chosen for each category. Table 6 shows the random seed sets in each category. The starred terms were filtered out automatically at the pre-filtering step when creating the representative vector for a given category.

For medication names, using a small set of random seeds was very successful, achieving 66% precision on the top 50 ranked results (74% if names of treatments are included), as compared to 44% and 62% when using RxNorm as basis for the seed. This demonstrates that if the target class is well defined, our method can learn accurate information from only a small number of examples, and a large, manually compiled lexicon is not necessary. For the category of signs and symptoms, the small randomly selected seed sets were also very effective. For single-word terms, the small seed set achieved 44% precision on the top 50 ranked results, significantly higher than that achieved by using UMLS and SIDER as seed sets, where the accuracy was 38% and 34%, respectively. For two-word terms, the randomly selected seed set achieved similar precision to using UMLS (62% on the top 50), but was not as effective as using SIDER (88%).

Table 6. Random seed set for Medications, Signs and Symptoms single words, and Signs and Symptoms two words. The terms with asterix were filtered out automatically during the step for construction of the representative vector.

Medications	**Signs & Symptoms Single word**	**Signs & Symptoms Two words**
tamoxifin	Pain	allergic reactions
herceptin	Leakage	mood swings
taxol	Cyst	* distended abdomen
carboplatin	Nausea	mouth ulcers
taxotere	neuropathy	hot flashes
tylenol	Baldness	high fever
xeloda	Blisters	scar tissue
zofran	Fatigue	* temple pain
percoset	headaches	abdominal pain
avastin	exhaustion	back pain

Discussion

Principal Findings. The primary objective of this study was to evaluate the use of lexical semantics in creating lexicons for use in content analysis of online health communities. Existing lexicons, like RxNorm, UMLS, and SIDER are fairly static resources, with potentially low coverage of the particular sublanguage of online health communities, whose informality often includes unique jargon, misspellings and abbreviations created by

community members. Our method aims to fill in these gaps, by generating lexicons to represent the language of members in a given community with respect to different semantic categories.

Our study suggests that using context vectors trained on a small seed set is a viable, robust method to expand existing medical lexicons across a range of potential semantic categories. The method was robust across semantic categories as long as seeds were good representatives of those categories. Furthermore, we showed that the seed set can be very small (e.g., six terms like in our experiments with detecting emotion terms) and still generate viable lexicons with good coverage. Finally, our experiments with UMLS and SIDER suggest that seed set selection should take into account surface characteristics like number of words in phrase. Finally, our study's experimental setup assessed the validity of only terms that were not already covered by existing lexicons. Thus, in the case of a semantic category and a lexicon with good coverage, our method has less opportunity to identify new terms (e.g., RxNorms and medications), but when the existing lexicons are scarce, our method identifies new terms with high accuracy (e.g., emotions).

Implications for Quantitative Research on Online Health Communities. As online health communities become a standard data source for mining information about patients, the underlying lexicons used to retrieve or assess prevalence of different terms must be representative of the way community members communicate. The lexicons we generated contain variations of known terms, which would be difficult to discover otherwise, as well as terms, which are not covered by existing lexicons. By making our code and generated lexicons available, we hope to contribute a valuable resource to the research community.

Limitations. Although the current work can be viewed as an important first step for augmenting lexicons to reflect online health community sublanguage, the results do not have high enough precision to be used without a manual annotator in the loop. Our hope is that the generated lexicons are still useful to researchers, since it is much easier to cross off terms in a generated list that should not in the lexicon, than it is to browse through thousands of posts manually to identify terms representative of the way members communicate (in our experiments, manual review was short and easy, on average 20 minutes per 100 terms). In our future work, we plan to experiment with other unsupervised methods and improve the accuracy of our generated lexicons. Second, our experiments focused on three specific semantic categories. While we chose them, because they are important types of content to know about for the content analysis of a breast cancer community and results indicate that a small seed set can generate valid lexicons for all three categories, we have not generalized our work to other types of semantic categories. Finally, the experiments presented in this paper focus on a single community as underlying corpus. In the future, we would want to compare lexicons learned from different communities specific to the same disease. We have conducted preliminary experiments indicating that, for instance, the lexicon learned in one breast cancer community is a useful resource for another, but further work is needed to generalize this finding. We also plan to test our method on other health communities specific to diseases different from breast cancer, to test the generalizability and robustness of our method.

Conclusion

This paper describes a method to generate a lexicon to represent the language of patient users in a given online health community with respect to specific semantic types. We experiment with a breast cancer forum and detect terms that belong to three semantic types: medications, symptoms and side effects, and emotions. Experimental results show that our method captures the sublanguage of members in these communities with more coverage than existing, general-purpose terminologies do. Furthermore, even with a very small number of seed terms, the method can generate reliable lexicons. This work contributes a building block to quantitative research on online health communities.

Acknowledgements. This work is supported by an NSF (National Science Foundation) award #1027886 (NE).

References

1 Fox S, Duggan M. Health online 2013. Pew Research Center's Internet & American Life Project 2013. http://pewinternet.org/Reports/2013/Health-online.aspx
2 Rozmovits L, Ziebland S. What do patients with prostate or breast cancer want from an Internet site? A qualitative study of information needs. *Patient Educ Couns* 2004;53:57.
3 Civan A, Pratt W. Threading together patient expertise. In: *Proc AMIA Annual Symposium*. 2007. 140–4.
4 Meier A, Lyons EJ, Frydman G, *et al.* How cancer survivors provide support on cancer-related Internet mailing lists. *J Med Internet Res* 2007;9.

5 Overberg R, Otten W, De Man A, *et al.* How breast cancer patients want to search for and retrieve information from stories of other patients on the internet: an online randomized controlled experiment. *J Med Internet Res* 2010;12.
6 Wang Y, Kraut R, Levine J. To stay or leave? The relationship of emotional and informational support to commitment in online health support groups. In: *Proc ACM Conference on Computer-Supported Cooperative Work (CSCW)*. 2011.
7 Portier K, Greer GE, Rokach L, *et al.* Understanding topics and sentiment in an online cancer survivor community. *J Natl Cancer Inst Monogr* 2013;47:195–8.
8 Harris ZS. *A theory of language and information: a mathematical approach*. Clarendon Press Oxford: 1991.
9 Friedman C, Kra P, Rzhetsky A. Two biomedical sublanguages: a description based on the theories of Zellig Harris. *J Biomed Inform* 2002;35:222–35.
10 Nguyen D, Rosé CP. Language use as a reflection of socialization in online communities. In: *Proc Workshop on Language in Social Media (LSM)*. 2011. 76–85.
11 Unified Medical Language System (UMLS). http://www.nlm.nih.gov/research/umls/
12 NCI Metathesaurus. http://ncim.nci.nih.gov/ncimbrowser/
13 Zeng QT, Tse T, Divita G, *et al.* Term identification methods for consumer health vocabulary development. *J Med Internet Res* 2007;9:e4.
14 Maclean D, Heer J. Identifying medical terms in patient-authored text: a crowdsourcing-based approach. *J Am Med Inform Assoc* 2013;20:1120–7.
15 Yangarber R, Lin W, Grishman R. Unsupervised learning of generalized names. In: *Proc International Conference on Computational Linguistics (COLING)*. 2002. 1–7.
16 Slaughter L, Ruland C, Rotergard AK. Mapping cancer patients' symptoms to UMLS concepts. In: *Proc AMIA Annual Symposium*. 2005. 699–703.
17 Alfonseca E, Manandhar S. Extending a lexical ontology by a combination of distributional semantics signatures. In: *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*. 2002. 281–93.
18 Zhang S, Elhadad N. Unsupervised biomedical named entity recognition: experiments with clinical and biological texts. *J Biomed Inform* 2013;46:1088–98.
19 Jonnalagadda S, Cohen T, Wu S, *et al.* Using empirically constructed lexical resources for named entity recognition. *Biomed Inform Insights* 2013;6:17–27.
20 Harris ZS. Mathematical structures of language. 1968.
21 Vanteru BC, Shaik JS, Yeasin M. Semantically linking and browsing PubMed abstracts with gene ontology. *BMC Genomics* 2008;9:S10.
22 Cohen A, Hersh W, Dubay C, *et al.* Using co-occurrence network structure to extract synonymous gene and protein names from MEDLINE abstracts. *BMC Bioinformatics* 2005;6:103.
23 Elvevaag B, Foltz PW, Weinberger DR, *et al.* Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia. *Schizophr Res* 2007;93:304–16.
24 Pivovarov R, Elhadad N. A hybrid knowledge-based and data-driven approach to identifying semantically similar concepts. *J Biomed Inform* 2012;45:471–81.
25 Garla VN, Brandt C. Semantic similarity in the biomedical domain: an evaluation across knowledge sources. *BMC Bioinformatics* 2012;13:261.
26 Cohen T, Widdows D. Empirical distributional semantics: Methods and biomedical applications. *J Biomed Inform* 2009;42:390.
27 Turney PD, Pantel P, others. From frequency to meaning: Vector space models of semantics. *J Artif Intell Res* 2010;37:141–88.
28 Erk K. Vector Space Models of Word Meaning and Phrase Meaning: A Survey. *Lang Linguist Compass* 2012;6:635–53.
29 Jha M, Elhadad N. Cancer stage prediction based on patient online discourse. In: *Proc BioNLP ACL Workshop*. 2010. 64–71.
30 Driscoll P, Lipsky Gorman S, Elhadad N. Learning Attribution Labels for Disorder Mentions in Online Health Forums. In: *Proc SIGIR Workshop on Health Search and Discovery*. 2013. 3–6.
31 Manning CD, Raghavan P, Schutze H. *Introduction to Information Retrieval*. Cambridge University Press 2008.
32 Nelson SJ, Zeng K, Kilbourne J, *et al.* Normalized names for clinical drugs: RxNorm at 6 years. *J Am Med Inform Assoc* 2011;18:441–8.
33 Kuhn M, Campillos M, Letunic I, *et al.* A side effect resource to capture phenotypic effects of drugs. *Mol Syst Biol* 2010;6:343.
34 Pennebaker JW, Francis ME, Booth RJ. *Linguistic inquiry and word count: LIWC 2001*. 2001.

Integrated Multisystem Analysis in a Mental Health and Criminal Justice Ecosystem

Dr. Erin Falconer[1], Dr. Tal El-Hay[2], Dr. Dimitris Alevras[3], Dr. John Docherty[1], Dr. Chen Yanover[2], Alan Kalton[4], Dr. Yaara Goldschmidt[2], Dr. Michal Rosen-Zvi[2]
[1] Medical Affairs, Otsuka America Pharmaceutical, Inc., Princeton, NJ, USA
[2] IBM Research - Haifa, Israel
[3] IBM Global Business Services, West Chester, PA, USA
[4] IBM Research – Africa, Nairobi, Kenya

Abstract
Patients with a serious mental illness often receive care that is fragmented due to reduced availability of or access to resources, and inadequate, discontinuous, and uncoordinated care across health, social services, and criminal justice organizations. These gaps in care may lead to increased mental health disease burden and relapse, as well as repeated incarcerations. Further, the complex health, social service, and criminal justice ecosystem within which the patient may be embedded makes it difficult to examine the role of modifiable risk factors and delivered services on patient outcomes, particularly given that agencies often maintain isolated sets of relevant data. Here we describe an approach to creating a multisystem analysis that derives insights from an integrated data set including patient access to case management services, medical services, and interactions with the criminal justice system. We combined data from electronic systems within a US mental health ecosystem that included mental health and substance abuse services, as well as data from the criminal justice system. We applied Cox models to test the associations between delivery of services and re-incarceration. Using this approach, we found an association between arrests and crisis stabilization services in this population. We also found that delivery of case management or medical services provided after release from jail was associated with a reduced risk for re-arrest. Additionally, we used machine learning to train and validate a predictive model linking non-modifiable and modifiable risk factors and outcomes. A predictive model, constructed using elastic net regularized logistic regression, and considering age, past arrests, mental health diagnosis, as well as use of a jail diversion program, outpatient, medical and case management services predicted the probability of re-arrests with fair accuracy (AUC=.67). By modeling the complex interactions between risk factors, service delivery and outcomes, we may better enable systems of care to meet patient needs and improve outcomes.

Introduction

The mental healthcare system in the United States is fundamentally broken – it is fragmented, inconsistent, underfunded, and rapidly deteriorating. Dominated by a lack of both consistency and continuity of care, the system allows many clients to mentally decompensate, ultimately leading to three negative crisis outcomes – homelessness, emergency hospitalization and incarceration. This fragmentation in the treatment of the mentally ill is particularly evident when one considers that it has been estimated that between 44 and 61 percent of jail inmates in the United States have a mental health problem (James et al Mental health problems in prison). Many individuals with severe mental illness are released from prison every year in the United States and re-enter the community with a need to continue treatment for their mental health issues. Continuous mental health treatment of these individuals may help prevent relapse and recidivism.

Lack of continuous care for adults with serious mental illness may not only result in more decompensation and crisis as the individual navigates the mental health, social, and criminal justice systems, but it also limits our understanding of the impact and interaction between modifiable risk factors and access to multiple services on patient outcomes such as re-arrest. Given that different agencies often maintain isolated datasets, the fragmentation of data systems and lack of access to continuous patient-level data means that it is difficult to collate data across health, social, and criminal justice agencies to evaluate the interplay between multiple services and outcomes. There is a critical need to evaluate patient-level and service-level data across multiple agencies in order to understand the mechanisms through which we may intervene to prevent or delay psychiatric crises.

Previous work evaluated data from US Medicaid claims files and arrest records and found a reduced risk of re-arrest with receipt of outpatient services[1,2,3] and psychotropic medication possession[3] in adults with mental illness. Other research using county and statewide criminal justice records and archival data from health and social services found that individual risk factors including being homeless, not having outpatient mental health treatment, and having involuntary psychiatric evaluation in the previous quarter, and being black, younger than 21 years and having a co-occurring substance abuse problem increased the odds of arrest[4].

Recent studies for other medical applications have used electronic medical records data to establish predictive models for illness severity in various disease domains, including preterm infants[5], congestive heart failure[6], septic shock[7] and HIV[8].

In this study, we describe an approach to modeling the interplay among services and outcomes across an ecosystem of medical and social services providers and the criminal justice system in which there is a constant flow of individuals with serious mental illness in and out of the criminal justice system. We explored associations between the occurrence of arrest and crisis services outcomes tested using hazard modeling with both fixed- and time-dependent covariates. We demonstrate the utility of such a combined dataset for predictive modeling by training and testing a model for arrest prediction.

Preliminaries

In this study, the sets of covariates used for prediction include both basic risk factors as well as indicators of access to specific services. Cox models were applied to test the associations between access to services and outcomes. Predictive models were constructed using elastic net regularized logistic regression.

Association analysis using Cox models

Cox proportional hazard models were used for testing associations between risk factors and the expected time for failure events to occur[9].This association is modeled using a *hazard rate* that represents the amount of risk as a function of time. The effect of each risk factor is assumed to be multiplicative with respect to the hazard rate.

In addition to predicting the effect of services given immediately after release, we examine the effect of continuous access to services. These tests involve time dependent covariates such as access to services in every month after

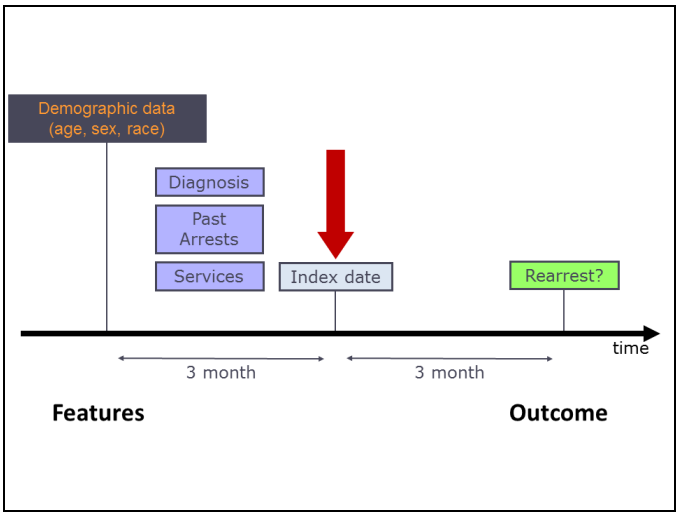


Figure 1: Data preprocessing and feature extraction framework

release from jail. Association tests with such covariates were performed using extended Cox models (see [11] for more details).

Predictive modeling using elastic nets

An elastic net is a method that allows classical regression models to deal with high dimensionality of observations.

This method performs data-driven variable selection and results in a sparse model that includes the most informative covariates. Learning in such models involves tuning of two parameters: (1) alpha – which controls the sparseness and stability of the model, where a higher alpha increases the tendency of the learning algorithm to filter out non-informative covariates; (2) beta – a regularization parameter that prevents over-fitting of the model to the data which is employed to obtain optimal generalization performance. These parameters are usually tuned using internal cross-validation on a training data set. The accuracy of the model is assessed on a test set. For more details see[12].

Methods

Framework – integrated view of patients ecosystem

This study uses information extracted from electronic systems resident within a mental health ecosystem in the southern US. This included data used to support the claims process for medical and social services delivered to mental health and substance abuse patients, and data collected by the criminal justice system. All individual patient data used for the analysis were collected after obtaining appropriate consents and agreements, and personal information was removed to protect patient privacy.

Patient data collected for this analysis span 21 months and describe the engagements that patients have with service providers in the ecosystem. An essential part of analyzing such longitudinal data is defining an index date, where data collected before that date serve as input to decision making and data from after that point in time define the outcome values and characterizes the treatment. We then define the target populations on which to focus and a quantitative outcome measure. Finally, we extract and filter features and risk factors to drive the modeling (Figure 1).

The main outcome addressed in this study was a re-arrest, as experienced by individuals that had become involved in the criminal justice system. The study explored two main questions: (i) What are the modifiable and non-modifiable risk factors associated with this outcome? (ii) How well can we predict the likelihood of re-arrest using such risk factors? Our approach employed an association analysis to explore the answers to the first question and a machine learning analysis to explore the answers to the second question.

We further test the association of various risk factors to admission to acute service, namely crisis stabilization unit.

We hypothesized that the relationship between mental health services and the criminal justice system may be bidirectional; to explore this hypothesis, we test associations between various risk factors including previous arrest records and the risk of admission to acute mental health treatment facilities (namely crisis stabilization unit).

Data sources

Datasets for mental health and substance abuse admissions and events were included which span the time window from October 1, 2010 to June 30, 2012. All mental health admissions before October 1, 2010 have been recorded with an admission date of October 1, 2010. Substance abuse admissions data span the period between July 1, 2009 and June 30, 2012. In this study we focus on Seriously Persistent Mentally Ill (SPMI) individuals that have one of the following diagnoses:

- Bipolar disorder (ICD9 codes: 296 to 296.19 and 296.40 to 296.99)
- Schizophrenic disorder (ICD9 codes: 295 to 295.99 and 297 to 298.99)
- Major depression (ICD9 codes: 296.20 to 296.39)

The dataset includes data on an SPMI population of 29,558 individuals.

Arrest data were supplied by the Department of Law Enforcement and was extracted from the Criminal Justice Information Services (CJIS). These data span a period from January 1, 2007 to September 6, 2012, and include records on 184,470 individuals. Out of these, 5,148 overlap with the SPMI population in the health ecosystem studied. The court runs a program that helps identify and divert detainees with a mental illness into a Jail Diversion Program (JDP), which seeks to reintroduce individuals into a sustained care environment, combining mental health and housing services as part of a structured year-long engagement. The court provided a list of participants for approximately ten years, overlapping the data contained in the other data sources.

Population Selection Criteria
To analyze the relations between arrest and behavioral health service events we focused on a subset of the adult population having records both in the CJIS and mental health ecosystem datasets. We excluded 281 individuals from this cohort because of inaccurate and inconsistent timeline data. Out of the remaining individuals, a total of 3,274 were released from an arrest after October 1st, 2010, which is the starting date of the mental health services recorded in the dataset. Of these, 3,171 were adults at the time of release[a].

In addition to viewing arrest as an outcome, we analyzed the association of past arrests with the outcome of admission to an acute mental health treatment using the SPMI cohort. We excluded individuals whose first recorded admission had ambiguous or unknown dates, creating a subset of 15,930 subjects. Of these, we further focused on the adult population (N=14,228).

Statistical Analysis

Association between non-modifiable risk factors, receiving services after release from jail, and the risk of re-arrest
The initial association analysis examined non-modifiable risk factors including gender, age, race, mental health diagnosis and past arrests using a Cox proportional hazard model. The association of receiving different service types with the risk of re-arrest was evaluated, adjusting for these non-modifiable risk factors. The dataset contains thirty nine types of services, out of which fourteen service types were given to more than 20 patients in the cohort. Each of these service types was represented using an *indicator covariate* equal to one if an individual received the service at least once in the first quarter after release from jail and equal to zero otherwise.

Continuous access to services
Extended Cox models were used to examine the association of services given throughout the entire period after release from jail to the risk of re-arrest. More specifically, for each patient, all release dates after Oct 1, 2010 were listed and corresponding re-arrest dates were identified (or, if the patient was not re-arrested, the end-of-study date was determined). Starting from each such release date, the number of times each service was given to the patient in each consecutive 90 day time period was tabulated.

Subsets of these time-varying covariates, in addition to the non-modifiable factors, were then used to infer the parameters of extended Cox models. In particular, models were constructed with the following features:

1. An indicator covariate identifying whether or not a specific service was given within the last ninety days (including non-modifiable risk factors) to predict re-arrest within the coming ninety days.
2. An indicator covariate identifying whether or not a specific service was given since the last release from jail (including non-modifiable risk factors) to predict re-arrest within the coming ninety days.

Predictive modeling using elastic nets
To test the predictability of the arrest outcome, data were partitioned into a training set which contained approximately 80% of the cohort and a test set which contained the remaining 20%. An elastic net regularized regression model was used where alpha was tuned to balance sparseness and stability on the training set.

Because the goals of the analysis were set to predict re-arrest probability in the second quarter after release, the target population was similar to the one described in re-arrest risk factor analysis. However, individuals were excluded for which data were not available for two quarters. Applying this additional criterion, the cohort size was established as 1,679 individuals in the training set and 421 in the test set. We evaluated the predictive power of the model using a receiver operating characteristic (ROC) curve which compares the likelihood of correctly and incorrectly predicting re-arrest.

[a] Due to the removal of exact birth dates, we use estimated ages at different time points. Here we include individuals with estimated age at release > 18.

Results

Preliminary associations of demographic and historical factors with re-arrest

The association between non-modifiable risk factors was estimated using Cox proportional hazard models. Past arrests factors were modeled as indicator variables whose value was one if the individual was arrested and released from jail between January 2007 (start date of the CJIS data) and October 2010, and zero otherwise.

Preliminary associations (i.e., without adjusting for other variables) between these factors and the risks of re-arrest are summarized in Table 1. Associations with p-value < 1e-4, will remain significant at a 0.05 level after a Bonferroni correction for 500 hypotheses. In particular, schizophrenia, history of arrests, male gender, black race, and younger age are shown to be risk factors for increased likelihood of re-arrest, in agreement with previous studies[4].

Table 1. Preliminary associations between baseline characteristics and the risk of re-arrest

Factor	P-value	Crude Hazard ratio	95% confidence interval
Gender (Female vs. Male)	<1e-4	0.71	0.63-0.81
Race (Black vs. Other)	<1e-4	1.31	1.18-1.47
Diagnosis (vs. Major Depression)			
Bipolar Disorder	0.02	1.22	1.03-1.45
Schizophrenic Disorders	<1e-4	1.50	1.30-1.74
Past arrests	<1e-4	2.04	1.80-2.31
Age	<1e-4	0.99	0.99-1.00

Association between arrests and crisis services

In the ecosystem studied, a large proportion of individuals were first admitted into the system of care through a Crisis Stabilization Unit (CSU; representing more than 30% within a week from first admission). A subset of the population was examined that did not record an admission to a CSU in the first quarter (N=10,307). In this sample, the association of a later CSU admission with past arrests, adjusted for age, gender, race and mental health diagnosis was significant (p<1e-4) with a high hazard ratio of 2.46 (95% confidence interval 2.00-3.02).

Services associated with reduced risk of re-arrest

To test associations with services given in the first quarter after release inmates that were re-arrested within the quarter were excluded. Out of 3171 adults, a total of 2,377 (~75%) remained out of jail during this period. This test therefore was able to examine conditional probabilities of future re-arrests given that the individual remained out of jail in the first quarter.

Associations between service indicator variables and the risk of re-arrest adjusting for gender, age, race, mental health diagnosis and past arrests as defined in the baseline test model were also tested. Results, summarized in Table 2, indicate an association of case management and medical services with a reduced risk of re-arrest[b]. Figure 2 presents Kaplan-Meyer plots of arrest probability for individuals that stayed out of jail in the first quarter after release, given their access to these services in this quarter.

Table 2. Associations among services and the risk of re-arrest

Factor	N	P-value	Adjusted Hazard ratio	95% confidence interval
Case management	172	0.00018	0.45	0.30-0.68
Medical services	491	<1e-4	0.59	0.47-0.74
Outpatient group[c]	45	0.038	0.46	0.22-0.96

[b] The CJIS data include both booking and arrest records. The table above refers to booking dates. Using arrest dates instead, gives a 0.48 hazard ratio for case management and 0.64 for medical services.

[c] This association is not considered significant as it does not pass a multiple hypothesis correction.

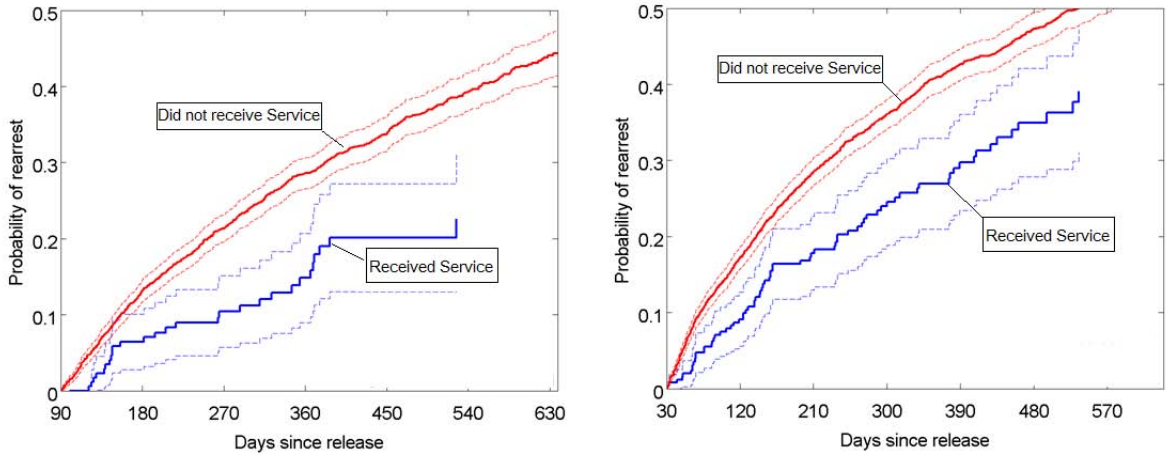


Figure 2: Kaplan-Meyer estimators for arrest probability demonstrating the effect of case management (left) and access to medical services (right)

Continuous access to care and continuous monitoring of patients states

Extended Cox model analysis shows that the indicator for Medical Services, either in the past 90 days or since release, is significantly associated with a reduced risk for re-arrest, with hazard ratios of 0.68 (confidence interval 0.58-0.80, p-value<1e-5) and 0.67 (0.58-0.78, p-value<1e-7), respectively. Conversely, the indicator for Crisis Stabilization, in both time periods, is associated with an elevated effect on the risk of re-arrest, with hazard ratios of 1.43 (confidence interval 1.22-1.69, p-value<1e-4) and 1.23 (1.07-1.42, p-value=0.003), respectively. Schonenfeld residuals for all these indicators, except for Crisis Stabilization in the past 90 days, attested to the correctness of the proportional hazard assumption.

Predictive Modeling of Re-Arrests

Elastic net regularized logistic regression models were trained using a training set containing 1,679 individuals. The regularization parameters alpha and beta were tuned using cross validation and the model was retrained on the entire training set with optimal parameters. Testing this model on a test set of 421 individuals resulted in an AUC of 0.67 (see 'Full model' in Figure 3 below). Informative covariates selected by the training procedure included age, past arrests, mental health diagnosis, enrollment to the JDP as well as utilization of outpatient group services, medical services and case management. The probability of re-arrest is modeled as function of a weighted sum of these factors. As the ROC curve in Figure 3 indicates, the model correctly predicts 50% of individuals in the ecosystem at risk for re-arrest based on the defined risk factors, while mis-characterizing 30% of individuals at risk. To assess the predictability of re-arrest from basic demographic data, namely, age, gender and race, we trained a simpler model using the same cohort and an elastic net model. This model was inferior to the full model, with an AUC of 0.60 and 42% true positive rate at the 30% false positive threshold ('Basic model' in Figure 3). The difference between the two ROCs illustrates the additional predictive power of the judicial and mental health related factors.

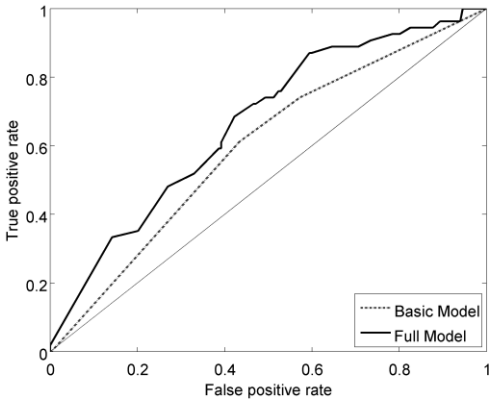


Figure 3: ROC curve of elastic net predictive model for re-arrest outcome

Discussion

In this analysis, we found that characteristics including schizophrenia, history of arrests, male gender, black race, and younger age were risk factors for increased likelihood of re-arrest. Further, factors such as past arrests increased the risk of having a crisis stabilization event, whereas receipt of case management services or medical services following release from jail reduced the likelihood of re-arrest. Using this knowledge of non-modifiable and modifiable core risk factors associated with re-arrest, we developed a model that predicted the probability of re-arrest after the first 90 days of release from jail with reasonable accuracy.

These findings support previous work that examined US Medicaid claims and arrests records, and other research, which found that factors associated with greater risk of arrest include minority racial-ethnic status[3] or African American race[1,4], male gender,[3] younger age. [3] and receipt of outpatient services (which may have included individual or group behavioral health services, medication checks,[2] and/or case management[1,3]). While we found that the receipt of medical and case management services predicted a reduced risk of arrest, the receipt of other behavioral health services such as outpatient group therapy was not considered significant. These findings therefore extend previous research by considering the independent contributions of case management and medical services in predicting re-arrest. However, while our findings suggest that providing case management and medical services within 90 days of release from jail may reduce the chance of re-arrest in adults with serious mental illness, it is important to remember that these are correlational results, and differential assignment to services may be due to some unmeasured variable associated with a better prognosis.

In the current analysis, we chose to include individuals with major mental illness including depression, bipolar disorder, and schizophrenia because these diagnoses are found at a high rate in the criminal justice system.[13] Future work should explore the role of other mental health diagnoses in predictive models, as well as the effect of comorbid conditions. Future studies should also be conducted to refine the model by integrating other sources of data (e.g. additional medical claims, pharmacy and hospitalization data). In the current analysis, medical services included primary medical care, psychiatric assessment and services, and administration of psychotropic drugs and other medications. It would be of particular interest for future analyses to examine the specific role of pharmacy/medication administration in predicting re-arrest, particularly given recent data showing that post-hospitalization medication possession reduced the likelihood of arrest in adults with serious mental illness in a Florida Medicaid population.[3]

Understanding how we can best intervene to address modifiable risk for incarcerations may be important not only for improving the quality of life of adults with serious mental illness, but also for reducing costs within the systems that support their care. Using Florida Medicaid data and records from Florida's Department of Children and Families (DCF) and the Florida Department of Law Enforcement (FDLE), Van Dorn et al (2013) compared the costs associated with criminal justice system involvement with those for mental health treatment, and found that overall system costs were lower for adults with serious mental illness who did not get arrested.[3] Taken together with our current findings, the results suggest that increasing the provision of case management and medical services in a

SPMI population at risk for arrest may be an important strategy for reducing overall system cost burden. This should be explored in future research.

Conclusions

In conclusion, our findings illustrate the complex interactions between modifiable and non-modifiable risk factors and delivery of services on outcomes in adults with serious mental illness. The data-driven approach defined in this analysis demonstrates the value of integrating data across disparate datasets from healthcare, social services, and criminal justice agencies. Further development of this predictive model may help us to identify those individuals who are at greater risk for re-arrest and crisis, and to intervene in a timely manner to help improve outcomes for the mentally ill. A reduction in arrests in this seriously mentally ill population may not only improve patient outcomes, but also diminish the burden on the judicial and health systems.

References

1. Gilbert, A.R., Moser L.L, Van Dorn R.A., et al. "Reductions in arrest under assisted outpatient treatment in New York." *Psychiatric Services*. 2010;61(10):996-999.
2. Morrissey JP, Cuddeback GS, Cuellar AE, et al. "The role of Medicaid enrollment and outpatient service use in jail recidivism among persons with severe mental illness." *Psychiatric Services*. 2007; 58(6):794-801.
3. Van Dorn RA, Desmarais SL, Petrila J, et al. "Effects of Outpatient Treatment on Risk of Arrest of Adults With Serious Mental Illness and Associated Costs," *Psychiatric Services,* pp. 856-862, 2013.
4. Constantine R, Andel R, Petrila J, et al. "Characteristics and experiences of adults with a serious mental illness who were involved in the criminal justice system". *Psychiatric Services*.2010; 61(5):451-457.
5. Saria S, Rajani AK, Gould J, et al. "Integration of early physiological responses predicts later illness severity in preterm infants." *Science translational medicine*.2010;2(48):48ra65.
6. Sun, J., Hu, J., Luo, D., et al. Combining knowledge and data driven insights for identifying risk factors using electronic health records. In *AMIA Annual Symposium Proceedings* (Vol. 2012, p. 901-910). American Medical Informatics Association.
7. Paxton, C., Niculescu-Mizil, A., & Saria, S. Developing Predictive Models Using Electronic Medical Records: Challenges and Pitfalls. In *AMIA Annual Symposium Proceedings* (Vol. 2013, p. 1109-1115). American Medical Informatics Association.
8. Zazzi, M., Kaiser, R., Sönnerborg, A., et al. "Prediction of response to antiretroviral therapy by human experts and by the EuResist data-driven expert system (the EVE study)." *HIV medicine*, 2011, *12*(4), 211-218.
9. Cox, D. R.. "Regression Models and Life Tables (with Discussion)." Journal of the Royal Statistical Society, Series B 34:187-220, 1972
10. Fisher, L. D. ve Lin D. Y. "Time-dependent covariates in the Cox proportional hazards regression model." *Annual Review of Public Health* 20, 145-157, 1999.
11. Pettitt, A. N. and Daud, I. "Bin Investigating time dependence in Cox's proportional hazards model." *Applied Statistics* 39, 313-329, 1990.
12. Zou, H. and T. Hastie. "Regularization and variable selection via the elastic net." Journal of the Royal Statistical Society, Series B, Vol. 67, No. 2, pp. 301–320, 2005.
13. Ditton PM: Mental Health and Treatment of Inmates and Probationers. Report NCJ 174463. Washington, DC, US Department of Justice, Bureau of Justice Statistics, 1999.

TagLine: Information Extraction for Semi-Structured Text in Medical Progress Notes

Dezon K. Finch, PhD,[1,2] James A. McCart, PhD,[1] Stephen L. Luther, PhD[1]
[1]James A. Haley Veterans Hospital, Tampa, FL; [2]University of South Florida, Tampa, FL

Abstract

Statistical text mining and natural language processing have been shown to be effective for extracting useful information from medical documents. However, neither technique is effective at extracting the information stored in semi-structure text elements. A prototype system (TagLine) was developed to extract information from the semi-structured text using machine learning and a rule based annotator. Features for the learning machine were suggested by prior work, and by examining text, and selecting attributes that help distinguish classes of text lines. Classes were derived empirically from text and guided by an ontology developed by the VHA's Consortium for Health Informatics Research (CHIR). Decision trees were evaluated for class predictions on 15,103 lines of text achieved an overall accuracy of 98.5 percent. The class labels applied to the lines were then used for annotating semi-structured text elements. TagLine achieved F-measure over 0.9 for each of the structures, which included tables, slots and fillers.

Introduction

The analysis of text from the electronic health record (EHR) is an important research activity in medical informatics and particularly in the Veterans Healthcare Administration (VHA) because of its large integrated EHR. A wide variety of methods have been employed to extract information from text. Information retrieval (IR) and information extraction (IE) have been shown to be useful for detecting patterns in patient care[1], patient treatment patterns[2] and outcomes. IR has been used to identify co-morbidities,[3] smoking status,[4] as well as detecting fall-related injuries.[5] Regular expressions have been used to extract blood pressure values from progress notes.[6] Natural language processing (NLP) has been used to extract medical information such as principal diagnosis[4] and medication use[7] from clinical narratives. This work has led to a better understanding of the conditions patients face and how to treat them.[8]

Raw medical text passages are voluminous and heterogeneous as is their structure.[9] Some of the information is in free-text form, written as full sentences or phrases, but much of it is in the form of semi-structured data, or templates.[10] Semi-structured data is defined as data that has some structure but is inconsistent or does not adhere to any rigorous format.[11] While some work has been done extracting information from semi-structured data, most of that research focused on extracting data from web pages or research articles and is not easily adapted to the medical domain.[12-15] Part-of-speech parsers in off-the-shelf NLP programs do not perform well on semi-structured data because it does not adhere to grammatical rules. If the structures within documents could first be accurately identified, then extraction methods that do not depend on English grammar could be developed to extract the information in these structures.

The goal of this study was to evaluate a method of processing information in semi-structured text in medical progress notes by first, classifying each line of text using machine learning, then using the line classifications in a rules-based parser, annotate the semi-structured text elements. This will allow the information in these structures to be further processed or stored in structured form. To achieve this goal, a prototype system "TagLine" was developed. We exploit non-grammatical features derived from the text in progress notes to apply a class label to each line of text and use these labels in a rule-based annotator to identify semi-structure text elements in the text. Our system combines methods already familiar in IE, such as concept look-ups, regular expressions, rules, and machine learning on features from the text to accurately identify information contained in semi-structured text elements.

Background

Information Extraction. IE is defined as the extraction of predefined types of information from text.[16] There are four primary methods available to implement an information extraction system, including NLP, pattern matching,

rules, and machine learning. The primary means of performing IE is NLP. NLP research focuses on developing computational models for understanding natural language.[17] The use of rules and pattern-matching exploits basic patterns over a variety of structures, such as text strings, part-of-speech tags, semantic pairs, and dictionary entries.[18] Regular expressions are effective when the structure of the text and the tokens are consistent, but tend to be one-off methods tailored to the extraction task. Hand coding of complex regular expressions can be a very time consuming effort that requires *a priori* knowledge of all possible patterns that represent the concept being sought. Machine learning techniques can be an effective method for IE through automated knowledge acquisition.[19] Features extracted from the text, such as parts of speech and sentence length, are fed into a learning machine to assist in tasks downstream like word sense disambiguation. The primary disadvantage of using machine learning is that it requires a labeled dataset for training a model.

Semi-Structured Data. Well-structured data, as found in a typical database, conforms to a schema or data model and can be queried using a structured query language to answer questions. Semi-structured data is data that has some structure but is inconsistent or does not adhere to any rigorous format,[36] and is very difficult to query. In semi-structured data, information normally associated with a schema is contained ***within*** the data, which is sometimes called "self-describing."[11] Semi-structured data can break the conventions for structured data in a number of ways. The structure is often irregular, implicit, or partial.

Table 1. Previous work on semi-structured data

Method Used	Document Type
Finite state machines[20]	HTML
Rule generation within specified constraints[21]	HTML
Example guided object decomposition[22]	HTML
Example guided structure induction[23]	HTML
Machine learning on object exchange models[24]	HTML
Schemas and wrappers on document structures[25]	HTML
Path expressions[26]	HTML
Labeled ordered trees on tag structure[27]	HTML
Ontology and graph based modeling[28,29]	HTML
Descriptive logics[30]	Journal
Graph modeling on schemas[31-33]	Journal
Link grammars connect features with numbers	EHR
ID3 trees on NLP features[34]	EHR
Standard data cleansing techniques[35]	EHR

Efforts to perform IE on semi-structured data are well developed for web pages but less so on research articles from peer-reviewed journals and notes for the electronic medical record. Table 1 shows a summary of the published methods previously used for extracting semi-structured data. Pages on the World Wide Web are written in HTML, which is a standard and provides a healthy measure of reliable structure that these studies used in developing extraction routines. Research papers also adhere to a certain amount of structure. There is an order to the flow of the paper and very specific formatting conventions with journals for section headers, graph labels and tables. For IE on semi-structured data in the electronic medical record, the work is limited and it has been noted that locating the data is difficult, since no standard way to enter the data in the EHR system is reinforced. Furthermore, there are no built-in edit checks available to facilitate data entry.

The Electronic Health Record at the VA. The VHA EHR, VistA, records information regarding a patient's clinical encounters, in both structured data tables and text. Each line of 80 characters or less is stored as a string associated with a specific document, such as a medical progress note. This preserves the formatting of the document to make it easier to read in the Computerized Patient Record System (CPRS),[37] the user interface for VistA. It also provides an artifact useful in text processing. Each note can be separated into a group of individual lines of text. Because progress notes are written for a variety of purposes, notes are assigned descriptive names. Users can create their own custom designed notes using the Progress Notes Construction Set[38] and design their own templates for the notes. This causes tremendous variation in the way notes are structured, which means developing extraction routines using techniques like regular expressions and handcrafted rules are typically useful only as one-off solutions.

Document element ontology. This study is guided by an ontology developed by investigators in the Consortium for Health Informatics Research (CHIR) to define the text elements to be targeted. The ontology described here is the result of an error analysis from the 2010 i2b2 challenge submission and a CHIR Information Extraction Methods (IEM) initiative.[39] A document is made up from a set of document elements. Sections, slot-value pairs, paragraphs, sentences, phrases, content, questions, lists, tables, and address blocks are examples of document elements. Figure 1 graphically depicts a sample of the structures and their component parts. For purposes of this study, we selected the text elements tables and slot-values. Table 2 presents a sample of the line-of-text classifications being used, along with the parent "is-a" class from the text element ontology and the larger structural (part-of) class. It gives the relationship of the line type to structure type in "is a" or "is part of" relationships. There were a total of 75 distinct class labels derived from the text and ontology.

Table 3. Sample line classes and their relationships to document elements.

Line Class	Is-A	Part-Of
Double Xed Items	CheckBox	Question
Multiple Xed Items	CheckBox	Question
Xed item	CheckBox	Question
Line list with Header	List	List
Comma separated list on a line	List	List
List Header	ListHead	List
Numbered Item	ListItem	List
Medication Footer	MedFooter	Table
Medication List Directions	TableRow	Table
Med List Item (not numbered)	TableRow	Table

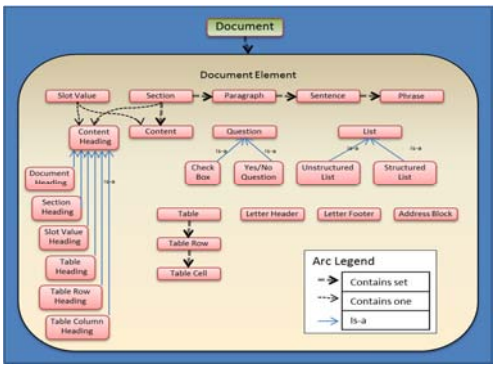


Figure 1. Document element ontology

The Line of Text as the Unit of Analysis. Progress notes in VistA are stored as a series of strings at a length of 80 characters or less. These text strings can be assigned recognizable roles as defined in the text element ontology. Some of these elements stand on their own as atomic text elements, while others are part of larger more complex text element structures. Both tables and slot values can be defined by their constituent parts. Slot values consist of a content heading and content separated by a delimiter and can be contained within a single line of text. Slot value content is a short value, typically a word, numeric value or phrase. The elements of a table can be identified by their parts at the line level. A table, as defined by the document element ontology, is a collection of related items arranged in columns and rows. Tables may have column labels and row labels as well as a caption for the subject of the table. The first line in Table 5, `"---- CBC PROFILE ----"` stands complete as the table header (THE). The second and third lines are used to label the information in the columns (CLA). The remaining lines are each identifiable table items (TBI) in the table.

Table 4. Table Example

Class	Line of Text			
THE		---- CBC PROFILE ----		
CLA	BLOOD	01/19		Reference
CLA		14:20	Units	Ranges
TBI	WBC	5.7	G/L	4.2-10.3
TBI	RBC	4.8	T/L	4.2-5.8
TBI	HGB	14.2	g/dL	14-17
TBI	HCT	43.3	%	39-50

Table 5. Slot Filler Examples

Class	Line of Text
SLF	CURRENT LEVEL OF PAIN: 5
DSV	LMP:11-26-06 PMP:221
TSV	Grava: 2 Para: 2 Abortion: 0
SLT	Indicate HOW YOUR PAIN FEELS:
FRT	Aching

The basic slot-filler consists of a label and a value separated by a delimiter, usually a colon (see Table 6). In its simplest form the slot-filler appears as shown on a single line. This line would be given the label "SLV" for slot-filler or slot-value pair. However, slot-fillers do not always appear in this form; there may be two or three sets of slot-fillers on the same line. These two examples are then labeled "DSV" and "TSV" respectively, for double slot-filler and triple slot-filler. The two elements, the label and the value may also appear on separate lines. The values that fill the slot are not always present, so there may only be a slot. Since each line tends to stand as a unit, we chose this as our unit of analysis in machine learning. Each of these variations must be handled by a specific set of parsing rules, and the class labels identify which set of rules to employ.

Table 6. Feature Examples

Feature	Description
AllCaps	All uppercase letters
Title	In title case
NumCaps	Number of uppercase letters
Hyphens	Number of hyphens
Spaces	Number of spaces
Slashes	Number of slashes
DecPos	Offset position of decimal
ColPos	Offset position of colon
QmPos	Offset position of "?"
Bar	Formatting bar
YesNo	Ends with "Yes" or "No"
Icd	Presence of IDC9 code
Bullet	Line is bulleted
Numbed	Line is numbered

Feature selection. The selection of features that were used in machine learning to assign classes to each line was crucial to success. We reviewed the work of other studies in this area[20,40-42] and derived and tested a number of additional features to detect structure in lines of text. Table 6 shows a sample of the features we adopted. We looked for similar clues in each of the text elements that help us tell them apart from other text elements. These clues fall into one of several types of text features: formatting features, special character usage, term usage, and document structural features. Examples of formatting features include whether the line was in all uppercase letters or in title case, as well as the number of uppercase letters in the line. Special character features

included items such as the total number of spaces, slashes, and hyphens in the line and where the first, second, and third decimal and colon could be found in the line. Examples of term usage-based features included position of a question mark, if the line of text ends with a "Yes" or "No" and if the token "ICD" was found in the line. Finally, document structural features included items such as if a line was numbered, had a text bullet, and if a line had a formatting bar (e.g., **********). In all, there were over 70 features defined to describe the differences between classes.

Data Preparation

Document Selection. A set of 162 notes formed the corpus for this study. These notes are a subset of the 5,048 medical progress notes collected in a separate and unrelated study in the VHA to identify patients who have suffered a fall-related injury[5]. The notes were randomly selected from note types containing the greatest number and variety of semi-structured text elements. The variety of note types selected represents note types that are most frequently used in the VHA for treatment of falls, primarily notes from emergency room and primary care visits. They included "Primary Care Notes", "Primary Care Nursing Notes", "Primary Care H&P Notes", "Primary Care Home Health Consult Notes," "Emergency Room Nursing Notes", "Emergency Room Triage Notes" and "Nursing Discharge Notes." These note types tend to have more structured elements than those written in free text. To evaluate machine learning on assigning line classes, all lines (n=15,103) from the selected notes were randomly split into sets of 10,000 and 5,103 for training and test respectively. To evaluate the structure annotator on identification of targeted semi-structured data, the 162 notes were randomly split into 115 notes for training and 47 notes for testing.

Labeling Lines of text. To provide data for machine learning each of the 15,103 lines in the 162 notes, classes were assigned to lines of text as an intermediate step to finding and extracting information from specific predefined types of semi-structured text elements. Class labels were determined by membership in or relationship to, the structure types defined in the CHIR ontology, and were derived empirically from the text. The classes may describe a part of a structure, such as a table or contain multiple structural parts. Class determination and text line labeling were iterative simultaneous tasks. First we examined the line for a relationship to a structure type in the ontology. If there was a class that describes this relationship, then we applied that class. If not, we created a new class within the ontology that describes the relationship and apply the class label. Structure types in the ontology can be associated with multiple classes of lines. This is grounded in the fact that any given structure type identified by the ontology may appear in the text in multiple forms as in the slot-fillers example above. Rigorous steps were taken to ensure that the classes were unique and that they related to only one text element in the ontology. Each class was evaluated individually and discussed as it was added to the set. As the number classes grew we evaluated their usefulness them by using them in some preliminary machine learning models. Classes that were misclassified were re-examined and if the label applied by the machine learning model was valid, it was relabeled with the predicted label. If the label was not valid, then we applied a new label and new features were added to the machine learning models to improve performance.

When the labeling was completed a count was made of each occurrence of each class and the distribution of the class frequencies was checked. It was found in initial models developed that classes with fewer than six instances in the dataset achieved an F-Measure of less than 0.6. Classes with less than six occurrences in the dataset were removed and the lines with those labels were re-labeled with the next best class. This required that the new class label be re-defined to include the new instances. There were a total of 75 possible class labels derived from the text and ontology. A total of 13 classes were eliminated from the list, leaving 58 remaining classes for use in machine learning models. None of the removed classes were relevant for the structures examined in this effort. The classes that most frequently appeared in the text were free text (FRT), slot-value (SLV), medication list item (MLI), and table item (TBI). Since the annotator uses the predicted line labels in its parsing routines, the errors in line label predictions are likely to cause errors in the structural annotations. It is therefore crucial to achieve the highest possible prediction accuracy in the first step.

TagLine

To accomplish the goal of this study, the prototype TagLine system was designed and implemented. TagLine consists of a series of interacting software modules written in Python. Shown in Figure 2, the paths in yellow show the flow sequence for training a new model and the blue paths show the sequence for structure annotation.

Extraction Module. Notes are converted to a list of text lines and each text line is subjected to a series of functions that extract values for the features to be used in machine learning. When a new model is trained, all features are selected for extraction. The final model determines which features will be extracted when using the system for annotation. The resulting new dataset was used to train the learning machine and make predictions. Each line in the dataset described a line of text in the note. This dataset was sent to the C5.0 module when training a new model or to the next stage for use in classification by the tree classifier.

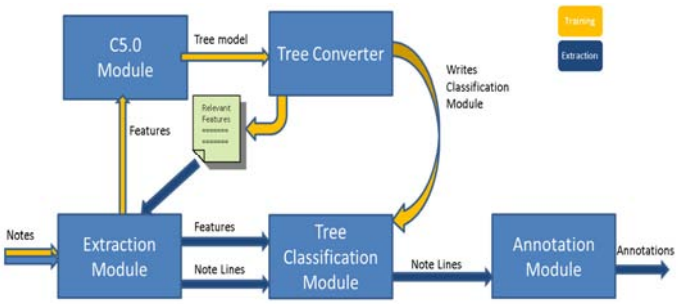


Figure 2. TagLine system architecture

C5.0 module. Decision trees are a good fit for the work in this study for the following reasons[43]: They handle high dimensional data well, are computationally efficient, easily interpreted by human beings, and most importantly for this study, can be readily converted to computer executable code. This study uses an updated version of the C4.5 algorithm developed by Ross Quinlan as C5.0.[44] The C5.0 module is a Python wrapper for a console application. The C5.0 application was compiled from the GPL C code distributed freely by Ross Quinlan. The C5.0 module develops a decision tree model based on the features provided by the extraction module. The C5.0 algorithm performs winnowing of the attributes before building a tree by evaluating each attributes' effect on error rate when the attribute is removed. Error-based pruning is also performed to cut back on branches that do not contribute to the models overall efficacy. Finally, this module writes a tree model file out to disk for use later.

Tree converter module. The tree converter parses the original tree model file written by the C5.0 module and writes two new files; a text file and a Python executable classification module. The text file is a list of the features determined by the C5.0 decision tree model found to be useful in prediction. When the new classification module is used, the extractor will only extract those features that are needed.

Tree classification module. The tree classification module is automatically created by the tree converter module by parsing the tree model and writing executable Python code. This module takes the note as a list of text lines and classifies them using a series of if-then rules. Then it passes the classification results along with the lines of text to the annotation module.

Annotation module. The annotation module takes the classification results and the lines of text from the tree classification module and uses them to locate structures in the text based on options submitted by the user. For each type of structure indicated in the options, the annotation module loops through each line of the note looking for the appropriate line labels for the targeted structures. When targeted labels are encountered, a rule-based approach is used to parse and annotate the structure and return values in the form "ElementType|StartOffset|StopOffset." The annotations can be written to a file for review, recorded in a database for storage and used later as structured data, used as features in another classification task downstream, or sent to an NLP pipeline where annotations can be used for extracting concepts from the text elements using a structured vocabulary. TagLine can also extract the elements and record the notes with the annotated structures removed as a text reduction method.

TagLine Evaluation

TagLine was evaluated in two separate experiments. First, we evaluated the use of decision trees for predicting the classes assigned to the lines of text, then we evaluated the accuracy of our rules based annotator for identifying slot-value pairs and tables based on the classes assigned to the lines.

Results for Line Classification. A decision tree was constructed on the line-level training data and evaluated on unseen test data for prediction accuracy. Winnowing was used for feature reduction before a tree was constructed. Winnowing removes any feature that does add to the models efficacy. The C50 algorithm constructs an initial model on half of the training data and calculates the increase in error rate for each feature when it is left out of the model. Global pruning was performed after tree induction keeping only those branches that had at least one instance associated with it. Error based pruning was also employed; a branch was pruned if its prediction errors exceeded a

level of 25 percent. A series of fifty models were tested using graduating numbers of lines in increments of 200 examples starting at 200 and ending with 10,000. The overall prediction accuracy was calculated for each test using a hold out set of 5,103 lines. The results are presented as a learning curve in Figure 4.

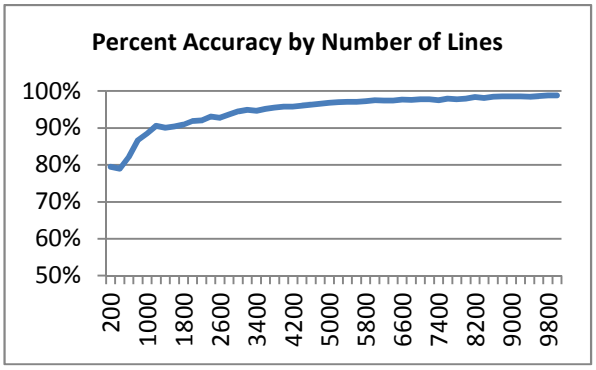


Figure 3. Learning Curve for Machine Learning

Table 7. Line Classification Statistics

Class	N	Precision	Recall	F-Measure
Free-text	934	0.9789	0.9936	0.9862
TableItem	900	0.9944	0.9878	0.9911
Slot-filler	508	0.9882	0.9882	0.9882
MedLine	362	0.9564	0.9696	0.9630
Slot	273	1.0000	0.9963	0.9982
AphaLine	6	0.8333	0.8333	0.8333
NumbedQue	4	0.8000	1.0000	0.8889
NumbMed	3	0.6667	0.6667	0.6667
LabeledDS	2	0.6667	1.0000	0.8000
QuestHead	1	0.5000	1.0000	0.6667

With a sample size of 200 lines, the decision tree model was able to achieve an accuracy rating of 80 percent. The accuracy increased to 90 percent at 1,200 lines. An overall accuracy level of 98.5 percent was achieved with a sample set of 10,000 lines, only 3.5 percent higher than at the 3,000-line sample size. The performance results for the five most and least frequent classes in the test set are shown in Table 7. Only five classes did not achieve an overall F-measure of 0.9 or above.

Prior to tree construction, the winnowing process eliminated 20 features. The top 10 remaining features are shown below in Table 8 with their respective importance ratings. The importance rating is C5.0's estimate of the factor by which the true error rate or misclassification cost would increase if that attribute were excluded. The number of colons in a line of text appears to be the most important feature in the model. Colons are important when looking for slot-value pairs, as well as dates and timestamps. The second most predictive feature is the line number for the line of text. Line numbers describe how far into the note a line appears; beginning, middle or end. Interestingly, all of the tests done by the model using the line number feature took place at the top of the note (see figure 4). Question marks, capital letters, and white space gaps also help in distinguishing structured text from un-structured text.

Table 8. Top 10 Features

Importance	Feature
835%	Colons
264%	LNum
255%	Slot
179%	LSpc
166%	Bull
150%	Med
148%	Gaps
139%	QM
137%	Caps
127%	Time

```
QM > 0:
:...Quest > 0:
:   :...LNum > 0: NQU (10)
:   :   LNum <= 0:
:   :   :...Colons <= 0:
:   :       :...SLow <= 0: QUE (268)
:   :       :   SLow > 0: QUF (28)
:   :       Colons > 0:
:   :       :...UpSlot <= 0: HQU (4)
:   :           UpSlot > 0: SLV (1)
```

Figure 4. Decision Tree Fragment

The occurrence of a question mark was the most used feature in the model. This is shown in the tree fragment in Figure 4 where the feature QM is at the top of the decision tree, indicating it is the first attribute the tree splits on.

The tree classification module was used to predict the 5,103 line class labels in the test data. Many classes achieved a perfect score. The poor performers were all low prevalence classes, each occurring less than ten times in the entire test set. The class ALI is alpha list item, or an item in a list that is delineated by an alpha character and some delimiter. One line labeled ALI was misclassified as FRT causing one false negative. The only feature that would distinguish Free-text from an alpha-labeled-line is the use of the delimiter after the delineating character. In most applications a 95 percent accuracy rate would be considered acceptable, so it would not be necessary to use more than 3,000 lines for an acceptable result. However, because the results of the annotation in TagLine is dependent on the accuracy of the line labels it was decided to use the full 10,000 lines for the training set and the remaining for testing in the next section.

Results for Annotating Tables and Slot-Fillers. For this experiment, the data were split into training and test sets segregated at the note level. In the next stage, the parsing routines were tested on two types of structures; slot-value

pairs and tables. Since tables are multi-line units, all lines are needed to identify a unit. There were a total of 115 notes and 10,048 lines of text in training set, while the test set had the remaining 47 notes and 5,055 lines of text. A record was constructed noting the number of tables, slots and fillers in the 47 notes. Table 6 shows the occurrences of each structure in the note set. The test set of notes has a total of 96 tables, 770 slots and 566 fillers for the slots. Not all of the slots have fillers associated with them so the numbers of slots and fillers will not match. The complete set of 47 notes was used in a GATE pipeline that called TagLine using a remote procedure call (RPC) server.

As the notes were processed, the server recorded the annotation actions. The extraction results were compiled for evaluation. Table 9 summarizes performance of TagLine on tables, slots, and slot fillers. The TagLine server reads the options, extracts the features from the note, uses the prediction module constructed from the decision tree, applies class labels to the lines, and returns the annotations on the structures found. When tables are targeted, start and stop rules are used to identify the boundaries of the table. The routine sequentially examines the labels applied to each line and when a class is encountered that signals the beginning of a table, a flag is set and all successive lines are included until a line label signaling a stop rule is encountered and the table end boundary is set and the annotation is returned. If a line in the middle of the table is classified as free text, it would prematurely trigger the table's end and close off the boundaries of the table and erroneously start a new table at the next table line. For this reason, a rule was included to allow for a one-line misclassification gap in the table. While traversing the lines of text after a table beginning has been encountered, if a line is found that is not a table line, the end of table is marked, but held until the next line is checked. If the line after the non-table line is a table line then the table end is cleared and the boundary of the table is extended until there are no more table lines. In the event the next line is a table header or a column labels, then the current table is completed and a new table begins. A constraint enforced by this method is that a table must have at least one table line (TBL) to be annotated as a table.

Table 9. TagLine Annotation Performance

		Exact Match			Partial Match		
Structure	**Count**	**Precision**	**Recall**	**F-Measure**	**Precision**	**Recall**	**F-Measure**
Tables	96	0.9250	0.9250	0.9250	0.9896	0.9896	0.9896
Slots	770	0.9709	0.9974	0.9840	0.9735	1.0000	0.9865
Fillers	566	0.9543	0.9965	0.9749	0.9560	0.9982	0.9767

Discussion

TagLine performed well on slots and fillers. There was little difference when evaluating performance on exact versus partial matching criteria. There were 770 slots in the 126 notes used for the test. TagLine found 768 of the slots matching the start and stop end points exactly. The two remaining slots were found but the parsing routine failed to set the offsets properly. One was due to colon placement in the string. The text line in the middle of the note: "`(R): 0.7cm X 0.5cm  (L): cm X cm`" was marked as a DSV, or a double slot value. The "`(R):`" is the slot, and the filler is "`0.7cm X 0.5cm`." TagLine failed to parse this line appropriately for both of the slots and fillers. While they were found, they were not considered exact matches. There were 21 false positives for slots. A total of 19 of the false positives were due to date entries across several notes that were misclassified as slot-values. There were a total of 566 fillers in the 47 test notes. TagLine successfully annotated all but one of the fillers, but there were 26 false positives, 19 associated with the date misclassifications mentioned above, resulting in a lower precision (0.9543) than recall (0.9965). Figure 5 shows the slot-filler annotations highlighted for one of the test notes.

There were 96 tables distributed throughout the 47 test notes. TagLine achieved an F-measure of 0.9250 for exact matching, lower than the number achieved for slots and fillers, but encouraging. Of the 96 tables in the test set, 86 were matched exactly according to the start and stop codes. Of the remaining 10 tables, 9 were identified with partial matches and one table was missed completely. Many partial match cases were due to column labels (CLA), or table item (TBI) being misclassified as free text (FRT), causing the parser to miss that portion of the table. In Figure 6, it can be seen that three lines in the table were not annotated. The first of the three is the column label and the next two descriptive entries did not conform to the format of the other table lines and were labeled FRT. Because the lines "`color yellow`" and "`appeara sl cldy`" do not conform to the format of the rest of the table, the table was partially captured in two parts. An allowance is made in the case that one line is misclassified, so tables are not broken apart. However, if more than one line is misclassified then the table will only be a partial match. Making allowances for more than one misclassified line causes problems when tables are found stacked directly on top of another table. In these cases, the two tables erroneously become one contiguous table.

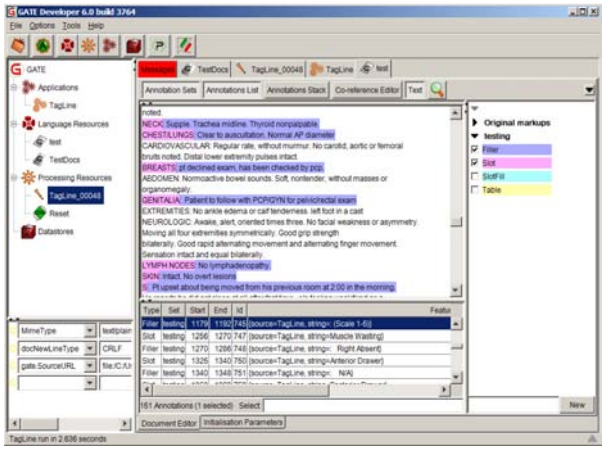


Figure 5. Slot-filler Annotation Example in GATE

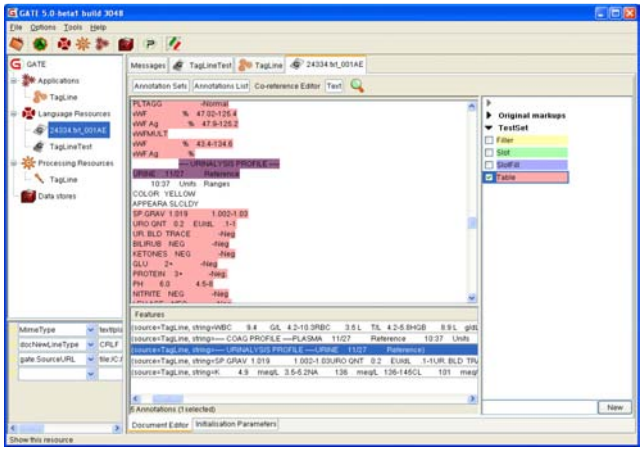


Figure 6. Table Annotation Example in GATE

Conclusions

TagLine was evaluated for parsing and annotating semi-structured text elements both at the macro level, on multi-line structures, and at the micro level, on within-line structures. TagLines' performance identifying slots and fillers was impressive. The annotator achieved an F-measure of 0.9840 for exact matching criteria for slots and 0.9965 for fillers. The results for tables were not as impressive, achieving an F-measure of 0.925 for exact matches and 0.9896 for partial matches. Since even partial matches are likely to be useful in practice, the results were quite good.

Contributions. In this study, we have shown how using the line of text as the unit of analysis can play an important role in semi-structured elements and that text analysis can be beneficial at this level. TagLine has been shown to be effective for distinguishing specific classes of lines based on their structural roles and for identifying semi-structured text elements in medical progress notes. This will enable researchers to use more of the information in the text than was possible before. The information in the annotations can be stored in a database and used for other analyses as fully structured data. Once identified, semi-structured text elements can also be removed from the document so NLP can focus on the free-text sections which may result in more accurate concept extraction, as well as faster processing times for each note. The counts of these structures could be used as features in a machine learning approach for document classification tasks. TagLine can now be used to do ad hoc concept extraction from the semi-structured text by using rule as implemented in the JAPE module of GATE. Rules such as "if slot = <search term> lookup <filler>."Only those structures that contained the search terms or concepts would be annotated and returned. Since more information can be extracted from the notes, more complete information is available for patient analysis or document classification. This additional data could enable researchers to explore topics better and perhaps improve the healthcare for veterans easier and faster than before.

Limitations. This study has two primary limitations. First the dataset, while adequate for purposes of providing sufficient examples for developing and evaluating learning machines, is still limited since the sample was taken from an existing study and may not include many note types and their specific challenges. TagLine needs further testing and development to ensure good generalizability. Also, samples of documents from other hospitals should to be included in the corpus to train models for increased generalizability.

Future Work. The parser in TagLine will be expanded to include other types of structures like full templates, questions and checkboxes. The language across note sections tends to differ.[45] This feature is often used in NLP systems for "word sense disambiguation," and is a fertile area for research[46]. There is a need to improve section identification or section header identification.[47,48] We will also extend Tagline to function as a "sectionizer," which accurately determines which section of the note a line of text belongs to. A sectionizer would be useful in word sense disambiguation and for text reduction, potentially saving significant amounts of time and money on creating annotated data sets. Human annotators would have smaller, more concentrated notes to cover if it is known that certain sections held no interest. This may increase productivity by shortening the time necessary to cover the note

and decrease the likelihood of mental fatigue that occurs with longer notes, as well as reducing the amount of time spent on irrelevant documents.

Funding for this work was provided by the Veterans Healthcare Administration Health Services Research & Development grants IIR05-120-3 and SDR HIR 09-002. The views expressed in this paper are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the US government.

References

1. **Pakhomov S, Bjornsen S, Hanson P, Smith S. Quality performance measurement using the text of electronic medical records. *Med Decis Making.* Jul-Aug 2008;28(4):462-470.**
2. **Rao RB, Krishnan S, Niculescu RS. Data mining for improved cardiac care. *SIGKDD Explor. Newsl.* 2006;8(1):3-10.**
3. **Guillen R. Identifying Obesity and Co-morbidities from Medical Records. AMIA Symposium; November 14-18, 2009; San Fransisco.**
4. **Zeng Q, Goryachev S, Weiss S, Sordo M, Murphy S, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC Medical Informatics and Decision Making.* 2006;6(1):30.**
5. **McCart JA, Berndt DJ, Jarman J, Finch DK, Luther SL. Finding falls in ambulatory care clinical documents using statistical text mining. *Journal of the American Medical Informatics Association.* December 15, 2012 2012.**
6. **Turchin A, Kolatkar NS, Grant RW, Makhni EC, Pendergrass ML, Einbinder JS. Using regular expressions to abstract blood pressure and treatment intensification information from the text of physician notes. *Journal of the American Medical Informatics Association.* 2006;13(6):691.**
7. **Xu H, Shane S, Doan S, Johnson K, Waitman L, J D. Extract Medication Information from Clinical Narratives. AMIA Symposium; November 14-18, 2009; San Fransisco.**
8. **Cerrito P. Data and text mining the electronic medical record to improve care and to lower costs. Paper presented at: SUGI-312006; San Fransisco, CA.**
9. **Honigman B, Lee J, Rothschild J, et al. Using computerized data to identify adverse drug events in outpatients. *Journal of the American Medical Informatics Association.* 2001;8(3):254.**
10. **Chowdhury G. Template Mining for Information Extraction from Digital Documents. *Library Trends.* 1999;48(1):182-208.**
11. **Buneman P. Semistructured data. Paper presented at: Syposium on Principles of Database Systems1997; Tucson, AZ.**
12. **Singh L, Chen B, Haight R, Scheuermann P. An algorithm for constrained association rule mining in semi-structured data. *Lecture Notes in Computer Science.* 1999:148-158.**
13. **Chang CH, Hsu CN, Lui SC. Automatic information extraction from semi-structured web pages by pattern discovery. *Decision Support Systems.* 2003;35(1):129-147.**
14. **Bratko A, Filipi B. Exploiting structural information for semi-structured document categorization. *Information Processing and Management.* 2006;42(3):679-694.**
15. **Hemnani A, Bressan S. Extracting information from semi-structured Web documents. *Advances in Object-Oriented Information Systems.* 2002:389-396.**
16. **DeJong G. An overview of the FRUMP system. *Strategies for natural language processing.* 1982:149-176.**
17. **Hayes PJ, Carbonell J. Natural Language Understanding. *Encyclopedia of Artificial Intelligence.* 1987:660-677.**
18. **Pakhomov S, Buntrock J, Duffy P. High throughput modularized NLP system for clinical text. 2005.**
19. **Lehnert W, Soderland S, Aronow D, Feng F, Shmueli A. Inductive text classification for medical applications. *Journal of Experimental and Theoretical Artificial Intelligence.* 1995;7:49-49.**
20. **Hsu J, Yih W. Template-based information mining from html documents. Paper presented at: National Conference on Artificial Intelligence1997.**
21. **Singh L, Chen B, Haight R, Scheuermann P, Aoki K. A robust system architecture for mining semi-structured data. Paper presented at: International Conference on Knowledge Discovery and Data Mining1998.**

22. **Ribeiro-Neto B, Laender A, da Silva A. Top-down extraction of semi-structured data. Paper presented at: String Processing and Information Retrieval Symposium and Internationa; Workshop on Grioupware1999.**
23. **Ribeiro-Neto B, Laender A, Da Silva A. Extracting semi-structured data through examples. 1999.**
24. **Xufa C. Semi-structured Data Extraction and Schema Knowledge Mining. Paper presented at: EUROMICRO 1999.**
25. **Gao X, Sterling L. Semi-structured data extraction from heterogeneous sources. 2000.**
26. **Taniguchi K, Sakamoto H, Arimura H, Shimozono S, Arikawa S. Mining semi-structured data by path expressions. *Lecture Notes in Computer Science.* 2001:378-388.**
27. **Kashima H, Koyanagi T. Kernels for semi-structured data. 2002.**
28. **Wessman A, Liddle S, Embley D. DW: A generalized framework for an ontology-based data-extraction system. 2005.**
29. **Imtiaz H, Darlington J, Zuo L. Ontology Driven Web Extraction from Semi-structured and Unstructured Data for B2B Market Analysis. 2009.**
30. **Calvanese D, De Giacomo G, Lenzerini M. What can knowledge representation do for semi-structured data? Paper presented at: Artificial Intellignce/Innovative applications of Artificial Intelligence1998; Madison, WI.**
31. **Calvanese D, De Giacomo G, Lenzerini M. Semi-structured data with constraints and incomplete information. *Networking and Information Systems Journal.* 1999;2:253-273.**
32. **Calvanese D, De Giacomo G, Lenzerini M. *Queries and constraints on semi-structured data.* Vol 1626/2009. Berlin: Springer; 1999.**
33. **Calvanese D, De Giacomo G, Lenzerini M. Modeling and querying semi-structured data. *Networking and Information Systems Journal.* 1999;2:253-273.**
34. **Zhou X, Han H, Chankai I, Prestrud A, Brooks A. Converting semi-structured clinical medical records into information and knowledge. Paper presented at: Proceedings of the International Workshop on Biomedical Data Engineering2005.**
35. **Kristianson K, Ljunggren H, Gustafsson L. Data extraction from a semi-structured electronic medical record system for outpatients: A model to facilitate the access and use of data for quality control and research. *Health Informatics Journal.* 2009;15(4):305.**
36. **Abiteboul S. Querying Semi-Structured Data. 1997.**
37. **Fletcher RD, Dayhoff RE, Wu CM, Graves A, Jones RE. Computerized medical records in the Department of Veterans Affairs. *Cancer.* 2001;91(S8):1603-1606.**
38. **Brown S, Hardenbrook S, Herrick L, St Onge J, Bailey K, Elkin P. Usability evaluation of the progress note construction set. 2001.**
39. **Divita G. A Medical Document Text Element Ontology Paper presented at: 2011 AMIA Annual Symposium2011; Wash DC.**
40. **Chieu HL, Ng HT. A maximum entropy approach to information extraction from semi-structured and free text. 2002.**
41. **Pakhomov SV, Ruggieri A, Chute CG. Maximum entropy modeling for mining patient medication status from free text. *Proc AMIA Symp.* 2002:587-591.**
42. **Rao BR, Sandilya S, Niculescu R, Germond C, Goel A. Mining time-dependent patient outcomes from hospital patient records. *Proc AMIA Symp.* 2002:632-636.**
43. **Korting TS. C4. 5 algorithm and Multivariate Decision Trees. *Image Processing Division, National Institute for Space Research–INPE Sao Jose dos Campos–SP, Brazil.***
44. **Quinlan JR. Rulequest Free Software Downloads. [http://www.rulequest.com/download.html]. 2013. Accessed August 1, 2011.**
45. **Zeng Q, Redd D, Divita G, Jarad S, Brandt C. Characterizing Clinical Text and Sublanguage: A Case Study of the VA Clinical Notes. *J Health Med Informat S.* 2011;3:2.**
46. **Patterson O, Igo S, Hurdle JF. Automatic Acquisition of Sublanguage Semantic Schema: Towards the Word Sense Disambiguation of Clinical Narratives. 2010.**
47. **Denny JC, Miller RA, Johnson KB, Spickard III A. Development and evaluation of a clinical note section header terminology. 2008.**
48. **Denny J, Spickard III A, Johnson K, Peterson N, Peterson J, Miller R. Evaluation of a Method to Identify and Categorize Section Headers in Clinical Documents. *Journal of the American Medical Informatics Association.* 2009:M3037v3031.**

Predicting Electrocardiogram and Arterial Blood Pressure Waveforms with Different Echo State Network Architectures

Allan Fong, MS[1,3], Ranjeev Mittu, MS[2], Raj Ratwani, PhD[3], James Reggia, MD, PhD[1]
[1]University of Maryland, College Park, MD; [2]Naval Research Laboratory, Washington DC; [3]National Center for Human Factors in Healthcare, Washington DC

Abstract

Alarm fatigue caused by false alarms and alerts is an extremely important issue for the medical staff in Intensive Care Units. The ability to predict electrocardiogram and arterial blood pressure waveforms can potentially help the staff and hospital systems better classify a patient's waveforms and subsequent alarms. This paper explores the use of Echo State Networks, a specific type of neural network for mining, understanding, and predicting electrocardiogram and arterial blood pressure waveforms. Several network architectures are designed and evaluated. The results show the utility of these echo state networks, particularly ones with larger integrated reservoirs, for predicting electrocardiogram waveforms and the adaptability of such models across individuals. The work presented here offers a unique approach for understanding and predicting a patient's waveforms in order to potentially improve alarm generation. We conclude with a brief discussion of future extensions of this research.

1. Introduction

Intensive Care Units (ICUs) are designed to handle some of the most physiologically fragile patients in the hospital. As a result, ICUs utilize a wide spectrum of machines, technologies, and tests to help medical staff better understand and care for patients. However, the wide array of stand-alone machines often collect data and produce alarms and alerts independently, leaving the difficult integration tasks for the medical staff[1]. Time sensitive decisions, including identifying non-critical alarms, are just some of the problems faced by ICU medical staffs. Studies have shown that staffs in ICUs face an extraordinary number of alarms each day, some as many as 1,000 alarms a day, many of which are non-actionable or not necessary for patient care[2,3]. Excess amounts of non-critical alarms can lead to alarm fatigue which can adversely affect patient care[4-6]. While there is ongoing research to effectively minimize false alarms, such as allowing nurses to adjust alarm thresholds, much work is still needed to improve classification techniques and systems in order to reduce false alarms[5,7]. A remaining challenge is to develop algorithms robust enough to understand, integrate, and predict multiple physiological waveforms from patient data to better classify and interpret alarms.

The purpose of this paper is to explore the value of using of Echo State Networks (ESN), a particular type of recurrent neural network, to predict an individual's waveforms. Being able to forecast an individual's waveform could potentially offer much more information to the medical staff in addition to alarm classifications. ESNs were chosen because of their ability to accurately predict chaotic time series[8-11]. In our research, these networks were trained to predict an individual's electrocardiogram (ECG) and arterial blood pressure (ABP) waveform data, which can potentially help prioritize alarms as well as predict life-threatening situations in the ICU. Our research uses clinical ICU patient data to develop, train, and test various ESN architectures for prediction tasks, and establishes the benefits of using ESN architecture designs for predicting ECG and ABP waveforms. Waveform and alarm classification and prediction is very important and we hope that this work will be helpful in providing additional insight into this problem.

2. Background

The recent release of clinical ICU patient data Multiparameter Intelligent Monitoring in Intensive Care II(MIMIC II) makes it possible to develop better models and support tools to aid medical workers in understanding and filtering the abundance of information and alarms[11]. This publicly available database has already been analyzed and used in several ways, for example to develop decision support systems to better categorize and classify mortality rates in the ICU[6, 11-14]. A study previously used this data and an expert review panel to reclassify five common ICU alarms into true alarms and false alarms. These authors then developed an algorithm that classified alarms based on both electrocardiogram (ECG) and arterial blood pressure (ABP) waveforms immediately prior to the machine generated alarms. When tested, the algorithm suppressed approximately 59.7% of the false alarms and 0% of true alarms (except for true ventricular tachycardia alarms which was reduced by 9.4%)[15]. While their algorithm is focused on classification, this project aims to complement their work by developing a neural network that can learn to predict

the more continuous waveforms. An effective predictive model can help medical staff better anticipate a patient's condition, which includes the occurrence of alarms and false alarms.

It is difficult to predict time series data, especially with chaotic waveforms such as ECG and ABP. Previous studies that have focused on predicting ECG and ABP waveforms have used structures or approaches that simplified the waveforms, usually with higher order measures[15-19]. While there have been previous studies modeling and predicting complex time series, only a few studies have explored the relationship between ECG and ABP waveforms during classification or prediction tasks[15, 20-24]. Our work explores the use of ESNs, a type of recurrent neural network, to explore this ECG-ABP relationship further. Neural networks have been shown to be good at predicting time series data, especially in situations where building proper heuristic models is difficult[25-27]. ESNs were chosen for this research because they have been previously shown to accurately predict chaotic time series without the need to train the specific internal representations of the system[9, 10]. This computational advantage makes ESNs very attractive for predicting ECG and ABP time series which are both chaotic and difficult to learn.

2.1 Echo State Network

An Echo State Network (ESN), Figure 1, is an example of a recurrent neural network capable of modeling and predicting non-linear behaviors[9, 10]. A typical ESN has four sets of unique weights: W_{in_hidden}, W_{hidden}, W_{hidden_out}, and W_{in_out}. W_{in_hidden} are randomly assigned fully connected weights between the input node to the reservoir nodes. A distinct property of ESNs is the sparely connected, randomly assigned weights between the reservoir nodes, W_{hidden}. These sparsely connected nodes allow for pockets of local resonances, or echoes, to develop, which together can model complex waveforms[9,10]. Furthermore, W_{hidden_out} are randomly assigned fully connected weights from the reservoir nodes to the output node which contributes to the learning of the network from the teacher (or training) signal. Unlike more typical recurrent neural networks, ESNs do not have connections from the output nodes back to the hidden nodes; this greatly reduces the model complexity and convergence time for ESNs[10].Lastly, W_{in_out} are randomly assigned weights from the input node to the output node.

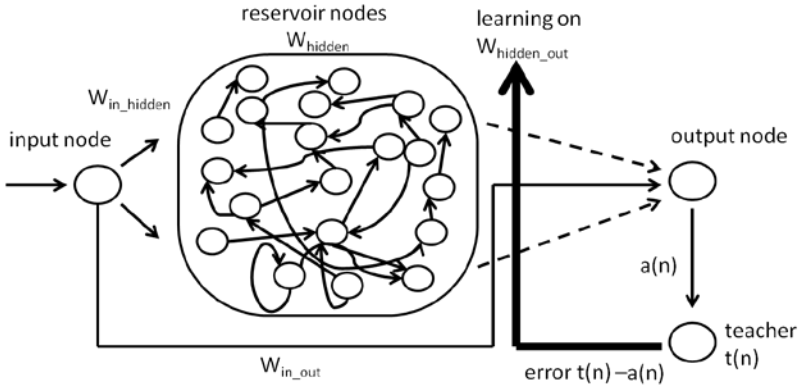


Figure 1: ESN architecture with learning on W_{hidden_out} (dashed arrows).

3. Methods

We comparatively evaluate the performance of three different types of ESN architectures at predicting two related physiological waveforms. The following sections first describe the data and preprocessing of the data. Next we discuss the analysis to identify reasonable ranges for key model parameters (reservoir size, activation rule, and learning rates). We then describe the three ESN architecture designs and our evaluation criteria.

3.1 Data Source and Pre-Processing

The data used for this project comes from the MIMIC II database which is publicly available[11, 14, 28]. The complete database currently contains data from approximately 33,000 de-identified patients collected over 7 years (beginning in 2001) from Boston's Beth Israel Deaconess Medical Centers. It combines both clinical and physiological data. The adult patients range in age from 18 to over 90 years old (mean 68 years), and were collected from 48 medical, surgical, and coronary intensive care beds. Each patient record typically contains data from two electrocardiogram (ECG) leads, arterial blood pressure (ABP) and pulmonary arterial pressure (PAP) stored at 125Hz over time intervals that can range between a few hours to a few days. The ECG was originally sampled at 500Hz but was compressed to 125Hz while still preserving the peaks[15]. The resulting database is quite large (over 3TB). Because

we were interested in evaluating ESNs for individual patients, we focused on data from the ECG II (ECG lead II) and ABP readings from six randomly selected patients. ECG II data was selected because it appeared to be more available from a cursory look at the patient records. ABP was selected because of its relationship to ECG in the interpretation and classification of alarms[15].

Although the data is publicly available, a specialized WaveForm DataBase (WFDB) software package was required to download, interpret, and format the data[29]. Cygwin was used to connect directly to the server to download and format the data. The downloaded data was converted to comma separable version files which were compatible with Matlab. Some basic preprocessing was needed to make the magnitudes for the two waveforms comparable. A simple smoothing function, that averaged the data points in a 5-time step moving window, was applied to both the ECG and ABP data. This window size was effective at smoothing the waveform while maintaining the important features of the waves. The ABP data was also normalized to fall within the values 0 and 1. The ECG was vertically shifted up by the minimal value so it would be within the same range as the ABP data. These transformations were necessary in order to make the two waveforms similar in magnitude while maintaining unique features to allow for comparison.

3.2 Defining Baseline Model

An Echo State Network (ESN), Figure 1, was first built in Matlab based on previous work[9, 10]. Weights were initially assigned random values between -0.5 and 0.5. Similar to previous work, approximately 20% of the possible connections in the reservoir have non-zero weights and are scaled with a spectral radius of 0.98, using:

$$W_{hidden} = \frac{\alpha W'_{hidden}}{|\lambda_{max}|}$$

where α is a scaling factor, W'_{hidden} is the weight matrix for the reservoir prior to transformation, and $|\lambda_{max}|$ is the maximum eigenvalue of W'_{hidden}, i.e., the spectral radius[10].

The activation for the hidden nodes, A_{hidden}, and output nodes, A_{out}, is[10]:

$$A_{hidden}(t) = tanh(W_{in\_hidden}A_{in}(t) + W_{hidden}A_{hidden}(t\text{-}1))$$

$$A_{out}(t) = W_{in\_out}A_{in}(t) + W_{hidden\_out}A_{hidden}(t)$$

The learning rule and training only applied to the connections between the hidden nodes and the output nodes; all other weights remained unchanged through initialization, training, and testing. Although various learning techniques to train these weights were tried, such as linear regression and simplified error back propagation, a simple delta learning rule that incrementally changed the weights based on the product of the learning rate and the training error was shown to be both effective and fast. To prevent excessive oscillations in weights, a minimal error threshold was applied such that weights would not change if the absolute value of the error was less than 0.0001 (determined empirically).

To help validate this initial implementation, it was first tested by training it to model a simulated sine wave. The data was divided into training data (2,000 time-steps) and testing data (1,000 time-steps). The ESN, with 600 reservoir nodes and a learning rate of 0.0001, was initialized by passing the simulated sine wave through the reservoir once to let the internal system transients dissipate. Next, the training data was introduced to the network and the hidden-to-output weights were allowed to learn. The goal was to develop a model that could predict a waveform; hence the teacher signal to be predicted was the input signal 100 time-steps to the right (i.e., 100 time-steps in the future). This trivial example demonstrated the basic workings of this network. Mean Square Error (MSE) was used to evaluate this and subsequent test predictions:

$$MSE = {}^{1}/_{N}\sum_{n=1}^{N}\big(t(n) - a(n)\big)^2$$

where N is the total number of time-steps in the test prediction, *t(n)* is the actual teacher value at time-step *n*, and *a(n)* is the output predicted value at time-step *n*. This trivial but useful prediction demonstration resulted in a low MSE test of 0.025.

3.3 ESN Architecture Designs

Three ESN architectures were designed to predict ECG and ABP waveforms. Three architectures were chosen to explore how coupling and integrating related waveforms effects overall predictive performance, Figure 2. The first

ESN consisted of two independent reservoirs (ESN1) which served as a control for the experiment. This architecture consisted of two ESNs time-synchronized and running in parallel with no connections between the networks. These reservoir sizes were determined experimentally as discussed later. The second coupled reservoir architecture (ESN2) was similar to the first architecture. However, the nodes in both reservoirs were connected to each of the output nodes. The learning of the hidden to output weights were specific to the error generated by the corresponding waveform. For example, only the error between the predicted and actual ECG waveforms was used to update the W_{hidden_out} connection to the ECG output node. Both reservoirs were initialized synchronously with their corresponding waveform. It was interesting to investigate if the predictions of one reservoir could benefit from the other reservoir with this architecture, and whether two separate reservoirs would make the overall network more robust to limitations associated with the randomly assigned weights.

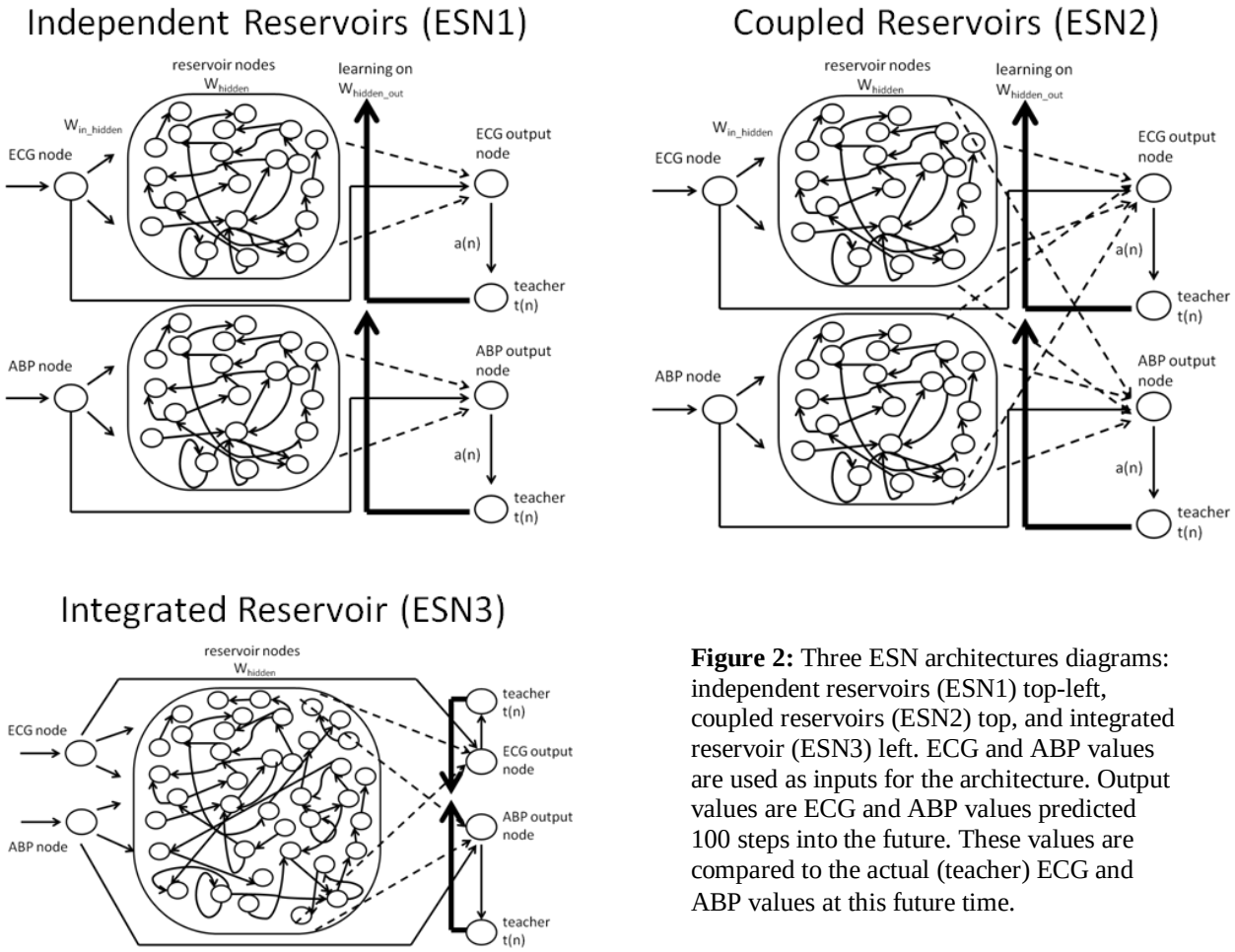


Figure 2: Three ESN architectures diagrams: independent reservoirs (ESN1) top-left, coupled reservoirs (ESN2) top, and integrated reservoir (ESN3) left. ECG and ABP values are used as inputs for the architecture. Output values are ECG and ABP values predicted 100 steps into the future. These values are compared to the actual (teacher) ECG and ABP values at this future time.

The third architecture consisted of one large integrated reservoir network (ESN3) with two input and two output nodes. This architecture was chosen to investigate how single reservoir systems compared to multiple reservoir systems when predicting multiple related waveforms. It was interesting to test if initializing and training one reservoir with two related waveforms could lead to better performance. To be consistent with the independent and coupled reservoirs, both the inputs and outputs were fully connected to the hidden nodes but inputs were only connected to their corresponding output waveform node. The other network parameters, such as spectral radius, W_{hidden}, sparsity, were the same with the independent and coupled reservoirs architectures. This ensured that the analysis would primarily focus on differences resulting from the network architecture. The integrated reservoir network was initialized with both waveforms, and similar to coupled reservoirs architecture, the fully connected

weights from the hidden to each output node were different and were trained based on the error associated with their corresponding predicted waveforms.

3.3.1 Determining Model Parameters

Experiments were first completed to identify some useful ESN parameter ranges for the ECG and ABP waveform data, especially reservoir sizes and learning rates. Identifying which parameters to set and the ranges of interest made the comparisons of different ESN architectures more appropriate. Two separate ESNs with similar architectures to the ESN mentioned above were used to predict ECG and ABP waveforms 100 time-steps into the future. These models were evaluated on MSE and maximum prediction error (max error). Different combinations of reservoir sizes and learning rates were tried because of their influence on how the waveforms were represented, decomposed, and learned by the system.

3.3.2 Methods for evaluation

The performance of these three architectures was assessed on MSE and the maximum prediction error (max error) between the performance and the actual waveform for both ECG and ABP data. A 10,000 time-step sample from patient record a41278 was used for this analysis. The data was divided into initializing (1-5,000), training (5,001-8,000), and testing segments (8,001-10,000). Initialization, training, and testing with all the networks followed the sample protocol. The Kolmogorov-Smirnov (KS) test was used to assess both the ECG and ABP MSE results for normality. This test is necessary to determine which statistical tests would be appropriate to use. Assuming normality, the ECG and ABP MSE and max error would be evaluated using one-way analysis of variance (ANOVA). This was used to determine if there were statistically significant differences between the ECG and ABP MSE and max error for the different architecture types (ESN1, ESN2, ESN3-900, ESN3-800, ESN3-700). Lastly, randomly sampled data from five individuals (a41325, a40416, a40076, a40432, and a41563) were used to evaluate the consistency and performance of the largest integrated reservoir network (ESN3-900).

4. Results

4.1 Model Parameters

To determine useful learning rates and reservoir sizes for the ESNs, a cursory assessment was first completed with the ECG data, varying the number of hidden nodes (100, 500, 750, and 1000) and the learning rates (0.01, 0.001, and 0.0001). The data was divided into initializing (1-5,000 time-steps), training (5,001-8,000 time-steps), and testing segments (8,001-10,000 time-steps). The data was tested 10 times for each combination. As shown in Table 1, reservoirs with nodes ranging between 100 and 750 and learning rates ranging from 0.001 and 0.0001 had on average better performance, prompting additional investigation as follows.

Table 1: ECG MSE test results (standard deviations)

MSE test	100	500	750	1000
0.01	3.78e42 (6.55e42)	7.83e7 (1.26e8)	4.1e13 (7.1e13)	4.6e76 (8.0e76)
0.001	0.026 (0.023)	18.39 (31.07)	1.42 (2.32)	1.6e6 (2.7e6)
0.0001	0.013 (0.0018)	0.02 (0.0029)	0.043 (0.031)	0.11 (0.14)

Reservoir size and learning rate ranges were further investigated with higher fidelity. The network ran ten more times with randomly initialized weights with new combinations of reservoir sizes (100, 200, 300, 400, 500, 600, and 700) and learning rates (0.001, 0.0005, and 0.0001). The results suggested that learning rates of 0.0001 and reservoir sizes of 500 or less than 300 tended to have better performance. A learning rate of 0.0001 and a reservoir size of 500 were used for the ECG components of the test ESN architectures. A reservoir size of 500 was chosen because it had slightly less variability in the results compared to sizes of 300 nodes or less. A similar analysis, conducted using ABP data, suggested a learning rate of 0.0001 and a reservoir size of 400 for the ABP components of the test ESN architectures. As a result, ESN1 and ESN2 both had two separate reservoirs with 500 and 400 nodes for the ECG and ABP waveforms respectively. Furthermore, we investigated the ESN3 architecture with three different reservoir sizes (700, 800, and 900 nodes) because a single reservoir system can perform differently based on its size.

4.2 Evaluating ESN Architecture Designs

The five ESN models were run thirty times with randomly initialized weights. Figures 3 and 4 show data from a training and testing run of ESN1 (independent reservoirs). The networks were evaluated on their prediction/test MSE and the maximum error values for both ECG and ABP waveforms. We found that the ESN architectures were able to predict ECG and ABP waveforms with varying levels of accuracy. We summarize the MSE results from both the ECG and ABP predictions in Figures 5 and 6 respectively. Outliers (results greater than three standard deviations from the mean) were most likely caused by poor randomly initialized weights and were removed, approximately five in each factor level, for the remainder of the analysis.

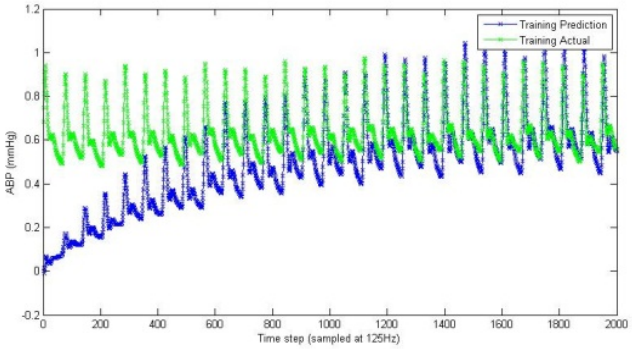


Figure 3: Sample ABP training run

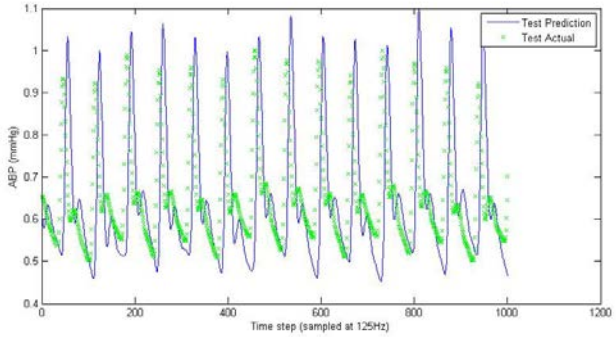


Figure 4: Sample ABP test prediction

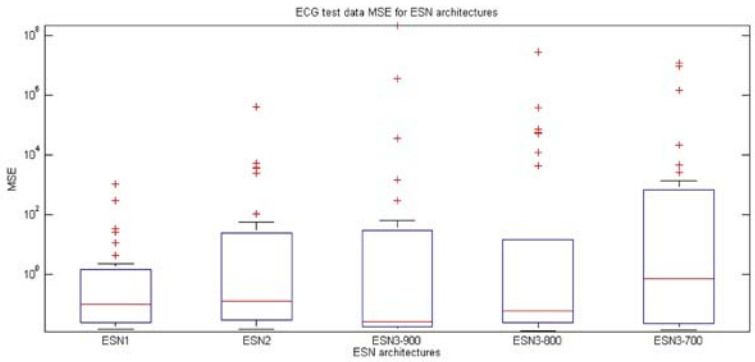


Figure 5: ECG MSE results for the different ESN architectures

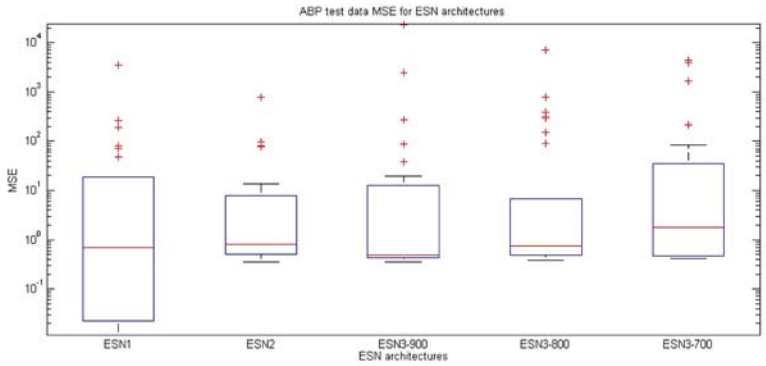


Figure 6: ABP MSE results for the different ESN architectures

Although the KS values for the ECG and ABP distributions were 0.5 and 0.53 respectively, parametric statistical tests could still be applicable considering the sample size. One-way ANOVAs were used to evaluate the MSE and max error of the predictions, Table 2. The factor levels were the different architectures (ESN1, ESN2, ESN3-900, ESN3-800, and ESN3-700).

Table 2: ANOVA results when evaluating ECG and ABP predictions using MSE and max error metrics

F-value (p-value)	ECG	ABP
MSE	1.48 (0.21)	1.19 (0.32)
Max Error	1.65 (0.17)	2.22 (0.071)

The results from the ABP max error analysis tended to be more meaningful (p-value = 0.071) compared to other metrics. The ABP max error results were mostly driven by the poor performance from ESN3-800 and ESN3-700 (integrated reservoir) as shown in Figure 7. We also noted that ESN3-900 (the largest integrated reservoir) tended to predict ECG waveforms slightly better than the independent and coupled reservoirs, Figure 7. However, ESN3's performance decreased with less hidden nodes. This may be due to the inability of the reservoir to correctly learn the two waveforms. In general, the independent reservoirs gave a much better prediction for the ABP waveform, although ESN3-900 had comparable performance in terms of ABP max error predictions. These results showed that there was no significant difference in the architectures at predicting two related waveforms together. There may also be increases in performance when combining waveforms in a single reservoir that is approximately similar in size to the ESN1 and ESN2. This showed the potential benefits of combining reservoirs to predict different but related waveforms.

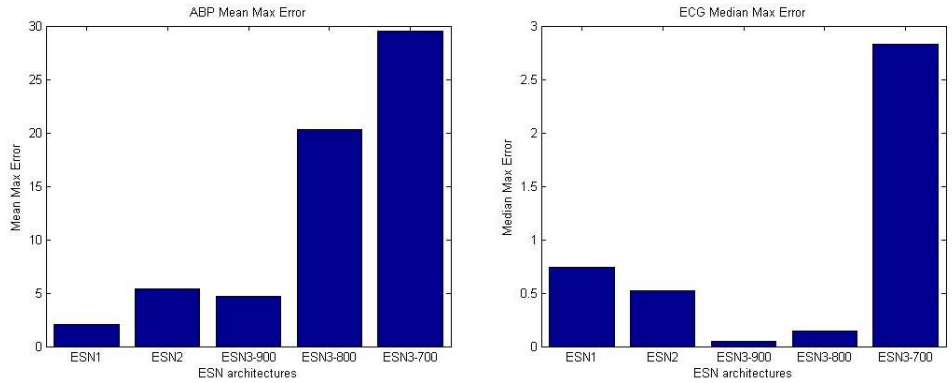


Figure 7: ABP mean max error (left) and ECG median max error (right).

4.3 Evaluating Across Individuals

Lastly, ESN3-900 (the largest integrated reservoir) was tested and compared using randomly sampled data from five individuals. ESN3-900 was chosen because its performance was similar to the basic independent reservoir architecture and it was faster to implement. The results showed that the ECG and ABP, Figure 8, performance across subjects were fairly similar, with the exception of patient 3. Though further work is needed, these results suggest that this architecture may be robust enough to be applied to different patients with little customization.

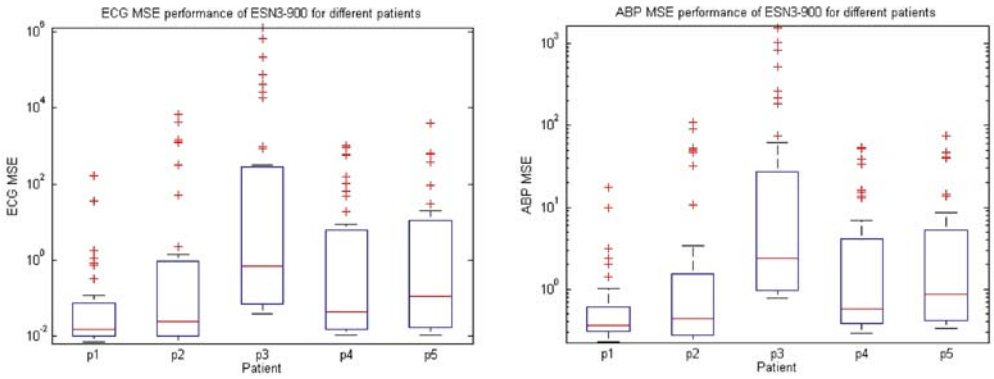


Figure 8: ECG MSE and ABP MSE results (left and right respectively) from five randomly selected patients

5. Discussion

The ability to predict and classify the ECG and ABP waveforms of patients is extremely important for the medical staff in Intensive Care Units. This paper approached this problem by demonstrating the effectiveness of different ESN architectures at predicting ECG and ABP waveforms. Although there was little significant difference in performance when predicting ECG and ABP waveforms between the different architectures, our results suggest that integrated reservoirs tended to have slightly better performance, especially when predicting the ECG waveform. Integrated reservoirs were also tested with different individuals with comparable results.

The ESN architectures in general tended to have much better performance at predicting ECG compared to ABP. This might be because the ECG waveforms have higher frequencies than the ABP waveforms. The loosely coupled subsystems in the reservoir might resonate or echo better at higher frequencies compared to lower frequencies. Furthermore, the ability of one larger reservoir to predict the waveforms was dramatically reduced as the size of the reservoir decreased. As the reservoir size decreased, the performance degradation was much more apparent with ABP then with ECG. This might also suggest that any coupling of ECG and ABP waveforms is biased toward the faster waveforms. This is an area that will require further investigation.

Furthermore, analysis of the largest integrated reservoir with five different individuals showed that this network architecture could be adaptable to different subjects. Without changing the parameters of the model, the largest integrated reservoir resulted in similar performance for four of the five randomly selected patients. Having a model that can be applied to different patients with little or no tailoring is attractive, and could be very beneficial in developing a predictive system for hospitals.

Although this research was exploratory, it does highlight some potential advantages of having one large reservoir for learning and predicting two related waveforms. The performance of the largest integrated reservoir was comparable to independent and coupled reservoirs, and was easier to implement. This analysis hints at interactions between how these waveforms are learned and stored in the reservoir. This could be investigated further and in more detail, perhaps starting with simpler, less chaotic waveforms. There are many questions to ask: for example, could initializing two reservoirs separately and then combining and reinitializing them lead to more robust internal subsystems in the reservoir? These and other queries can help further the understanding of Echo State Networks and make them more applicable for bioengineering and other applications.

There are limitations associated with this exploratory work, primarily concerning the limited predictive time-steps and the number of unique patients tested. This work aimed at exploring the application of ESNs for ECG and ABP waveforms. Although, these waveform predictions were limited to 100 time-steps, which may be too small to clinically trigger alarms, this work demonstrated the utility of this approach and could be refined in future work. It

would also be helpful to try different combinations of waveforms (not just the ECG II and ABP waveforms) and evaluate the models with more data from different patients. Furthermore, the parameters for each network (besides the learning rates and reservoir sizes) could be optimized. This would greatly increase the number of factors to control for, but a detailed factor level analysis of the network parameters would be very insightful. Expanding these models to predict higher level alarm states would be helpful. In addition, data transformations of the results might provide more normalized data. However, more research is needed to understand what these transformations mean intuitively for the results before applied.

A model that can predict ECG and ABP waveforms can naturally be extended to classify other waveforms and predict alarms. A model that can forecast the accuracy of ECG alarms or classify false alarms based on predicted waveforms might be extremely beneficial to medical staff, especially those in the ICU. There are several types of alarms in the ICU, each with unique waveform patterns, and this work can be extended to investigate the differences between predicted waveforms during a false alarm and a true alarm. Models that can extrapolate what a patient's waveforms will be like even a few seconds after an alarm can help medical staff and hospital systems better understand and classify alarms, with the ultimate goal of reducing false alarms, alarm fatigue, and improving patient care.

6. Conclusion

Alarm fatigue caused by bedside machines is a serious issue in Intensive Care Units. Part of this problem is the inability of these machines to accurately predict and classify a patient's ECG and ABP waveforms. In this study, we explored and demonstrated the ability of different Echo State Network architectures for predicting ECG and ABP waveforms with varying levels of accuracy. The most accurate predictions were generally by the largest integrated reservoir and the independent reservoirs architectures, which often had comparable results. Furthermore, results showed potential benefits for applying large integrated ESN reservoirs for different individuals. This paper also discussed limitations of this research, as well as suggestions for future work, and in particular the investigation of one versus two reservoir interactions and applications for predictive alarm classifications.

References

1. Mathews SC, Pronovost PJ. The need for systems integration in health care. Journal of the American Medical Association. 2011; 305(9): 934-5.
2. Graham KC, Cvach M. Monitor alarm fatigue: standardizing use of physiological monitors and decreasing nuisance alarms. American Journal of Critical Care. 2010; 19(1): 28-34.
3. Tsien CL, Fackler JC. Poor prognosis for existing monitors in the intensive care unit. Critical care medicine. 1997; 25(4): 614-619.
4. Cvach M. Monitor alarm fatigue: an integrative review. Biomedical Instrumentation & Technology. 2012; 46(4): 268-77.
5. Edelson, M. Safety First. Hopkins Medicine. 2013; 24-31.
6. Fuchs L, Chronaki CE, Park S, Novack V, Baumfeld Y, Scott D, et al. ICU admission characteristics and mortality rates among elderly and very elderly patients. Intensive care medicine. 2012; 38(10): 1654-61.
7. Konkani A, Oakley B, Bauld TJ. Reducing hospital noise: a review of medical device alarm management. Biomedical Instrumentation & Technology. 2012; 46(6): 478-87.
8. Jaeger H. Reservoir riddles: Suggestions for echo state network research. Proceedings of the IEEE International Joint Conference of Neural Networks; 2005; 3: 1460-2.
9. Jaeger H, Harald H. Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. Science. 2004; 304(5667): 78-80.
10. Tong MH, Bickett AD, Christiansen EM, Cottrell GW. Learning grammatical structure with echo state networks. Neural Networks. 2007; 20(3): 424-32.
11. Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman LW, Moody G, et al. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): a public-access intensive care unit database. Critical care medicine. 2011; 39(5): 952.
12. Celi LA, Galvin S, Davidzon G, Lee J, Scott DJ, Mark RG. A Database-driven decision support system: customized mortality prediction. Journal of personalized medicine. 2012; 2(4): 138-48.
13. Celi LA, Mark RG, Lee J, Scott DJ, Panch T. Collective experience: a database-fuelled, inter-disciplinary team-led learning system. Journal of computing science and engineering. 2012; 6(1): 51.
14. Scott DJ, Lee J, Silva I, Park S, Moody GB, Celi LA, Mark RG. Accessing the public MIMIC-II intensive care relational database for clinical research. BMC Medical Informatics and Decision Making. 2013; 13(1): 9.

15. Aboukhalil A, Nielsen L, Saeed M, Mark RG, Clifford, GD. Reducing false alarm rates for critical arrhythmias using the arterial blood pressure waveform. Journal of biomedical informatics. 2008; 41(3): 442-51.
16. Fetics B, Nevo E, Chen CH, Kass DA. Parametric model derivation of transfer function for noninvasive estimation of aortic pressure by radial tonometry. Biomedical Engineering. 1999; 46(6): 698-706.
17. Keogh E, Lin J, Fu A. Hot sax: Efficiently finding the most unusual time series subsequence. Proceedings of the IEEE International Joint Conference on Data Mining; 2005;8.
18. Keogh E, Lin J, Lee SH, Van Herle H. Finding the most unusual time series subsequence: algorithms and applications. Knowledge and Information Systems. 2007; 11(1): 1-27.
19. Lonardi S, Lin J, Keogh E. Efficient discovery of unusual patterns in time series. New Generation Computing. 2006; 25(1): 61-93.
20. Zong W, Moody GB, Mark RG. Reduction of false arterial blood pressure alarms using signal quality assessment and relationships between the electrocardiogram and arterial blood pressure. Medical and Biological Engineering and Computing. 2004; 42(5): 698-706.
21. Saria S, Duchi A, Koller D. Discovering deformable motifs in continuous time series data. International Joint Conference on Artificial Intelligence. 2011; 22(1): 1465.
22. Williams C, Quinn J, McIntosh N. Factorial switching Kalman filters for condition monitoring in neonatal intensive care. Neural Information Processing. 2005; 18: 1513-1520.
23. Gather U, Imhoff M, Fried R. Graphical models for multivariate time series from intensive care monitoring. Statistics in medicine. 2002; 21(18): 2685-2701.
24. McSharry PE, Clifford GD, Tarassenko L, Smith L. A dynamical model for generating synthetic electrocardiogram signals. IEEE Transactions on Biomedical Engineering. 2003; 50(3): 289-294.
25. Kaastra I, Boyd M. Designing a neural network for forecasting financial and economic time series. Neurocomputing. 1996; 10(3): 215-36.
26. Maguire LP, Roche B, McGinnity TM, McDaid LJ. Predicting a chaotic time series using a fuzzy neural network. Information Sciences. 1998; 112(1): 125-36.
27. Zhang GP, Qi M. Neural network forecasting for seasonal and trend time series. European Journal of Operational Research. 2005; 160(2): 501-14.
28. MIMIC II [Internet]. MIT (MA): Multiparameter Intelligent Monitoring in Intensive Care [cited 2013 March 1]. Available from: http://physionet.org/mimic2/
29. WFDB Software Package [Internet]. MIT (MA): PhysioNet WFDB Software Package [cited 2013 March 1]. Available from: http://physionet.org/physiotools/wfdb.shtml

Evaluation of RxNorm for Medication Clinical Decision Support

Robert R. Freimuth, PhD[1,2], Kelly Wix[3], PharmD, RPh, Qian Zhu, PhD[1], Mark Siska, RPh[3], Christopher G. Chute, MD, DrPH[1]
[1]Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, [2]Office of Information and Knowledge Management, [3]Pharmacy Services Information Systems; Mayo Clinic, Rochester, MN

Abstract

We evaluated the potential use of RxNorm to provide standardized representations of generic drug name and route of administration to facilitate management of drug lists for clinical decision support (CDS) rules. We found a clear representation of generic drug name but not route of administration. We identified several issues related to data quality, including erroneous or missing defined relationships, and the use of different concept hierarchies to represent the same drug. More importantly, we found extensive semantic precoordination of orthogonal concepts related to route and dose form, which would complicate the use of RxNorm for drug-based CDS. This study demonstrated that while RxNorm is a valuable resource for the standardization of medications used in clinical practice, additional work is required to enhance the terminology so that it can support expanded use cases, such as managing drug lists for CDS.

Introduction

Clinical decision support systems (CDSS) have emerged as an essential component of clinical information systems and electronic medical records (EMRs). Not only do CDSS serve a variety of functions to assist clinicians provide better care and prevent adverse events, but also they are required to meet regulatory requirements including the Meaningful Use measures of the American Recovery and Reinvestment Act (1). A number of health care organizations have described positive outcomes associated with deploying CDSSs and have leveraged their capabilities as a strategic tool for attaining institutional quality and safety improvement goals and objectives (2-4). Despite the growing advantages associated with CDSS, however, significant challenges impede its widespread adoption and the development of even more sophisticated computerized alerts and types of CDSSs (5). Sittig et al. identified ten grand challenges associated with CDSSs, which included improving the user interface to streamline clinical workflow and disseminating best practices for CDS design, development, and implementation (6). Fine-tuning CDS rules to deliver the most useful information at the appropriate time without causing alert fatigue remains a significant challenge, particularly when medications are involved (7).

One of the primary challenges related to the implementation of a robust, scalable CDSS for medications is the development of a comprehensive knowledge base that effectively accommodates maintenance and interoperability, supports rigorous data quality management principles, and enables the development of "free form" rules that are EMR-agnostic and accessible to ancillary medication management and supporting systems (8). In an effort to more effectively manage our medication-related CDS rules, including those related to pharmacogenomics, across the diverse systems within the Mayo Clinic enterprise we sought to adopt a standard, vendor-neutral drug terminology that would support interoperability and allow us to manage a single drug list for each CDS rule that is implemented. This manuscript describes our evaluation of RxNorm to better understand its capabilities and limitations in the context of a prototypic use case: the management of the drug list for an existing, active CDS rule.

Motivating Use Case: CDS Rule for Deep Vein Thrombosis (DVT) Prophylaxis

The following scenario illustrates a motivating use case for this study. To minimize the maintenance of drug-based CDS rules as formularies change, medications are referenced by generic drug name. To minimize alert fatigue and improve the specificity of the rules, the route of administration is used as an additional criterion for the rule trigger.

A CDS rule was developed to ensure that hospitalized inpatients receive deep vein thrombosis (DVT) prophylaxis. Providers receive a pop-up alert in the ordering system if no DVT prophylaxis was ordered.

The rule scans the active order list to determine if an order for heparin exists, using an RxNorm identifier (RxCUI) for heparin injectable solution. Testing reveals the rule is not identifying some orders for heparin subcutaneous injection. After detailed review of the data, it is determined that the RxCUI used for heparin injectable solution is not used for heparin in a prefilled syringe.

In this example, a logical assumption was made that all products that were injectable solutions for a given drug would have that corresponding defined relationship. In reality, drugs that were available in prefilled syringes were assigned a different relationship than those that were not in prefilled syringes. Regardless of the reason for this difference, this example emphasizes the necessity of knowing the intricacies and limitations of the data source, especially when used for clinical applications.

Background

RxNorm is a standard terminology developed by the National Library of Medicine (NLM) that provides normalized names for clinical drugs (9). It is intended to be used to facilitate the exchange of medication-related information among clinical systems and it is part of the federal Meaningful Use standard. RxTerms is a drug interface terminology that is derived from RxNorm, intended to facilitate medication order entry. It contains a pruned set of drugs from RxNorm that are anticipated to be most useful in a prescribing environment (10).

The Veterans Affairs National Drug File Reference Terminology (NDF-RT) is developed by the Department of Veterans Affairs (VA) Veterans Health Administration (11). The NDF-RT contains information about drug characteristics, including ingredient(s) and dose form. Concepts in the NDF-RT are organized into taxonomies, which represent generalization relationships between concepts hierarchically. As of June 2010, the NDF-RT has been integrated into RxNorm.

The RxNorm and NDF-RT terminologies have been used extensively for the normalization of drug data (12-15). Many of these efforts focused on developing methods for mapping terms from local drug coding systems to the reference terminologies and assessing the overall coverage of content. In contrast, in this study we evaluated the ability of RxNorm, with associated data from relationships to NDF-RT and RxTerms, to provide standardized representations of drug name and route of administration to facilitate management of drug lists for CDS rules.

Materials and Methods

Data Sources

The RxNorm full monthly release dated February 3, 2014 was downloaded from the NLM Unified Medical Language System (UMLS) web site (16). The RxTerms data files, version 201401, were also downloaded (17). The RxNorm and RxTerms data were loaded into a local MySQL database using a custom version of the loader scripts that were provided. The "rxnconso" and "rxnrel" tables were cloned and populated with subsets of the original data from the respective tables to facilitate complex queries that required multiple joins across these tables.

Terms from the NDF-RT "Dose Forms" (NUI N0000010010) hierarchy that had the NDFRT_KIND property of "DOSE_FORM_KIND" were downloaded via the NLM NDF-RT REST API using http://rxnav.nlm.nih.gov/REST/Ndfrt/allconcepts?kind=DOSE_FORM_KIND. The NDF-RT unique identifier (NUI) was extracted for each term. Since the term "Dose Forms" contains two child hierarchies, "Drug Delivery Device" (NUI N0000177905) and "Orderable Drug Form" (NUI N0000135762), the BioPortal web interface (18, 19) was used to browse the NDF-RT terminology and determine the branch to which each dose form term belonged. The concept terms, codes, and hierarchical branch name were loaded into a custom table in the local database.

Identification of Clinical Drugs and Related Attributes

To evaluate RxNorm for our motivating use case, we needed to identify the term types (TTY) that corresponded to generic drug name and route of administration. The RxNorm technical documentation was used to determine which entities and relationships were most pertinent to this study (20, 21). We chose the Semantic Clinical Drug (SCD) to represent orderable drugs using generic drug name(s). The SCD also contains strength and dose form. The local RxNorm database was queried for all SCDs (TTY = "SCD") that did not have a "suppress" attribute of "O", "Y", or "E". This list of drugs was used as the starting point for evaluation.

By browsing the RxNorm data using the RxNav user interface (22), several attributes were found to contain information related to route of administration: "Dose Form" and "Dose Form Group" from RxNorm; "rxn_dose_form", "new_dose_form", and "route" from RxTerms; and "Dose Form" from NDF-RT (**Figure 1**). The NLM was contacted to clarify the relationship between these attributes. RxNorm Dose Form was originally based on HL7 vocabulary group dose forms. RxNorm Dose Form was equivalent to RxTerms rxn_dose_form, which was used to derive RxTerms route and new_dose_form. RxTerms route was used as the basis for RxNorm Dose Form Group. NDF-RT Dose Form was "initialized from RxNorm" but equivalency was not determined. Based on this

information, the attributes RxTerms rxn_dose_form and RxNorm Dose Form Group were omitted from the analysis but the other attributes were retained and used for comparison.

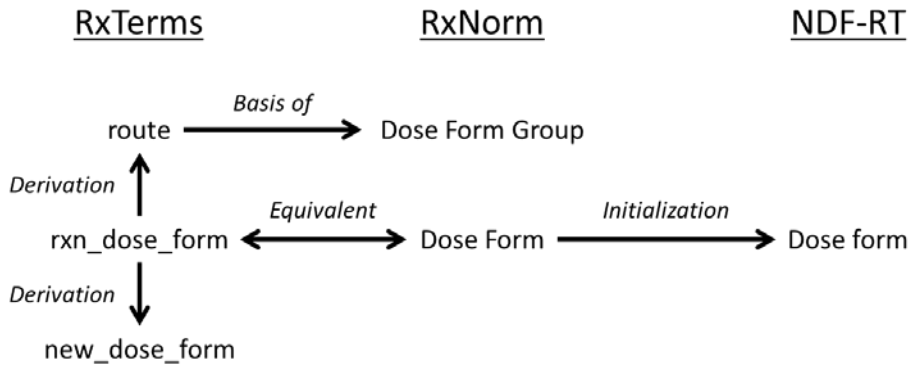


Figure 1: Attributes containing route of administration. The figure shows the attributes from each of the three data sources that were found to contain information related to route of administration, as well as the relationship between those terms.

Extraction and Evaluation of Defined Relationships

Instances of RxNorm Dose Form (TTY = "DF", source abbreviation (SAB) = "RXNORM") were obtained for each SCD using the RxCUI of the SCD and the relationship (RELA) "has_dose_form" (**Figure 2**). Terms for RxTerms route and new_dose_form were extracted directly from the RxTerms data using the RxCUI of the SCD. Terms for NDF-RT Dose Form (TTY = "PT", SAB = "NDFRT") were obtained for each SCD using the RxCUI of the SCD and the relationship (RELA) "dose_form_of". The resulting data set from each source was checked for missing or multiple terms for each SCD. All related terms were stored in new tables in the database to simplify subsequent analysis and exported to a spreadsheet for manual review of inter-source semantic consistency.

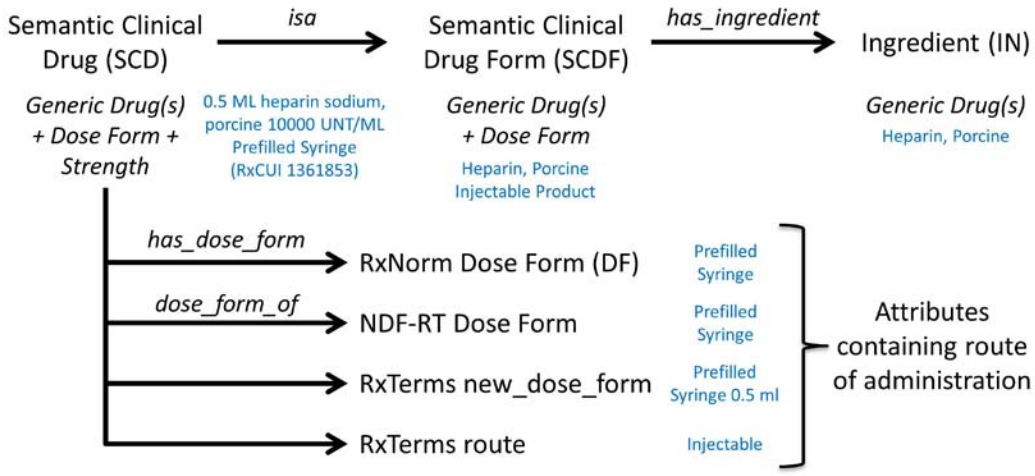


Figure 2: Drug entities, attributes, and relationships. The figure shows the entities, attributes, and relationships that are relevant to this study. Terms corresponding to the heparin use case are shown in blue text, as examples.

The Semantic Clinical Drug Form (SCDF) was used to compare dose forms for a given set of ingredients. The SCDFs (TTY = "SCDF"; "suppress" attribute not "O", "Y", or "E") associated to each SCD were obtained using the RxCUI of the SCD and the relationship (RELA) "isa". Ingredients (TTY = "IN", SAB = "RXNORM") for each

SCDF were obtained using the relationship (RELA) "has_ingredient". An ordered, concatenated form of ingredients was created for each SCD. Similarly, an ordered, concatenated list of dose forms was generated for each concatenated ingredient.

Results

Identification of Clinical Drugs and Related Attributes

The SCD contains generic drug name(s), strength, and dose form, and as such it serves as a convenient entity in RxNorm to which orderable drugs can be mapped. We found 20,268 SCDs in RxNorm, which were used as the starting point for the subsequent evaluation.

The authoring and maintenance of drug lists for CDS rules is most efficiently performed using the ingredient generic name and route of administration; therefore, it was necessary to obtain these attributes for each SCD. Generic ingredients were clearly represented within the RxNorm data model (TTY = "IN") but information about the route of administration was found in several entities (**Figure 1**). Two of these were eliminated from further consideration. Specifically, RxTerms rxn_dose_form was found to be redundant with RxNorm Dose Form, and was excluded from further analysis. Similarly, RxNorm Dose Form Group was defined as "a term type that serves as a grouping of dose forms (TTY=DF) related by route of administration (i.e., Topical) or dose form (i.e., Pill)" (23) and was also excluded since it was a less primitive concept than its source concepts, which were already included in the analysis.

Extraction and Evaluation of Defined Relationships

The RxNorm Dose Form was identified for each SCD in the data set. A total of 20,268 Dose Form terms were found, one for each SCD. No SCD had more than one RxNorm Dose Form and no SCDs were missing a related Dose Form. There were 104 different terms represented; the most frequently used terms are shown in **Table 1**.

RxCUI	Dose Form	Frequency
317541	Oral Tablet	4223 (20.8%)
316949	Injectable Solution	4215 (20.8%)
316965	Oral Capsule	1950 (9.6%)
316968	Oral Solution	1439 (7.1%)
316945	Extended Release Tablet	723 (3.6%)
721656	Prefilled Syringe	594 (2.9%)
316982	Topical Cream	589 (2.9%)
316969	Oral Suspension	512 (2.5%)
316986	Topical Solution	505 (2.5%)

Table 1: RxNorm and NDF-RT Dose Form. The most frequently used terms are listed for RxNorm and NDF-RT Dose Form.

The terms for RxTerms route and new_dose_form were extracted for each SCD. Only 15,077 SCDs had related terms from RxTerms (exactly one route and new_dose_form each); 5191 SCDs (25.6% of the SCDs in the data set) were missing corresponding terms from RxTerms. The most frequently used terms for route and new_dose_form are shown in **Table 2**; a total of 39 and 170 distinct terms were found, respectively.

route	Frequency
Oral Pill	6167 (40.9%)
Injectable	3754 (24.9%)
Topical	1584 (10.5%)
Oral Liquid	1510 (10.0%)
Chewable	305 (2.0%)

new_dose_form	Frequency
Sol	4820 (32.0%)
Tab	3954 (26.2%)
Cap	1667 (11.1%)
Susp	564 (3.7%)
Cream	396 (2.6%)

Table 2: Terms from RxTerms. The most frequently used terms are listed for RxTerms route and new_dose_form.

NDF-RT Dose Form terms were extracted for each SCD. Although we expected at most 20,268 results, 20,417 related terms were found. Further investigation revealed that 55 SCDs were related to two NDF-RT Dose Form terms and 47 SCDs were related to three terms. The 251 terms related to these 102 SCDs were instances of only 7 different terms that represented three distinct concepts: mouthwash, toothpaste, and topical cake (**Table 3**). Specifically, 47 SCDs were related to the NDF-RT orderable drug form "mouthwash", but they were also related to similar terms from two other NDF-RT hierarchies (pharmaceutical preparations and chemical ingredients). Similarly, 54 SCDs were related to the orderable drug form "toothpaste" as well as the chemical ingredient "toothpastes". A single SCD contained relationships to two terms from the orderable drug from hierarchy: the concept "cake" and its child concept "topical cake".

NUI	Concept	Location in NDF-RT Hierarchy
N0000029230	MOUTHWASHES	Pharmaceutical Preparations => Drug Products by VA Class => Dental and Oral Agents, Topical
N0000135733	Mouthwash	Orderable Drug Form => Liquid => Solution => Oral Solution
N0000011404	Mouthwashes	Chemical Ingredients => Biomedical and Dental Materials
N0000135791	Toothpaste	Orderable Drug Form => Solid => Paste
N0000171562	Toothpastes	Chemical Ingredients => Biomedical and Dental Materials => Dentifrices
N0000135686	Cake	Orderable Drug Form => Solid
N0000184140	Topical Cake	Orderable Drug Form => Solid => Cake

Table 3: Terms for NDF-RT Dose Form. The table lists groups of terms for NDF-RT Dose Form that were related to a single SCD.

All SCDs were related to a single term for Dose Form that was in the orderable drug form hierarchy, except for one SCD that contained relationships to both "cake" and "topical cake". When the extraneous terms from the pharmaceutical preparations and chemical ingredients hierarchies, and the parent concept "cake", were excluded from the query 20,268 results were obtained. Each SCD was found to contain exactly one term for NDF-RT Dose Form. There were 104 different terms represented; the terms and their frequencies were identical to those for RxNorm Dose Form (**Table 1**).

Since each SCD was related to at most one term of each type (RxNorm Dose Form, RxTerms route and new_dose_form, NDF-RT Dose Form), the results of the individual searches were merged into a single table that was indexed by SCD. The semantic consistency of the terms from each data source was reviewed by examining all 296 unique combinations of terms. Only 88 combinations were found to be consistent, which was defined to allow for some semantic precoordination of concepts. In 125 cases the terms from RxTerms were narrower than those from RxNorm and NDF-RT due to the explicit inclusion of route, count, and/or quantity information (**Table 4**). The remaining 83 combinations, corresponding to 5191 SCDs, were missing terms from RxTerms. In all cases, the terms for RxNorm Dose Form and NDF-RT Dose Form were identical.

RxNorm Dose Form	NDF-RT Dose Form	RxTerms route	RxTerms new_dose_form
Dry Powder Inhaler	Dry Powder Inhaler	Inhalant	DPI 60 puff
Extended Release Capsule	Extended Release Capsule	Oral Pill	24 HR XR Cap
Augmented Topical Lotion	Augmented Topical Lotion		
Augmented Topical Lotion	Augmented Topical Lotion	Topical	Lotion (Augmented)
Buccal Tablet	Buccal Tablet		
Buccal Tablet	Buccal Tablet	Buccal	Tab
Medicated Shampoo	Medicated Shampoo		
Medicated Shampoo	Medicated Shampoo	Shampoo	Medicated Shampoo

Table 4: Examples of semantic consistency among terms. The first two rows illustrate the inclusion of qualifiers (e.g., "60 puff", "24 HR") in the terms from RxTerms that are not present in those from RxNorm or NDF-RT. The remaining rows show pairs of term combinations that contain identical values for RxNorm and NDF-RT but differ overall because one of the entries is missing terms from RxTerms.

Of the 83 combinations that lacked terms from RxTerms, 76 were found to have an equivalent entry based on RxNorm and NDF-RT values that also contained terms from RxTerms (**Table 4**). These 76 instances corresponded to 5065 SCDs, which accounted for 97.6% of SCDs that were missing terms from RxTerms and 25.0% of all SCDs in this data set. For the remaining 7 combinations, which corresponded to 126 SCDs, a corresponding entry either was not found or could not be assigned due to ambiguity as a result of the semantic precoordination of route, count, or quantity (**Table 5**).

# SCDs	RxNorm Dose Form	NDF-RT Dose Form	RxTerms route	RxTerms new_dose_form
1	Crystals	Crystals		
2	Metered Dose Inhaler	Metered Dose Inhaler		
1	Ophthalmic Cream	Ophthalmic Cream		
1	Otic Ointment	Otic Ointment		
13	Prefilled Applicator	Prefilled Applicator		
107	Prefilled Syringe	Prefilled Syringe		
1	Rectal Solution	Rectal Solution		

Table 5: Missing terms from RxTerms. Term combinations that were missing terms from RxTerms. These combinations did not have an unambiguous corresponding entry that could be used to obtain these terms.

To evaluate the consistency of related terms at the drug level, generic ingredients (IN) were obtained for each SCD. Since there was no direct relationship between SCD and IN, the SCDF was used as an intermediate step. The SCDF, which contain generic drug name(s) and dose form but not strength (**Figure 2**), was obtained for each SCD. Each SCD had exactly one SCDF; 8717 distinct SCDFs were retrieved. The majority of SCDFs (4763, 54.6%) were utilized by a single SCD, which indicated that only one strength was available for the drug in the specified dose form (**Figure 3**). Another 1705 (19.6%) SCDFs were used by two SCDs.

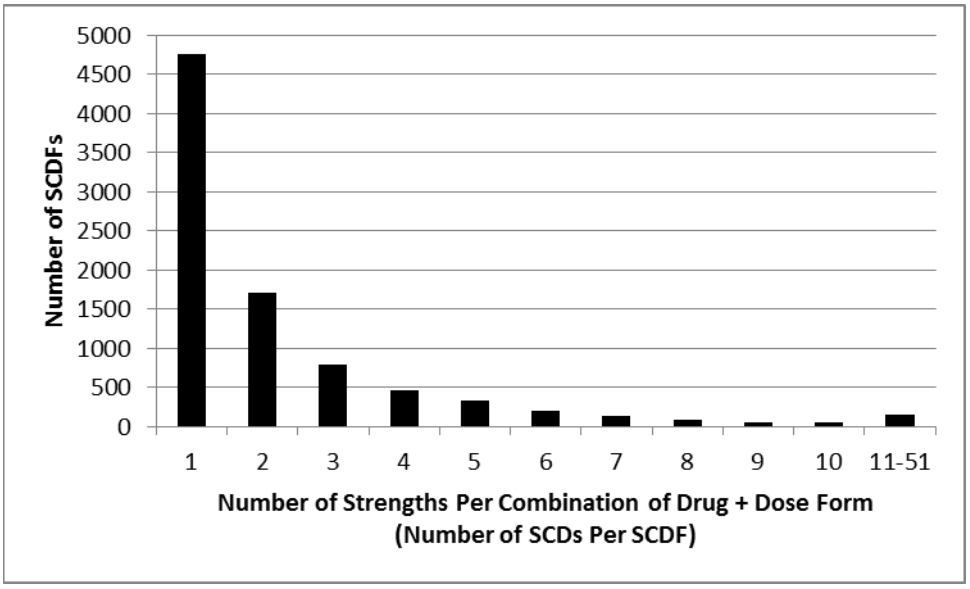


Figure 3: Relationship between SCDs and SCDFs. Most combinations of drug + dose form were mapped to a single SCDF but a small proportion of SCDFs represented many SCDs, indicating a large number of strengths available for the given combination of drug and dose form.

On the other end of the spectrum, 158 (1.8%) SCDFs were used by more than 10 SCDs each. The top 25 SCDFs (0.3%) were utilized by more than 20 SCDs each, representing 698 SCDs in total. The three most highly utilized SCDFs were "Menthol Lozenge" (RxCUI 374544), "Amylases / Endopeptidases / Lipase Enteric Coated Capsule" (RxCUI 402754), and "Guaifenesin / Phenylephrine Extended Release Tablet" (RxCUI 373391). The "menthol lozenge" SCDF was used by 51 SCDs that had strengths ranging from 1 mg to 18 mg. The "amylases/endopeptidases/lipase enteric coated capsule" SCDF was used by 45 SCDs that had varying combinations of strengths of the three ingredients. The "guaifenesin/phenylephrine extended release tablet" SCDF was used by 44 SCDs, distinguished not only by varying combinations of the two ingredients but also by different uses of the qualifier "12 HR". In fact, 24 of the 44 SCDs in this SCDF were identical pairs of drugs with the same strengths and dose form, but one specified "12 HR" and the other did not.

An ordered, concatenated list of ingredients was developed for each SCD. This list contained 8717 unique combinations of concatenated ingredients and dose form, which was expected based on the number of SCDFs. The list was grouped by ingredient(s) and the NDF-RT dose forms within each group were identified. Examples of the results are shown in **Table 6**. A total of 5021 unique combinations of ingredients were found.

Ingredient(s)	NDF-RT Dose Form(s)	NDF-RT Dose Form Type(s)
1-octacosanol	Oral Capsule	Orderable Drug Form
4-Aminobenzoate	Oral Capsule + Oral Tablet	Orderable Drug Form
4-Aminobenzoic Acid + Arginine	Topical Cream	Orderable Drug Form
Acetaminophen + Aspirin	Oral Powder + Oral Tablet	Orderable Drug Form
aclidinium	Dry Powder Inhaler	Drug Delivery Device
adalimumab	Injectable Solution + Prefilled Syringe	Drug Delivery Device + Orderable Drug Form

Table 6: NDF-RT Dose Forms, by Drug. The table lists examples of ingredient combinations and their respective related terms from NDF-RT Dose Form.

The type ("Drug Delivery Device" and/or "Orderable Drug Form") of dose form(s) available for each combination of ingredients was determined based on the NDF-RT hierarchy. The majority of ingredients (4702, 93.6% of the 5021 combinations) were related to one or more concepts from Orderable Drug Form only; these represented 15,776 SCDs. There were 47 (0.9%) ingredients representing 81 SCDs that were related to one or more concepts from Drug Delivery Device only. Interestingly, 272 (5.4%) ingredients were related to one or more concepts from both hierarchies, indicating inconsistency in how terms for dose form were related to a given drug. These cases represented 4411 SCDs, or 21.8% of the SCDs in this data set. Examples of drugs that were related to concepts from both hierarchies are shown in **Table 7**.

Ingredient(s)	NDF-RT Orderable Drug Form(s)	NDF-RT Drug Delivery Device(s)
ciclesonide	Inhalant Solution	Metered Dose Inhaler + Nasal Inhaler
Cromolyn	Inhalant Powder + Inhalant Solution + Nasal Solution + Nasal Spray + Ophthalmic Solution + Oral Capsule + Oral Solution	Metered Dose Inhaler + Nasal Inhaler
fluticasone	Inhalant Powder + Inhalant Solution + Topical Cream + Topical Lotion + Topical Ointment	Dry Powder Inhaler + Metered Dose Inhaler + Nasal Inhaler
heparin, porcine	Injectable Solution	Prefilled Syringe
Leuprolide	Injectable Solution + Injectable Suspension	Drug Implant + Prefilled Syringe

Table 7: NDF-RT Dose Forms, by Type and Drug. The table lists examples of drugs that were related to concepts from both NDF-RT Dose Form hierarchies.

Discussion

In this study we evaluated the potential ability of RxNorm to facilitate management of drug lists for CDS rules by providing standardized representations of generic drug name and route of administration. Generic drug name was clearly represented as ingredients. Anecdotal evidence suggested several different term types might contain information related to the route of administration. This study found route data in values of three attributes: RxNorm Dose Form, RxTerms route, and NDF-RT Dose Form. None of the value sets used by those attributes contained only route of administration, however, which presented a challenge when determining which to use to facilitate management of CDS rules.

The documentation for NDF-RT Dose Form indicates that the attribute is "initialized from RxNorm and is periodically resynchronized by computer algorithm" (11), which implies that some divergence may be anticipated over time. Nonetheless, we found this attribute to be identical to RxNorm Dose Form, with the caveat that some SCDs contained relationships to multiple terms from NDF-RT Dose Form. All but one of those related terms were outside the Dose Form hierarchy with similar spellings as the corresponding Dose Form term. This could suggest that those terms may have been introduced in error, a risk we recognized previously (24), and subsequently escaped the resynchronization process. The NUI codes for these terms are included in **Table 3** so they can be manually excluded from searches, if desired.

We found the content of the four term types was consistent between sources, although there was some variability in the semantic precoordination of concepts used by each attribute. In many cases the RxTerms new_dose_form attribute contained additional qualifiers that were not present in the other attributes, such as those which specified count (e.g., number of puffs for an inhaler), time-release (e.g., 12 HR), or volume (e.g., 1 ml). For example, the RxNorm/NDF-RT Dose Form "Prefilled Syringe" could not be mapped unambiguously to RxTerms new_dose_form, as the latter included a qualifier to indicate the volume of the syringe. In this case, there were 58 different values in RxTerms new_dose_form that map to "Prefilled Syringe", from "Prefilled Syringe 0.09 ml" to "Prefilled Syringe 125 ml". The inclusion of a volume metric within the term for dose form caused a significant expansion of the value set, which could complicate the management and use of these terms for CDS rules.

More than one quarter of the SCDs in the data set were missing relationships to RxTerms. The vast majority of them could be inferred, however, by using the RxTerms terms that were related to other SCDs that have the same value for RxNorm/NDF-RT Dose Form. Relationships for the remaining instances could not be assigned using this method largely due to ambiguity caused by count or volume qualifiers. In particular, most of those SCDs had a Dose Form of "prefilled syringe", which includes a specified volume in RxTerms. Without access to the algorithms used to assign RxTerms relationships, we could not ascertain the reason why so many relationships were missing from this data set. This may be partly due to the fact that RxTerms intentionally omits some drugs to improve query performance (10). Regardless, any query that uses terms from RxTerms as a parameter is likely to return incomplete results, which complicates or even prevents the use of RxTerms for the management of drug lists for CDS rules. We believe it would be better to define relationships for all drugs and allow implementers to filter the data set.

Our initial observations regarding injectable heparin suggested that the same drug could be classified as both an Orderable Drug Form and a Drug Delivery Device, based on the value of Dose Form. This study confirmed that observation and demonstrated that while this phenomenon affected a relatively small proportion of ingredients, those ingredients comprised nearly 22% of the SCDs in this data set. We were not able to find any documentation that describes the process by which relationships are assigned. Without a clear understanding of this process it would be difficult to confidently use those relationships for the management of drug lists used in CDS rules.

Finally, we observed extensive semantic precoordination of orthogonal concepts within the value sets used for each of the attributes in this study (RxNorm/NDF-RT Dose Form, RxTerms route, RxTerms new_dose_form). **Table 1** and **Table 2** illustrate terms that contain concepts for dose form, route, and/or drug delivery device, such as "oral tablet" and "topical cream". This precoordination complicates the computational use of these terms for managing drug lists and makes it difficult to use those terms appropriately.

Although RxNorm is intended to "assist with medication-related clinical decision support" (9), it was primarily developed as a standard for data exchange and semantic interoperability so it is not surprising that the data model may be more suited to this use case. In particular, the current structure and content for representing route of administration and dose form is not optimal for CDS. This is not unexpected, however, as the content and organization of biomedical ontologies may enable them to meet some use cases better than others (15, 25). As

RxNorm is adopted more widely it is likely that additional use cases will arise that will require extensions to the terminology.

Limitations

The limitations of this study include restrictions due to its scope. Specifically, we limited our evaluation to attributes that pertained to drug name and route of administration, and omitted analysis of RxNorm Dose Form Group (DFG) and Semantic Clinical Drug Form Group (SCDG). While it is possible these elements may be relevant to our use case, they are computational derivations of the attributes that were already included in the study. Similarly, we explored the hierarchy of NDF-RT Dose Form while evaluating semantic consistency but we did not investigate other classification schemes from that terminology. In particular, drug class would be relevant to drug-based CDS; this topic has been explored previously (24, 26, 27).

Finally, we intentionally scoped this evaluation to RxNorm, as it is a nationally-recognized standard that meets Meaningful Use requirements. Other terminologies with cross-references to RxNorm content might address some of the issues that were identified in this study. For example, the Drug Ontology (DrOn) is derived in part from RxNorm and follows formal ontological principles (28). It may be possible to extend RxNorm with data from other ontologies to meet the needs of our use case.

Conclusions

In this study we evaluated the potential ability of RxNorm to facilitate management of drug lists for CDS rules by providing standardized representations of generic drug name and route of administration. Generic drug names were clearly represented as ingredients. It was more difficult to find a robust representation for route of administration, however, and we explored several attributes to determine which might be most appropriate for our use case. None of those attributes provided an ideal, semantically "pure" form of route of administration.

In the course of these investigations we discovered several issues related to data quality, including erroneous or missing defined relationships, and the use of different concept hierarchies to represent the same drug. We also identified examples where the use of qualifier terms and the semantic precoordination of orthogonal concepts would complicate the use of the data.

This study demonstrates that while RxNorm is a valuable resource for the standardization of medications used in clinical practice, additional work is required to enhance the terminology so that it can support expanded use cases, such as managing drug lists for CDS. We encourage the NLM, content developers of drug knowledge bases, and the scientific community to continue to evaluate and collaboratively extend RxNorm for a variety of clinical uses.

Acknowledgements

This work was supported by the NIH/NIGMS (U19 GM61388; the Pharmacogenomics Research Network) and the Mayo Clinic Office of Information and Knowledge Management. The authors would like to thank the NLM staff that support RxNorm for providing critical clarifications regarding the content and relationships represented in the data set.

References

1. Services CfMM. Meaningful use. [1/23/2014]; Available from: www.cms.gov/Regulations-and-guidance/Legislation/EHRIncentivePrograms/Meaningful_Use.html.
2. Garg AX, Adhikari NK, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. Jama. [Research Support, Non-U.S. Gov't
Review]. 2005 Mar 9;293(10):1223-38.
3. Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. Bmj. [Research Support, U.S. Gov't, P.H.S.
Review]. 2005 Apr 2;330(7494):765.
4. Vora MB, Trivedi HR, Shah BK, Tripathi CB. Adverse drug reactions in inpatients of internal medicine wards at a tertiary care hospital: A prospective cohort study. J Pharmacol Pharmacother. 2011 Jan;2(1):21-5.

5. Technology ANRCfHI. Challenges and Barriers to Clinical Decision Support (CDS) Design and Implementation Experienced in the Agency for Healthcare Research and Quality CDS Demonstrations. 2010 [3/13/2014]; Available from: http://healthit.ahrq.gov/sites/default/files/docs/page/CDS_challenges_and_barriers.pdf.
6. Sittig DF, Wright A, Osheroff JA, Middleton B, Teich JM, Ash JS, et al. Grand challenges in clinical decision support. Journal of biomedical informatics. [Research Support, N.I.H., Extramural Review]. 2008 Apr;41(2):387-92.
7. Shah NR, Seger AC, Seger DL, Fiskio JM, Kuperman GJ, Blumenfeld B, et al. Improving acceptance of computerized prescribing alerts in ambulatory care. J Am Med Inform Assoc. [Research Support, U.S. Gov't, P.H.S.]. 2006 Jan-Feb;13(1):5-11.
8. Fox BI, Thrower MR, Felkey BG, American Pharmacists Association. Building core competencies in pharmacy informatics. Washington, D.C.: American Pharmacists Association; 2010.
9. Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R. Normalized names for clinical drugs: RxNorm at 6 years. J Am Med Inform Assoc. 2011 Jul-Aug;18(4):441-8.
10. Fung KW, McDonald C, Bray BE. RxTerms - a drug interface terminology derived from RxNorm. AMIA Annu Symp Proc. 2008:227-31.
11. U.S. Department of Veterans Affairs VHA. National Drug File – Reference Terminology (NDF-RT™) Documentation. In: U.S. Department of Veterans Affairs VHA, editor.2012.
12. Zhu Q, Freimuth RR, Pathak J, Chute CG. PharmGKB Drug Data Normalization with NDF-RT. AMIA Summits Transl Sci Proc. 2013;2013:180.
13. Zhu Q, Freimuth RR, Pathak J, Durski MJ, Chute CG. Disambiguation of PharmGKB drug-disease relations with NDF-RT and SPL. J Biomed Inform. 2013 Aug;46(4):690-6.
14. Li J, Lu Z. Automatic identification and normalization of dosage forms in drug monographs. BMC Med Inform Decis Mak. 2012;12:9.
15. Zhou L, Plasek JM, Mahoney LM, Chang FY, DiMaggio D, Rocha RA. Mapping Partners Master Drug Dictionary to RxNorm using an NLP-based approach. J Biomed Inform. 2012 Aug;45(4):626-33.
16. Medicine USNLo. RxNorm Files. [2/14/2014]; Available from: http://www.nlm.nih.gov/research/umls/rxnorm/docs/rxnormfiles.html.
17. Medicine USNLo. RxTerms Data Files. [2/17/2014]; Available from: http://wwwcf.nlm.nih.gov/umlslicense/rxtermApp/rxTermData.cfm.
18. Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res. [Research Support, N.I.H., Extramural]. 2009 Jul;37(Web Server issue):W170-3.
19. Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, et al. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Res. [Research Support, N.I.H., Extramural]. 2011 Jul;39(Web Server issue):W541-5.
20. Medicine USNLo. RxNorm Technical Documentation. [3/13/2014]; Available from: http://www.nlm.nih.gov/research/umls/rxnorm/docs/2014/rxnorm_doco_full_2014-1.html.
21. Medicine USNLo. Appendix 1 - RxNorm Relationships (RELA). [3/12/2014]; Available from: http://www.nlm.nih.gov/research/umls/rxnorm/docs/2014/appendix1.html.
22. Zeng K, Bodenreider O, Kilbourne J, Nelson S. RxNav: a web service for standard drug information. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2006:1156.
23. Medicine USNLo. Appendix 3 - RxNorm Dose Form Groups (TTY=DFG). [3/10/2014]; Available from: http://www.nlm.nih.gov/research/umls/rxnorm/docs/2014/appendix3.html.
24. Pathak J, Murphy SP, Willaert BN, Kremers HM, Yawn BP, Rocca WA, et al. Using RxNorm and NDF-RT to classify medication data extracted from electronic health records: experiences from the Rochester Epidemiology Project. AMIA Annu Symp Proc. 2011;2011:1089-98.
25. Bodenreider O. Biomedical ontologies in action: role in knowledge management, data integration and decision support. Yearb Med Inform. 2008:67-79.
26. Palchuk MB, Klumpenaar M, Jatkar T, Zottola RJ, Adams WG, Abend AH. Enabling Hierarchical View of RxNorm with NDF-RT Drug Classes. AMIA Annu Symp Proc. 2010;2010:577-81.
27. Pathak J, Chute CG. Analyzing categorical information in two publicly available drug terminologies: RxNorm and NDF-RT. J Am Med Inform Assoc. 2010 Jul-Aug;17(4):432-9.
28. Hanna J, Joseph E, Brochhausen M, Hogan WR. Building a drug ontology based on RxNorm and other sources. J Biomed Semantics. 2013;4(1):44.

Coverage of Rare Disease Names in Standard Terminologies and Implications for Patients, Providers, and Research

Kin Wah Fung[1], Rachel Richesson[2], Olivier Bodenreider[1]
[1]National Library of Medicine, Bethesda, MD
[2] Duke University, Durham, NC
kwfung@nlm.nih.gov|rachel.richesson@dm.duke.edu|obodenreider@mail.nih.gov

Abstract

Small numbers of patients are a special challenge for rare diseases research. Electronic health record (EHR) data can facilitate research if patients with rare diseases can be reliably identified. We estimate the coverage of the names of a set of 6,519 rare diseases. Using the UMLS, 697 (11%) diseases were matched to ICD-9-CM, 1,386 (21%) to ICD-10-CM and 2,848 (44%) to SNOMED CT. Using published mappings from SNOMED CT to ICD, we further estimate additional broader matches of 2,569 (39%) rare diseases to ICD-9-CM and 1,635 (25%) to ICD-10-CM. The number of codes that match one and only one disease are 1,081 (62%) for ICD-9-CM, 1,403 (73%) for ICD-10-CM, and 3,311 (85%) for SNOMED CT. Our findings confirm that SNOMED CT has the greatest coverage and specificity needed to identify patients with a rare disease from EHR-data, and can facilitate research and evidence-based care.

Introduction

Rare diseases are defined in the US as conditions that affect less than 200,000 Americans and in the European Union as those with a prevalence of 5 per 10,000 or less. They are largely, but not exclusively, genetic disorders. Because of the variation in definition, there is no globally authoritative list of rare diseases, but the number of recognized rare diseases is between 6-7,000 diseases.[1-4] Although each condition is uncommon, collectively rare diseases are more common. The National Organization for Rare Disorders estimates that up to 30 million (or 1 in 10) Americans are affected by a rare disease.[3] Consequently, rare diseases have emerged as priority research topics in both the US and the EU. Advances in genetic testing and the emergence of personalized medicine increase the number of subtypes of common diseases, further motivating development of research methods for rare diseases.

To identify sufficient numbers of rare diseases patients for research, multiple clinical sites and countries are often required and electronic health record (EHR) data can facilitate the identification of rare disease patients across multiple sites. Phenotype definitions that leverage widely adopted administrative terminologies (such as ICD-9-CM and ICD-10-CM) can potentially enable the consistent identification of patients with rare diseases from different providers and organizations. Our collective national capacity for rare diseases research, therefore, depends upon adequate coverage of rare diseases in these administrative terminologies. On the other hand, according to the "meaningful use" incentive program for the use of EHRs, clinical terminologies (such as SNOMED CT) are required for the encoding of clinical information in the EHR.[5] The encoded clinical information will in turn drive the EHR-embedded information and decision support (e.g., InfoButtons[6], clinical practice guidelines) and consumer health information (e.g., MedlinePlus Connect) functionalities. To comprehend the national capacity for rare diseases research, EHR-enabled clinical decision support and patient education, we estimate the coverage of rare diseases in ICD-9-CM, ICD-10-CM, and SNOMED CT for a set of 6,519 rare diseases, and explore the granularity of rare disease terms in these terminologies. We examine in detail the matches found for a set of rare diseases that are being studied in the national Patient Centered Outcomes Research network (PCORnet), recently established this year to create a national infrastructure for observational and clinical research in diverse and distributed healthcare organizations. [7]

Background

Rare diseases have become an increasingly important topic in health care research and policy contexts. Rare diseases are explicitly represented in important federally-funded research initiatives, such as the Rare Diseases Clinical Research Network[8] and the Clinical and Translational Science Awards (CTSA) program. Despite differences in

disease etiology and affected populations, there are common logistical challenges for research that can be addressed in part with data standards and informatics expertise and tools.[9-12] Generalizable research methods that address issues specific to rare diseases can impact the investigation of thousands of rare conditions and hundreds of thousands of Americans. With increased adoption and meaningful use of EHRs, there is renewed effort in leveraging EHRs for research. The Patient-Centered Outcomes Research Institute (PCORI) was funded from the Affordable Care Act to examine real-world treatment decisions.[13] PCORI network is specifically tasked to conduct observational and interventional research on the comparative effectiveness of various treatments using distributed and heterogeneous healthcare organizations and various EHR systems. PCORI currently supports the research for approximately 50 rare diseases (see appendix). The motivation for this paper was to explore the coverage of rare diseases in standard terminologies in order to characterize the current capacity for EHR-based research on those diseases, and to suggest strategies that will increase the national research capacity for all rare diseases.

There are several initiatives that have complete or partial inventories of rare disease names and terms. The Office of Rare Diseases Research (ORDR) of the National Center for Advancing Translational Sciences (NCATS) in the U.S. and Orphanet in the E.U. recognize 6 -7,000 disorders as rare diseases and support various efforts to link these disease names to standard terminologies. The ORDR also supports the Genetic and Rare Diseases Information Center (GARD), a web-based information resource for the public on more than 6,000 rare diseases.[14] The Genetics Home Reference, maintained by the National Library of Medicine, includes a smaller set of approximately 800 genetic diseases, most of which are rare. The NLM has identified and validated SNOMED CT codes for these 800 diseases to supports public retrieval of information. Orphanet, an EU-wide advocacy and information organization funded by national and European public institutions and patient organizations, foundations and corporations, provides information to the public on approximately 7,000 rare disorders on its web-based Portal for Rare Diseases and Orphan Drugs.[1] Orphanet also sponsors OrphaData, which provides the scientific community with data and tools to support the identification, quantification, and research of rare disorders. As part of this effort, Orphanet recently developed a rare disease ontology (ORDO) which serves as an inventory and classification of rare diseases, cross-referenced with OMIM, ICD-10, and SNOMED-CT and with genes in HGNC, OMIM, UniProtKB and Genatlas.[15, 16]

Existing standard terminologies, such as ICD and SNOMED CT are important components for EHRs and rare diseases research. Over 3,000 distinct concepts (including diagnoses, findings, treatments and procedures) from 4 medical centers were used to evaluate the content coverage of these and other clinical coding systems.[17] Although no coding system captured all concepts, SNOMED was the most complete. The authors concluded that both ICD-9-CM and ICD-10 fail to capture substantial clinical content, and warned that analytic conclusions that depend on these coding systems may be suspect. ICD-10 is critical for global surveillance of rare diseases. The Clinical Modifications (CM), e.g., ICD-9-CM and ICD-10-CM are critical for billing and reimbursement in the U.S. SNOMED CT is becoming increasingly adopted as a supporting clinical terminology in EHR systems worldwide[18], and gaining attention in the US since being named as a reporting standard for problem lists.[5] Previous studies have shown significant inclusion of rare diseases in SNOMED CT[19, 20], and anecdotally the IHTSDO (International Health Terminology Standards Development Organisation) is committing to increasing this coverage. SNOMED CT plays an important role in context-aware knowledge retrieval applications (i.e., InfoButtons) and the identification of patient-directed consumer information from the various information resources such as the Genetic Home Reference and MedlinePlus.

There are different approaches to identifying rare disease names in standard terminologies. Rare disease names from the Office of Rare Diseases have been mapped to the Unified Medical Language System (UMLS) to facilitate coding in other systems such as Medical Subject Headings (MeSH) for medical literature, ICD for public health surveillance, and SNOMED CT for use in clinical records documentation and clinical decision support.[19] In 2010, the NLM mapped 8,435 rare disease names (collected from ORDR, Orphanet, and the National Organization for Rare Disorders, a patient advocacy and voluntary health organization in the US) to the UMLS, and found different levels of coverage for Medical Subject Headings (MeSH) (5,663 ; 67%), Online Mendelian Inheritance in Man (OMIM) (3,802 ; 45%), SNOMEDCT (4,192 ; 50%), and ICD-10 (1,029 ;12%).[20]

In this investigation, we re-examine the current coverage of rare disease names in standard coding systems to support two use cases: 1) the identification of rare disease patients from EHR data for research, and 2) the identification of appropriate rare diseases information, including published medical literature, clinical practice guidelines for providers and authoritative consumer-directed information for patients, using coded data from EHRs.

Further, we explore differences in granularity between various terminologies, all in the context of how EHRs can support the consistent and reliable identification of rare disease patients, to enable evidence-based care and multi-site research.

Methods

Estimating the Coverage of Rare Diseases

We estimated the coverage of rare diseases in the three terminologies: ICD-9-CM, ICD-10-CM and SNOMED CT. To do this, we used two resources: the UMLS and the published maps from SNOMED CT to ICD-9-CM (developed by IHTSDO) and ICD-10-CM (developed by NLM). We first matched the 6,519 ORDR rare diseases by their names to the UMLS using lexical matching, utilizing both exact and normalized string matches, followed by semantic group validation (with restriction to the Semantic Group Disorders). Through the UMLS concept structure, we identified matches to SNOMED CT, ICD-9-CM and ICD-10-CM codes (we call these *UMLS-identified matches*). We anticipated that the UMLS-identified codes were mostly equivalent matches, since the UMLS concept structure is based on synonymy (i.e., not broader or narrower matches). For ORDR rare diseases with UMLS-identified SNOMED CT match but no ICD match, we further used the SNOMED CT to ICD-9-CM and ICD-10-CM published maps as an alternative path to match to ICD-9-CM and ICD-10-CM codes (we call these *map-identified matches*). (Figure 1) The published maps enabled us to identify matches other than equivalent matches, since the ICD map targets could be broader (often) or narrower (seldom) than the SNOMED CT concept. Since the published maps did not cover all of SNOMED CT, we extrapolated the results to estimate the number of map-identified matches that we could potentially find if all SNOMED CT concepts were included in the published maps.

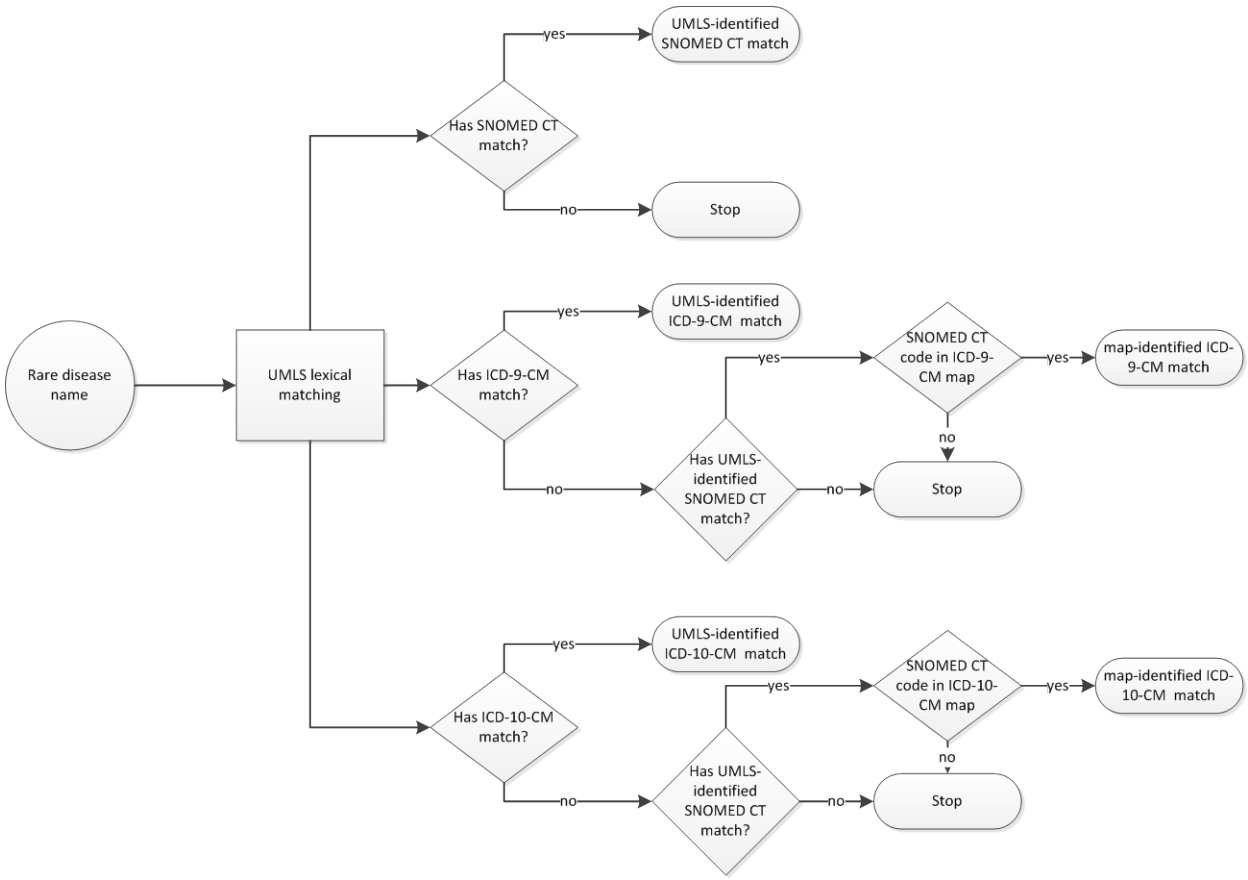


Figure 1. Overview of matching methods.

Estimating the Granularity of Matches for Rare Diseases
To study the impact of lack of equivalent matches for rare diseases, we looked at the ability of a specific code in the three terminologies to identify a specific disease. Using the matches identified above, we calculated the extent to which multiple rare diseases were included in a single code in the three terminologies.

Manually Validating a Sample of Matches
We reviewed a small set of rare diseases to validate our matching methods. We identified 46 rare disease categories under study in the PCORnet, specifically those diseases listed on applications for the 11 funded Clinical Data Research Networks (CDRNs) and 15 funded Patient Powered Research Networks (PPRNs).[7] We exploded some disease categories like 'vasculitis' and 'primary immunodeficiency diseases' to include specific diseases, such as Churg-Strauss Syndrome and Severe Combined Immunodeficiency.

The matches in the three terminologies found for the PCORnet rare diseases were reviewed by two authors (RR, medical informatician; KWF, physician) familiar with medical terminologies. Specifically, each assessed whether the match for the PCORnet diseases was an equivalent, narrower (more precise), broader (less precise) or related match. Where there was discrepancy between reviewers, consensus was reached through discussion.

As a general reference for comparison, we also calculated the coverage and granularity for the Orphanet's ORDO, based on the accompanying SNOMED CT mappings in the ontology.

Results

Estimating Rare Diseases Coverage
Using the names and synonyms of the 6,519 ORDR rare diseases for lexical matching in the UMLS, 697 (11%), 1,386 (21%) and 2,848 (44%) diseases were matched to ICD-9-CM, ICD-10-CM and SNOMED CT respectively. These were the UMLS-identified matches.

Among the 5,822 rare diseases with no UMLS-identified match to ICD-9-CM, 2,783 SNOMED CT matches were found, of which 80% (2,448 SNOMED CT codes) were included in the ICD-9-CM published map. This yielded map-identified matches for 2,055 (32%) diseases to ICD-9-CM. If all SNOMED CT concepts were included in the published ICD-9-CM map, the projected map-identified matches for ICD-9-CM would be 2,569 (39%) diseases. Similarly, among the 5,133 rare diseases with no UMLS-identified ICD-10-CM match, 1,841 SNOMED CT matches were identified, of which 56% (1,035 SNOMED CT codes) were included in the ICD-10-CM published map, which yielded map-identified matches for 919 (14%) diseases. The projected map-identified match for ICD-10-CM was 1,635 (25%) diseases. (Table 1)

As a comparison, Orphanet's rare disease ontology (ORDO) contained 6,750 diseases, among them 1,446 (21%) diseases were accompanied by matches to SNOMED CT.

Table 1. Coverage of rare diseases in the three terminologies.

	UMLS-identified match	Map-identified match (found)	Map-identified match (projected)
ICD-9-CM	697 (11%)	2055 (32%)	2569 (39%)
ICD-10-CM	1386 (21%)	919 (14%)	1635 (25%)
SNOMED CT	2848 (44%)	n/a	n/a

Estimating the Granularity of Matches for Rare Diseases

Using the matches identified above, we calculated the extent to which multiple diseases were included in a single code in the three terminologies. (Table 2) We define a *unique match* as a code that matches to only one rare disease,

and a *multiple match* as a code that matches to more than one rare disease. The number and proportion of unique matches were 1,081 (62%) for ICD-9-CM, 1,403 (73%) for ICD-10-CM, and 3,311 (85%) for SNOMED CT. Overall, 672 (38%) of the matched ICD-9-CM codes were multiple matches, which was lower for ICD-10-CM (n=526, 27%) and lowest for SNOMED CT (n=598, 15%). As for the cardinality of the broader matches, the maximum number of diseases matched to a SNOMED CT code was 5 diseases. The highest number of diseases matched to a single code was 208 for ICD-9-CM and 23 for ICD-10-CM. There were 117 ICD-9-CM and 40 ICD-10-CM codes matching to more than 5 diseases.

In Orphanet's ORDO, 1,446 rare diseases were matched to 1,748 SNOMED CT codes. Most of the SNOMED CT codes (1,735, 99%) were matched to a single disease, and 13 SNOMED CT codes were matched to two diseases.

Table 2. Number of Rare Diseases Included in Matched Codes from Source Terminologies.

# rare diseases matching to a code	# ICD-9-CM codes (% of total codes)	# ICD-10-CM codes (% of total codes)	# SNOMED CT codes (% of total codes)
1 (unique match)	1081 (62%)	1403 (73%)	3311 (85%)
2	319	328	478
3	125	88	84
4	68	45	33
5	43	25	3
> 5	117	40	0
# codes matching to > 1 disease (% of total codes) (multiple match)	672 (38%)	526 (27%)	598 (15%)
Examples	208 rare diseases matched to *759.89 Other specified congenital anomalies*	22 rare diseases matched to *Q82.8 Other specified congenital malformations of skin*	5 rare diseases matched to *28835009 Retinitis pigmentosa*

Table 3. Manual review of PCORnet rare diseases matches.

		ICD-9-CM	ICD-10-CM	SNOMED CT
UMLS-identified matches (% of all UMLS-identified matches)	equivalent	14 (93%)	13 (68%)	46 (74%)
	broader	0 (0%)	3 (16%)	2 (3%)
	narrower	1 (7%)	2 (11%)	9 (15%)
	related	0 (0%)	1 (5%)	5 (8%)
	total	15 (100%)	19 (100%)	62 (100%)
Map-identified matches (% of all map-identified matches)	equivalent	1 (2%)	9 (25%)	n/a
	broader	45 (87%)	23 (64%)	n/a
	narrower	4 (8%)	3 (8%)	n/a
	related	2 (4%)	1 (3%)	n/a
	total	52 (100%)	36 (100%)	n/a
# diseases with equivalent match (% of total diseases)		15 (28%)	22 (42%)	43 (81%)
# diseases with no equivalent match		30 (57%)	23 (43%)	2 (4%)
# diseases with no match		8 (15%)	8 (15%)	8 (15%)
Total # diseases		53 (100%)	53 (100%)	53 (100%)

Manual Review of Matches

Among the UMLS-identified matches, the proportions of equivalent matches were 93%, 68% and 74% for ICD-9-CM, ICD-10-CM and SNOMED CT respectively. Among the map-identified matches, 87% of the ICD-9-CM

matches and 64% of the ICD-10-CM matches were broader matches. Overall, 8 (15%) of the 53 diseases could not be matched to any of the three terminologies. Among the 45 diseases that could be matched, an equivalent match could be found for 15 (28%), 22 (42%), 43 (81%) diseases for ICD-9-CM, ICD-10-CM and SNOMED CT respectively. (Table 3)

Discussion

We first estimated the coverage of ORDR rare disease names in ICD-9-CM, ICD-10-CM, and SNOMED CT by lexical mapping to the UMLS. As expected, we found increasing coverage from ICD-9-CM (697; 13%) to ICD-10-CM (1,386; 26%), with the highest coverage for SNOMED CT (2,848; 53%). This is consistent with previous findings on the higher general clinical coverage of SNOMED CT. [17] The coverage of rare disease names in SNOMED CT is slightly lower than the 50% coverage seen in 2010, [20] although the earlier study used a larger set of rare disease names from multiple sources. As shown by the manual review, most of the UMLS-identified matches are equivalent matches.

This study differs from earlier work by Pasceri [20] in that we did not stop at lexical matching by the UMLS. In view of the low coverage of rare disease names in ICD-9-CM and ICD-10-CM, we explored new ways to match to these terminologies. We made use of the published maps from SNOMED CT to ICD-9-CM and ICD-10-CM to provide a cross-walk to the ICDs via SNOMED CT. This could considerably increase the matching rates to ICD-9-CM (from 11% to 50%) and ICD-10-CM (from 21% to 46%). However, most of these additional matches were not equivalent matches, and were either broader (majority) or narrower/related (minority) matches, as confirmed by the manual review. While these non-equivalent matches may still be useful in some use cases, e.g., to narrow down a large cohort of patients to those who may be suffering from a rare disease, they may not be precise enough to support direct patient care e.g. offering specific advice on the treatment of a particular disease.

Our work specifically supports use cases related to the use of EHRs to support the consistent and reliable identification of rare disease patients to enable evidence-based care and multi-site research. When searching health system data for rare disease patients, fine-grained and specific codes are preferable. Our data show that SNOMED CT has more codes (than ICD-9-CM or ICD-10-CM) that relate to one and only one disease. Due to the need to support statistical analysis in ICD-9-CM and ICD-10-CM, grouper concepts are more prevalent. This will cause problem when a code is required to identify a specific rare disease, such as linking an entry in the EHR to some disease-specific information. Our analysis shows that a higher percentage of ICD-9-CM and ICD-10-CM codes (38% and 27% respectively) lead to more than one rare disease, compared to 15% for SNOMED CT. One ICD code can lead to hundreds of diseases while the number is much smaller in SNOMED CT.

Although the PCORnet rare diseases that we explored in detail are not necessarily representative of all the rare diseases, these conditions are important in that they are being studied now. The proportion of diseases with equivalent match for the PCORnet list is considerably higher than what we saw for the 6,519 rare diseases overall. It is possible that the PCORnet funded research addresses more well-known or important diseases, so that more of them make their way into standard terminologies. Even if the small set of rare diseases that we used for validation are not representative of other rare diseases, their association with the PCORnet national research network that is actively exploring the use of EHR data in observational and interventional research will make them exemplars for refining strategies to increase the national capacity for rare diseases research and evidence-based care.

Limitations of our study include the following. We only focused on one source of rare disease names (ORDR). We did not do a comprehensive review of all the matches found in the three terminologies. The PCORnet rare diseases that we reviewed might not be representative of all rare diseases. Using the published maps to cross-walk to ICD-9-CM and ICD-10-CM was only possible for those diseases with UMLS-identified SNOMED CT matches.

The Way Forward

ICD and SNOMED CT are designed for very different purposes, and the rare disease community can benefit from knowing this distinction. As a classification system, ICD by definition includes categories that are designed to be exhaustive and mutually exclusive. For statistical purposes (in epidemiology and billing use cases) the idea of

"multiple counting" (e.g., classifying a disease in two different hierarchies) is discouraged, and residual categories (e.g., some diseases 'not elsewhere classified') can be meaningful. In fact, one reason for the residual categories is to avoid the need to assign codes to diseases with very low prevalence (i.e. rare diseases) and to maintain statistical balance of the coding categories. On the other hand, SNOMED CT as a clinical terminology is designed to support the representation of any concept to be stated about the patient. In the context of a terminology, "multiple counting" is desirable, and residual categories are meaningless. For example, for a patient suffering from Laurence-Moon syndrome, the diagnostic label "Other specified congenital malformation syndromes, not elsewhere classified" will not be useful in finding disease-specific patient education information or clinical practice guidelines. Another advantage of SNOMED CT is its shorter update cycle of 6 months, compared to yearly updates for ICD.

There is great potential benefit in the use of SNOMED CT as a source terminology in EHRs. Using a robust and fine-grained terminology such as SNOMED CT, clinicians can document patient data once at the point of care with fidelity to the clinical situation, at the appropriate level of granularity and certainty or uncertainty. These data, encoded in SNOMED CT, could be re-used for automated decision support and accessing customized information at the point of care. Others have shown that coarse disease classifications, such as ICD, are insufficient for these purposes, and our data indicate that SNOMED CT has higher proportion of fine-grained, or highly specific, codes for each disease. The use of SNOMED CT in the EHR can enable clinicians to record data only once at the point of care. Then the mappings of SNOMED CT to ICD classifications can be used to leverage these data for epidemiologic purposes, health system management, and billing. This would avoid duplicate effort of double-coding and potentially avoid the skewing of clinical data for billing purposes. To see how this will work for ICD-10-CM take a look at the I-MAGIC demo tool at NLM's website. [21]

Rare disease advocates should work to include rare diseases specifically in SNOMED CT. In future, a tighter integration between SNOMED CT and ICD-11 is anticipated. Concepts in SNOMED CT will then find a natural path into ICD, avoiding the risk of code translation or mapping errors. This is in line with Orphanet's exhortation for inclusion of more rare diseases in ICD.

Conclusions

To support patient care, patient education and research in rare diseases, adequate coverage of rare diseases in standard terminologies is essential. Existing coverage in SNOMED CT is higher than ICD-9-CM and ICD-10-CM, and with higher precision. More work is needed to improve coverage.

Acknowledgements

This work was partly supported by the Intramural Research Program of the National Institutes of Health and the National Library of Medicine.

Appendix. List of rare disease categories (46) studied in PCORI networks and grant awards.

Adrenoleukodystrophy	Granulomatosis with Polyangiitis	Phelan-McDermid Syndrome
Aicardi Syndrome	Hepatitis	Primary Immunodeficiency Diseases
alpha-1 antitrypsin deficiency	Hypoplastic left heart syndrome	Primary Nephrotic Syndrome (Focal Segmental Glomerulosclerosis)
Alström syndrome	Hypothalamic Hamartoma	Pseudoxanthoma elasticum
Amyotrophic Lateral Sclerosis	Inflammatory breast cancer	Psoriasis
Becker muscular dystrophy	Joubert syndrome	Pulmonary fibrosis
Chronic Granulomatous Disease	Juvenile Rheumatic Disease	Rare Cancers
Churg-Strauss Syndrome	Kawasaki Disease	Selective IgA Deficiency
Co-infection with HIV and hepatitis C virus	Klinefelter syndrome and associated conditions	Severe Combined Immunodeficiency
Common Variable Immunodeficiency	Lennox-Gastaut Syndrome	Severe Congenital Heart Disease
Cystic fibrosis	Membranous Nephropathy	Sickle Cell Disease
DiGeorge Syndrome	Metachromatic leukodystrophy	Sickle cell disease; Recurrent C. Difficile colitis
Dravet Syndrome	Microscopic Polyangiitis	Tuberous Sclerosis
Duchenne muscular dystrophy	Minimal Change Disease	X-Linked Agammaglobulinemia
Dyskeratosis congenital	Multiple Sclerosis	
Gaucher disease	Pediatric Transverse Myelitis	

References

1. Orphanet. *The portal for rare diseases and orphan drugs*. 2014 [cited 2014 July 21]; Available from: http://www.orpha.net/consor/cgi-bin/index.php.
2. NIH. *Office of Rare Diseases Research (ORDR) Brochure*. 2009 [cited 2014 July 21]; Available from: http://rarediseases.info.nih.gov/asp/resources/ord_brochure.html.
3. NORD. *Rare Disease Information*. 2014 [cited 2014 July 21]; Available from: http://www.rarediseases.org/rare-disease-information.
4. European Commission Public Health Policy. *Rare Diseases*. [cited 2014 July 21]; Available from: http://ec.europa.eu/health/rare_diseases/policy/index_en.htm.
5. Blumenthal, D. and M. Tavenner, *The "meaningful use" regulation for electronic health records.* N Engl J Med, 2010. **363**(6): p. 501-4.
6. Strasberg, H.R., G. Del Fiol, and J.J. Cimino, *Terminology challenges implementing the HL7 context-aware knowledge retrieval ('Infobutton') standard.* J Am Med Inform Assoc, 2013. **20**(2): p. 218-23.
7. PCORI. *PCORnet: The National Patient-Centered Clinical Research Network*. 2014 [cited 2014 July 21]; Available from: http://www.pcori.org/funding-opportunities/pcornet-national-patient-centered-clinical-research-network/.
8. Hampton, T., *Rare Disease Research Gets Boost.* JAMA 2006. **295** p. 2836-2838.

9. Griggs, R.C., et al., *Clinical research for rare disease: opportunities, challenges, and solutions.* Mol Genet Metab, 2009. **96**(1): p. 20-6.
10. Luisetti, M., et al., *The problems of clinical trials and registries in rare diseases.* Respir Med, 2010. **104 Suppl 1**: p. S42-4.
11. Seminara, J., et al., *Establishing a consortium for the study of rare diseases: The Urea Cycle Disorders Consortium.* Mol Genet Metab, 2010. **100 Suppl 1**: p. S97-105.
12. Eckfeldt, J.H. *Statement on the International Rare Diseases Research Consortium* Research and Innovation - Health 2011 [cited 2014 July 21]; Available from: http://ec.europa.eu/research/health/news-07_en.html.
13. Selby, J.V., A.C. Beal, and L. Frank, *The Patient-Centered Outcomes Research Institute (PCORI) national priorities for research and initial research agenda.* JAMA, 2012. **307**(15): p. 1583-4.
14. NIH. *Rare Diseases and Related Terms*. [cited 2014 July 21]; Available from: http://rarediseases.info.nih.gov/RareDiseaseList.aspx.
15. Orphanet. *Orphanet Rare Disease Ontology (ORDO)*. 2014 [cited 2014 July 21]; Available from: http://www.orphadata.org/cgi-bin/inc/ordo_orphanet.inc.php.
16. Rath, A., et al., *Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users.* Hum Mutat, 2012. **33**(5): p. 803-8.
17. Chute, C.G., et al., *The content coverage of clinical classifications. For The Computer-Based Patient Record Institute's Work Group on Codes & Structures.* J Am Med Inform Assoc, 1996. **3**(3): p. 224-33.
18. Lee, D., et al., *Literature review of SNOMED CT use.* J Am Med Inform Assoc, 2014. **21**(e1): p. e11-9.
19. Rance, B., et al., *Leveraging terminological resources for mapping between rare disease information sources*, in *Stud Health Technol Inform (Proc Medinfo)* 2013. p. 529-533.
20. Pasceri, E. *Analyzing rare diseases terms in biomedical terminologies. (LHNCB Medical Informatics Training Program Final Report; Dr. Olivier Bodenreider, Mentor)*. 2010 [cited 2014 July 21]; Available from: http://mor.nlm.nih.gov/pubs/alum/2010-pasceri.pdf.
21. NLM. *SNOMED CT to ICD-10-CM Map*. [cited 2014 July 21]; Available from: http://www.nlm.nih.gov/research/umls/mapping_projects/snomedct_to_icd10cm.html.

Validating Health Information Exchange (HIE) Data For Quality Measurement Across Four Hospitals

Nupur Garg, MD[1], Gil Kuperman, MD, PhD[2], Arit Onyile, MPH[1], Tina Lowry, MS[1], Nicholas Genes, MD, PhD[1], Charles DiMaggio, PhD[2], Lynne Richardson, MD[1], Gregg Husk, MD[3], Jason S Shapiro, MD[1].

[1]Icahn School of Medicine at Mount Sinai, New York, NY, [2]Columbia University and New York Presbyterian Hospital, New York, NY, Mount Sinai Beth Israel Medical Center, New York, NY

Abstract

Health information exchange (HIE) provides an essential enhancement to electronic health records (EHR), allowing information to follow patients across provider organizations. There is also an opportunity to improve public health surveillance, quality measurement, and research through secondary use of HIE data, but data quality presents potential barriers. Our objective was to validate the secondary use of HIE data for two emergency department (ED) quality measures: identification of frequent ED users and early (72-hour) ED returns. We compared concordance of various demographic and encounter data from an HIE for four hospitals to data provided by the hospitals from their EHRs over a two year period, and then compared measurement of our two quality measures using both HIE and EHR data. We found that, following data cleaning, there was no significant difference in the total counts for frequent ED users or early ED returns for any of the four hospitals ($p<0.001$).

Introduction

Health information exchange (HIE) is an important complement to electronic health records (EHRs), providing much needed outside clinical information to providers at the bedside.[1] Although comprehensive HIE at a state or national level is far from being fully realized, it is increasingly identified as a key element of our nation's approach to providing 21st century healthcare.[2-4] The Health Information Technology for Economic and Clinical Health (HITECH) portion of the American Recovery and Reinvestment Act (ARRA) has promoted the "meaningful use" (MU) of EHRs through three stages, each of which defines standards and provides incentives to compliant hospitals and providers. Whereas stage 1, meaningful use, focused on the technical ability to get information into a shareable electronic form, stages 2 and 3 focus on actual clinical use and driving improved outcomes with a progressively increased focus on HIE.[5]

HIE is not without challenges, facing questionable financial sustainability, adoption, and usage in many settings. [3,4,6-8] That said, the trend towards HIE use has continually increased,[6] with several recent studies showing that the *primary clinical use case* of clinicians reviewing the records of individual patients at the bedside can help prevent admissions, decrease testing, and ultimately save money.[7,9,10] However, studies showing that quality and safety are actually improved by HIE are still needed.[3,4,11,12]

As HIE expands, and we begin sharing data more broadly, health data is increasingly aggregated into ever larger nodes as we work toward a vision of national HIE. Through this progression, HIE data will increasingly serve as a data source for *secondary use cases* including care coordination,[13] population management,[14] public health surveillance,[15] broad community-wide quality measurement,[16] and research.[17] Ultimately these secondary uses may be a strong driving force for improving the safety and quality of patient care.

Because of the context-dependent nature of HIE data quality, meaning data may be of sufficiently high quality for one use case but not for another, HIE data that are "fit for use" in the primary clinical use case may not be "fit

for use" in secondary use cases.[18] As these secondary uses begin to test the limits of HIE data, the importance of ensuring HIE data quality should not be underestimated.

Background

Currently, EHRs are a common source of electronic data for quality measurement, many of these quality measurements are mandated by state and federal reporting requirements for research and for other secondary uses.[5,15,19,20] Because of this, the need for data quality analysis in the setting of EHRs has been recently described, and various methods have been explored for generating and rating data from EHRs to ensure a level of data quality.[18,21,22] For HIE, data quality assessment may be even more challenging, since aggregation of multiple data sources into an HIE multiplies the potential for data quality issues.

HIE networks are often initially formed to support the primary clinical use case, without much consideration for secondary use cases at their inception. HIEs often contain both clinical data from ancillary systems (e.g., lab results, diagnostic reports, clinical notes) and registration data from admission, discharge and transfer (ADT) registration systems (e.g. visit dates, patient demographics and other identifiers). For primary clinical use, these data are incorporated from multiple sites so that an individual clinician can view data belonging to an individual patient from multiple provider organizations. Initial HIE implementations may involve minimal validation, cleaning, or transformation. Testing may focus on interfaces between systems, and on assuring the proper display of clinical data at the presentation layer for individual patient data. Because HIE networks are usually designed around the primary clinical use case, data transformation to leverage the deep semantics at each site, mapping to appropriate terminologies to translate the meaning of data across sites, and detailed data quality and validation may be neglected. Once we begin to apply scenarios for the secondary use of HIE data, data quality issues are often discovered.[23,24]

In this analysis, we describe approaches to data cleaning, measuring data concordance before and after data cleaning, and validation of the secondary use of HIE data for two specific quality measures: identification of frequent emergency department (ED) users and early (72-hour) ED returns.

Methods

Setting

Healthix is a regional health information organization (RHIO) providing HIE to the New York metropolitan area and Long Island.[25] Healthix formed in 2012 through the merger of the Long Island Patient Information Exchange (Lipix) and the New York Clinical Information Exchange (NYCLIX).[26] Since then, the Brooklyn Health Information Exchange (BHIX) has also merged with Healthix,[27] and Healthix now has 9.2 million unique patients, > 6,500 users performing > 10,000 searches per month, and 107 participating organizations with 383 facilities comprising 29,946 acute and extended care beds as of January 2014.

To date, Healthix has enabled HIE primarily through the use of standard HL7 2.x messages,[28] written to a common specification, to send data from source systems at multiple provider organizations to edge servers at each site built with a common data structure. Registration information from each site's ADT system, including patient demographics and master patient index (MPI) functions, is housed in a centralized location.

As part of each site's implementation of HIE with Healthix, end-to-end interface testing was conducted. Artificial data was entered into test patient records in the site's source systems, and then followed downstream through the HIE interface, into the edge server, and finally to the display level in the HIE's portal viewer. Each step of the way, the data was checked for validity based on the source system, which consisted of making sure the interfaces were sending data properly and that the information was being properly displayed for clinicians. While this sort of testing is a standard aspect of HIE implementation, it may not detect problems that do not affect data display in the primary use case. For example, missing discharge date/time stamps might not be detected if all that is being tested is the ability to display laboratory and radiology results, but if the data are later employed for secondary use in measuring frequent ED users and early (72-hour) ED returns, then these missing data become much more important. As another example, free text diagnosis data might display properly during testing, but if analyses that require ICD-9 codes are attempted later, they may fail.

Data

EHR and HIE data elements from four hospitals were obtained for all ED visits from 3/1/09 to 2/28/11, including visit number (denotes a unique encounter), medical record number (MRN – denotes a unique patient), admission and discharge date and time, date of birth, and gender. All data were de-identified in accordance with HIPAA prior to analysis by the research team, and the protocol was reviewed by the Mount Sinai Program for the Protection of Human Subjects and given a not human research determination.

Analysis

HIE and EHR data for each site were merged on hashed visit numbers to evaluate concordance of unique encounters, and on hashed MRNs to evaluate concordance of unique patients. In this case, concordance is defined as having a unique match. Next, age, gender and the time-stamps for admissions/discharges were also tested for concordance. In order to adjust for differences between the EHR and HIE data for these latter four parameters, various data cleaning rules were systematically applied. These data cleaning techniques were derived from observation of the major areas of discrepancy noticed when concordance between the two data sets was tested, and through expert evaluation of workflow issues that were likely to have caused these discrepancies at each site. The three data cleaning techniques used were as follows:

1) Age was considered to match if the date of birth was less than or equal to one year difference between the HIE and EHR data.

2) Gender was considered to match if it was specified in either the EHR or HIE, and recorded as the same or "unknown" in the other system.

3) Admit and discharge times were considered to match if the difference in date/time was less than 6-24 hours. The number of hours from 6-24 was chosen on a site-specific basis by extending the data cleaning factor by the smallest multiple of 6 hours in which the concordance of encounters first became greater than 98%. This particular data cleaning technique was necessary because differences in clinical and registration staff workflows likely led to small but frequent discrepancies between the two data systems in which admit and discharge times were entered. For instance, when a clinician discharged an ED patient, a date/time stamp was immediately entered in the EHR but the registration staff member may wait until the end of his or her shift to remove the patient from the ADT system, causing the ADT date/time stamp to lag behind by a small number of hours.

The last part of the data analysis included measuring frequent ED users (patients with ≥ 4 visits in 30 days) and early (72-hour) ED returns (patients who return for a second ED visit within 72 hours of being discharged). The counts for each of these quality measures were then compared for statistically significant similarity between HIE and EHR datasets for each hospital using Chi square.

Results

Adjusted values (following data cleaning) were not significantly different between site-specific HIE and EHR datasets (Table 1). There was a high degree of concordance for unique encounters and patients between HIE and EHR data sets, so no data cleaning was employed.

Table 1	**Site 1**		**Site 2**		**Site 3**		**Site 4**	
	Unadj % Matched	Adj % Matched	Unadj % Matched	Adj % Matched	Unadj % Matched	Adj % Matched	Unadj % Matched	Adj % Matched
Unique Encounters (Visit #)	99.45	N/A	99.27	N/A	99.25	N/A	99.98	N/A
Unique Patients (MRN)	99.32	N/A	99.31	N/A	99.31	N/A	99.84	N/A
Age	99.71	100	97.61	100	97.69	100	97.94	100
Gender	99.25	100	99.16	99.91	99.16	99.91	99.61	99.63
Admit Date/Time	0.13	99.53	5.00	99.99	2.76	99.99	53.86	100
Discharge Date/Time	2.47	98.05	49.27	99.89	47.42	99.86	94.42	99.96

Table 1 shows the level of concordance between the EHR and HIE of the various data elements in the left column before and after adjustments were applied across all four sites. N/A signifies not applicable as no data cleaning was employed for these parameters.

The unadjusted match rate for age and gender ranged from 97.61% to 99.71% across sites (std. dev. 0.87%), and once the adjustment criteria were applied, all match rates increased and were greater than 99.6%. The admit and discharge date/time did not match well in the unadjusted data (range 0.13% to 94.42%, std. dev. 34.6%). Data cleaning adjustment allowed for a match in greater than 99.5% of admissions and 98% of discharges. The lowest discharge date/time unadjusted match rate was at Site 1, but only a six hour data cleaning time frame was needed to get the match percentage over 98%. Site 2-4 required a 24 hour adjustment of the time frame, and after adjustment had greater than 99.85% match rates across all sites.

When we measured the number of frequent ED users (patients with ≥ 4 ED visits in 30 days) and early (72-hour) ED returns, we found a high degree of concordance between HIE and EHR data sets (Table 2). All four sites have EHR counts that do not differ significantly from the HIE counts (p-value < 0.001).

Table 2	**Site 1**		**Site 2**		**Site 3**		**Site 4**	
	EHR Count	HIE Count	EHR Count	HIE Count	EHR Count	HIE Count	EHR Count	HIE Count
Frequent Users	1,204	1,221	1,060	1,035	1,746	1,708	936	924
72 Hour Returns	8,299	8,456	7,237	7,093	12,243	12,045	5,476	5,431

Table 2 shows the agreement between the EHR and HIE when the quality measures of the frequent ED users and the 72 hour returns were applied. EHR and HIE counts are not significantly different with a p-value < 0.001.

Discussion

In preparation for a project measuring frequent ED users and early (72-hour) emergency department (ED) returns across an HIE, this analysis was performed to validate the use of HIE data by comparing it to the electronic health record (EHR) data, since EHRs are often the source of data for these ED quality measures, and much of the encounter data in the HIE comes from ADT source systems. Our analysis shows that through some simple data cleaning and transformation, the level of concordance between EHR and HIE data sources for multiple data elements across four separate hospitals was very high. Furthermore, when we compared the performance of our two specific ED quality measures we found no statistical difference between EHR and HIE data.

There are several likely reasons that concordance between the data sets for demographic and encounter data exist, and the counts for frequent ED users and early (72-hour) ED returns were not identical. First, the data in the date and time stamps for the EHR are generally captured by clinicians directly as part of their EHR workflow, and the date and time stamps for the HIE were generally captured by registration staff in the sites' ADT source systems as part of their registration workflow. Differences in the timing of these two workflows likely led to some of the discrepancies. Second, age and gender data are generally captured by registration staff in ADT source systems, and then flow to both the sites' EHRs and to the HIE. It is possible that subsequent changes were occasionally made directly in the EHRs and not in the ADT systems, causing some of these discrepancies. Third, it is possible that some of the HL7 messages from various sites did not make it to the HIE due to occasional interface downtimes or other malfunctions, causing some small number of HL7 messages to be either dropped or altered. This may help explain why three of the four hospitals have lower counts for frequent ED users and early returns of ED patients from the HIE data than the EHR data. Some of the HIE cases were missing a discharge date/time, and in those we are unable to calculate a 72 hour return, so the counts would be lower. Regardless, the discrepancies between the HIE and EHR datasets was minimal in most cases, and in other cases could be addressed by simple data cleaning approaches.

This study has several important limitations. First, these analyses were performed on only a small subset of sites participating in Healthix (four out of more than 50 hospitals with EDs), and without further analysis of more sites, there is no way to determine if similar data quality issues would be encountered at the other sites, or if the data

cleaning techniques employed here would suffice. Also, the ED has traditionally been the focus of HIE interventions, but future studies should investigate data quality issues for inpatient and ambulatory domains. Second, some of the data quality issues here might be unique to this geographic region, though similar problems are likely to exist more broadly. Further analyses in other settings would need to be performed to make this determination. Finally, the analyses performed here were limited to HIE taking place using standard HL7 2.x interfaces. There is currently much work being done using newer XML-based continuity of care document approaches in HL7 Version 3 in RHIOs[28] and Integrating the Healthcare Enterprise protocols to exchange data between separate HIE networks.[29] There will likely be new and different data quality issues and data cleaning requirements that arise when these newer data transport standards are employed.

Secondary use of data gathered in electronic health records is increasing, raising the need for standardized data quality assessment. Without this, the validity of secondary uses that leverage electronic data may be questionable. Some of these secondary uses include quality measurement, chronic disease management and care coordination, population management, public health surveillance and observational comparative effectiveness research.

Health information exchange presents an appealing data source for these secondary uses of electronic data, and has the distinct advantage over EHR data in that it includes data from multiple provider organizations in a region. Patients often visit more than one provider organization, causing their healthcare data to become fragmented.[30-32] The use of HIE data for these secondary purposes therefore may more accurately reflect the manner in which patients interact with the healthcare system when compared to individual EHRs as a data source. However, the need for standardized data quality assessment and data cleaning is even greater when HIE serves as the data source because data from multiple sites are aggregated in an HIE, compounding data quality issues evident in individual EHRs. Further work should be done to determine if standardized data assessment and data cleaning techniques in the setting of an HIE can be developed, and if they differ from similar work that is being done at the level of the EHR.

Jason Shapiro was supported in part by the Agency for Healthcare Research and Quality (Grant No. 5R01HS021261). The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of the Agency for Healthcare Research and Quality

References

1. State Health Information Exchange Cooperative Agreement Program. 2011. (Accessed March 4, 2014, 2014, at http://www.healthit.gov/policy-researchers-implementers/state-health-information-exchange.)
2. Williams C, Mostashari F, Mertz K, Hogin E, Atwal P. From the Office of the National Coordinator: the strategy for advancing the exchange of health information. Health Aff (Millwood) 2012;31:527-36.
3. Vest JR. More than just a question of technology: factors related to hospitals' adoption and implementation of health information exchange. International journal of medical informatics 2010;79:797-806.
4. Genes N, Shapiro J, Vaidya S, Kuperman G. Adoption of health information exchange by emergency physicians at three urban academic medical centers. Applied clinical informatics 2011;2:263-9.
5. Meaningful Use Definition & Objectives. at http://www.healthit.gov/providers-professionals/meaningful-use-definition-objectives.)
6. Adler-Milstein J, Bates DW, Jha AK. Operational health information exchanges show substantial growth, but long-term funding remains a concern. Health Aff (Millwood) 2013;32:1486-92.
7. Vest JR. Health information exchange: national and international approaches. Advances in health care management 2012;12:3-24.
8. Kuperman GJ, McGowan JJ. Potential unintended consequences of health information exchange. J Gen Intern Med 2013;28:1663-6.
9. Wilcox AB, Shen S, Dorr DA, Hripcsak G, Heermann L, Narus SP. Improving access to longitudinal patient health information within an emergency department. Applied clinical informatics 2012;3:290-300.
10. Frisse ME, Johnson KB, Nian H, et al. The financial impact of health information exchange on emergency department care. Journal of the American Medical Informatics Association : JAMIA 2012;19:328-33.
11. Altman R, Shapiro JS, Moore T, Kuperman GJ. Notifications of hospital events to outpatient clinicians using health information exchange: a post-implementation survey. Informatics in primary care 2012;20:249-55.
12. Rudin R, Volk L, Simon S, Bates D. What Affects Clinicians' Usage of Health Information Exchange? Applied clinical informatics 2011;2:250-62.
13. Moore T, Shapiro JS, Doles L, et al. Event detection: a clinical notification service on a health information exchange platform. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium 2012;2012:635-42.
14. Shapiro JS, Mostashari F, Hripcsak G, Soulakis N, Kuperman G. Using Health Information Exchange to Improve Public Health. American Journal of Public Health 2011;101:616-23.
15. Shapiro JS, Genes N, Kuperman G, Chason K, Clinical Advisory Committee H1N1 Working Group NYCIE, Richardson LD. Health information exchange, biosurveillance efforts, and emergency department crowding during the spring 2009 H1N1 outbreak in New York City. Ann Emerg Med 2010;55:274-9.
16. Shapiro JS, Johnson SA, Angiollilo J, Fleischman W, Onyile A, Kuperman G. Health information exchange improves identification of frequent emergency department users. Health Aff (Millwood) 2013;32:2193-8.
17. Weiner MG, Embi PJ. Toward Reuse of Clinical Data for Research and Quality Improvement: The End of the Beginning? Annals of Internal Medicine 2009;151:359-60.

18. DQC White Paper Draft 1: A consensus-based data quality reporting framework for observational healthcare data. Data Quality Collaborative. http://repository.academyhealth.org/dqc/12013.
19. Frieling W. Beyond 'meaningful use'. Regional health information exchanges just as important to healthcare IT. Modern healthcare 2009;39:22.
20. Wright A, Feblowitz J, Samal L, McCoy AB, Sittig DF. The Medicare Electronic Health Record Incentive Program: provider performance on core and menu measures. Health Serv Res 2014;49:325-46.
21. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. Journal of the American Medical Informatics Association 2013;20:144-51.
22. Kahn MG, Raebel MA, Glanz JM, Riedlinger K, Steiner JF. A Pragmatic Framework for Single-site and Multisite Data Quality Assessment in Electronic Health Record-based Clinical Research. Medical Care 2012;50:S21-S9 10.1097/MLR.0b013e318257dd67.
23. Shapiro JS OA, Genes N, DiMaggio C, Kuperman G, Richardson LD. Validating health information exchange data for quality measurement. Ann Emerg Med 2013;62:S94.
24. Shapiro JS OO, DiMaggio C, Kuperman G. . Validating health information exchange data for quality measurement. In: AMIA, editor. Annu Symp Proc; 2013; Washington, D.C.
25. Healthix. 2013. (Accessed February 16, 2014, 2014, at https://services.lipixportal.org/HealthixPortal.)
26. Volpe S. LIPIX + NYCLIX Merge To form: Healthix. EHR PHR Patient Portals with Meaningful Use = Patient Centered Medical Home (PCMH). http://ehrphrpatientportal.blogspot.com/2011/11/lipix-nyclix-merge-to-form-healthix.html2011.
27. Becker AaVd. Healthix, Inc. and the Brooklyn Health Information Exchange (BHIX) announce plans to merge. https://services.lipixportal.org/Content/resources/RHIOMergerAnnouncment061113.pdf2013.
28. Health Level Seven International. Health Level Seven International, 2014. (Accessed March 6, 2014, 2014, at http://www.hl7.org/implement/standards/product_brief.cfm?product_id=185.)
29. Witting KaJM. Health Information Exchange: Enabling Document Sharing Using IHE Profiles 2012.
30. Bourgeois F OK, Mandl K. Patients treated at multiple acute health care facilities: quantifying information fragmentation. Arch Intern Med 2010:1989-95.
31. Finnell J OJ, Grannis S. All health care is not local: an evaluation of the distribution of emergency department care delivered in Indiana. AMIA Annual Symposium Proceedings; 2011; Washington, DC.
32. Grinspan ZM AE, Banerjee S, Kern L, Kaushal R, Shapiro JS. Potential value of health information exchange for people with epilepsy: crossover patterns and missing clinical data. In: AMIA, editor. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium; 2013; Washington, D.C. p. 527-36.

An Evaluation of a Natural Language Processing Tool for Identifying and Encoding Allergy Information in Emergency Department Clinical Notes

Foster R. Goss DO, MMSc[1], Joseph M. Plasek, MS[2], Jason J. Lau, BS[2], Diane L. Seger, RPh[3], Frank Y. Chang, MSE[4], Li Zhou, MD, PhD[2,4,5]

[1]Tufts Medical Center, Department of Emergency Medicine and Clinical Decision Making, Boston, MA; [2]Division of General Medicine and Primary Care, Brigham and Women's Hospital, Boston, MA; [3]Clinical & Quality Analysis, Partners HealthCare System, Boston, MA; [4]Clinical Informatics, Partners eCare, Partners HealthCare System, Boston, MA; [5]Harvard Medical School, Boston, MA

Abstract

Emergency department (ED) visits due to allergic reactions are common. Allergy information is often recorded in free-text provider notes; however, this domain has not yet been widely studied by the natural language processing (NLP) community. We developed an allergy module built on the MTERMS NLP system to identify and encode food, drug, and environmental allergies and allergic reactions. The module included updates to our lexicon using standard terminologies, and novel disambiguation algorithms. We developed an annotation schema and annotated 400 ED notes that served as a gold standard for comparison to MTERMS output. MTERMS achieved an F-measure of 87.6% for the detection of allergen names and no known allergies, 90% for identifying true reactions in each allergy statement where true allergens were also identified, and 69% for linking reactions to their allergen. These preliminary results demonstrate the feasibility using NLP to extract and encode allergy information from clinical notes.

Introduction

Allergic reactions result in over 1 million visits per year to emergency departments (ED)[1]. Foods and antibiotics are among the most common causes of these visits with an estimated 525,600 and 112,116 visits annually[2, 3]. Adverse drug events (ADEs) due to a patient receiving a medication to which they were known to be allergic is present in both the inpatient[4] and outpatient[5] settings. Careful documentation and encoding of patient allergies is critical to patient safety and ensuring drug allergy checking and clinical decision alerts are triggered. Natural language processing (NLP) has shown promise in this domain[6], yet few have adapted its use to detecting allergies from clinical notes, which often contains allergy information both within and outside the allergy section of the chart. This paper presents our early experience and preliminary findings in developing an allergy module for a general NLP system, named Medical Text Extraction, Reasoning, and Mapping System (MTERMS), to extract and encode allergy information from clinical text. We assessed the system performance of the MTERMS in processing free-text ED notes for allergen names and associated allergic reactions.

Background

Emergency physicians are often at the forefront of treating and managing acute allergic reactions. Their clinical notes typically convey the allergen, the patient's reaction and their response to treatment. While allergy information is usually recorded in the allergy section, it may not be included in the allergy section at the time of presentation or it may be located elsewhere (e.g., medical decision making or ED course). Additionally, the physician's narrative may include important details reflecting their certainty the patient is having an allergic reaction and the treatment given. As such, pertinent allergy information may be buried in clinical text, raising a concern for patient safety, as this information is not interoperable with computerized drug-allergy alerting or drug-drug interaction checking.

Most NLP applications to date have focused on other specific domains including clinical problems[7], medications[8], vaccination reactions[9], suicide ideation[10], smoking status[11], or ADE detection[12], etc. While allergy (a type of adverse event) is an important domain, it has not yet been widely investigated. In cTAKES[13], a NLP system by Savona et al, allergies to a given medication are handled by setting the

negation attribute of that medication to "is negated". In MedLEE, terms such as "allergy" or "toxic" are captured and assigned to the semantic type "reaction"[14]. Melton et al[15] used MedLEE[16, 17] to identify 45 adverse event types (e.g., loss or impairment of bodily functions) using discharge summaries. A set of computerized queries implemented the inclusion and exclusion logic for each event and identified the adverse events against the NLP output. Melton's approach achieved an overall sensitivity of 28%, specificity of 98.5%, and PPV of 45%. Most ADE detection algorithms use simple keyword search methods that search notes for relevant trigger words[14]. In general, these studies have reported low sensitivity (69%[18], 23%[19]) and PPV (7%[20], 12%[21], 52%[18], and 41%[19]). One study[14] searched keywords (e.g., "error," "mistake" or "iatrogenic") to detect medical errors in discharge summaries, residents' transfer of service notes, and outpatient visit notes with PPVs ranging from 3.4% to 24.4%, depending on the keywords used. Apart from performance issues, one major limitation of a keyword searching method is that it only considers a limited set of terms, hence most of these tools are unable to automatically extract and encode important clinical information (such as medications and symptoms). This lack of detailed clinical information as a part of the structured output limits its usage in further analysis or other research purposes.

Several studies[6, 22] using NLP techniques have been conducted on analyzing allergy repositories. Skentzos et al[22] developed NLP software to identify cholesterol lowering statin drug side effects documented in narrative provider notes and achieved a recall of 86.5% and precision of 91.9%. The Skentzos algorithm was further utilized to conduct a retrospective cohort study to determine the factors associated with documentation of statin side effects in a structured allergy repository[23]. Recently, a study by Epstein et al[6] evaluated the use of NLP to encode free-text food and drug allergens and mapped identified allergens to RxNorm. Epstein's study focused only on allergens within the patients allergy list and did not include reactions and environmental allergens. To date, no comprehensive NLP systems have been evaluated for processing allergy information in clinical notes with dynamic mapping to standard terminologies.

Automated encoding of allergy information is challenging as it requires the integration of multiple standard terminologies[24]. Initial standard terminology recommendations for encoding allergy have been put forth by the health information standards panel (HISTP). These include, SNOMED CT[25] for allergy/adverse reactions, RxNorm[26] for medications, National Drug File Reference Terminology (NDR-RT)[27] for drug classes and the Unique Ingredient Identifier[28] for food and substance allergens[29, 30]. The National Council for Prescription Drug Programs (NCPDP)[31] has proposed allergy value sets for encoding allergy that includes RxNorm as the source terminology. The Centers for Disease Control (CDC) has released a value set for encoding allergies that includes food, drug and environmental allergens, each mapped to SNOMED CT[32]. The coverage of five terminologies (RxNorm, SNOMED CT, UNII, NDF-RT and MedDRA) for encoding common allergies was recently evaluated by Goss et al who found SNOMED CT and RxNorm can satisfy most criteria for encoding common allergens with sufficient content coverage[24].

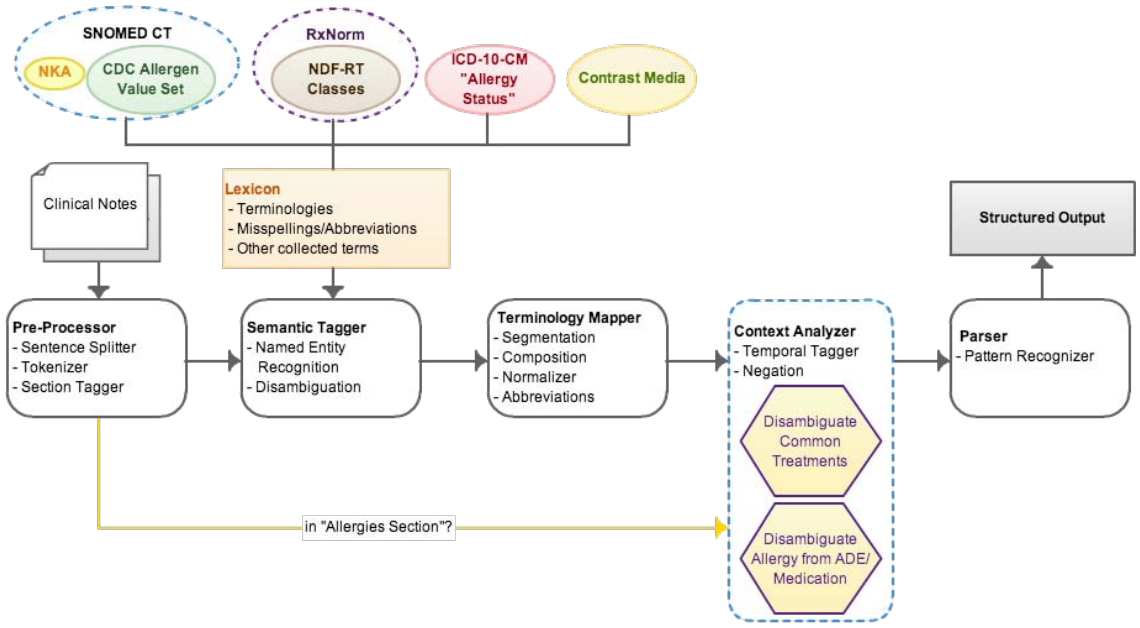
We have developed a generic NLP application, called Medical Text Extraction, Reasoning and Mapping System (MTERMS)[33], which is initially used for encoding medication information for medication reconciliation and for mapping between medication terminologies[33, 34]. MTERMS processes free-text entries into a structured XML output using a pipeline approach that includes a Pre-Processor, Semantic Tagger, Terminology Mapper, Context Analyzer, and Parser. The Terminology Mapper of MTERMS maps free-text medical concepts to multiple standard terminologies (e.g., SNOMED CT, RxNorm, NDF-RT, etc.) instead of a single specific terminology and where necessary, establishes dynamic mapping between these terminologies. In this study, we extended MTERMS by adding an allergy module for processing free-text allergy information.

Methods and System Design

System Design

The NLP allergy module is built on the MTERMS platform, and consists of a new lexicon for allergy concepts, an updated context analyzer consisting of new disambiguation algorithms, and updates to the XML output schema (Figure 1). MTERMS provides dynamic mapping of concepts across each terminology in its lexicon.

Figure 1: MTERMS system architecture for allergy module.



Lexicon Development

Our lexicon uses multiple recommended standard terminologies for encoding allergy concepts[24, 30] as well as local terminologies. For drug allergens, we use RxNorm and the same method to create the medication concept database described by Zhou et al[34]. We compiled a list of medications typically used to treat allergic reactions (e.g., Benadryl, Solumedrol) from allergy experts. Drug classes are inherited from NDF-RT (e.g., Ace Inhibitors), using those present in RxNorm and from ICD-10-CM "Allergy Status" terms (e.g., Antibiotics)[35]. For food and environmental allergen names, we use an available allergy value set from the CDC (Public Health Information Network Vocabulary Access and Distribution System (PHIN VADS)) as well as a subset of terms collected from the sub-tree (subClassOf) "allergen class" of SNOMED CT. Contrast media concepts were compiled from RxNorm and allergy experts. We use the 2014 versions of RxNorm, SNOMED CT, UNII, and the 2011 allergy value set from the CDC (Public Health Information Network Vocabulary Access and Distribution System (PHIN VADS)). To facilitate the integration of updates released from terminology sources, separate SQL tables were allocated to lexical variants or common misspellings.

For reaction names, we use a subset of terms compiled from the most frequently observed allergy reactions within Partners Enterprise-wide Allergy Repository (PEAR), a longitudinal allergy database shared within the Partners provider/hospital network[23]. These reactions were then mapped to SNOMED CT.

Search and Disambiguation Algorithms

Drug, food, and environmental agents mentioned in the clinical text can have different meanings. A drug may be a treatment for a condition or environmental agent may be the setting a patient was surrounded by. Since allergens and associated reactions can be found anywhere in a clinical note, our algorithms conduct disambiguation by considering contextual information. For example, when inside of the allergies section, our algorithms search and encode all matches to our lexicon for allergens and reactions due to the context of the section. Outside of the allergy section, our algorithms identify allergens and reactions using a set of rules (regular expressions) that search for the presence of indicators (e.g., allergic, allergy, caused, "started after", "likely due to"). If the pattern is recognized as being one indicating an allergy, then we add the annotations for the allergen and associated reactions. An example of a pattern outside of the allergies section is shown in Figure 2.

Medications used as common treatments for allergic reactions and their symptoms (e.g., hydrocortisone, Pepcid) often show up in the same sentence as the drug allergen, and thus need to be disambiguated. We

use a set of rules based on indicators (e.g., resolved, treated, relieved) taking into account position of indicator with treatment drug, relative to the reaction and allergen. These rules are applicable both within and outside the Allergies section of the clinical note. Our current disambiguation algorithms are limited to the sentence level.

We also compiled a subset of reaction terms or conditions that indicate an allergy (e.g., anaphylaxis, hives, Stevens-Johnson Syndrome, urticarial rash) to aid in disambiguation of allergic reactions from common reasons to visit the ED that are not necessarily immune-mediated (e.g., rash, itch). When any of these allergic reactions or conditions are found in the text, MTERMS will assert an allergy even if an allergen was not mentioned.

Allergy Module Structured XML Output Design

The allergy observation module was adapted from the HL7 allergy and intolerance working group model[36] and used to develop an XML schema for representing some of the key elements within an allergy observation (e.g., type of allergy, allergen, reaction). A simplified example of our XML representation from the NLP tool is shown in Figure 3.

Figure 2: XML representation of an allergy to Augmentin (the free-text input was mentioned in the Assessment section of an ED note)

Free-text within Assessment section:
"My impression is this is an allergic urticarial rash possibly from Augmentin"

NLP output:

```
<Allergy SectionText="Assessment ">
   <Allergen RxCUI="151392" SAB="RXNORM" SAB_Code="151392" TTY="BN"
   AllergyIndicator="allergic">augmentin</Allergen>
   <Reaction DESCRIPTIONID="369545012" DESCRIPTIONTYPE="Synonym"
   CONCEPTID="247472004" IndicatesAllergy="True">urticarial rash</Reaction>
</Allergy>
```

Evaluation Methods

Corpus

We evaluated free-text emergency department (ED) notes, which are often the initial point of care for treatment and management of acute allergic reactions. ED notes were randomly selected from years 2011 and 2012 from Brigham and Women's Hospital. A random set of 500 ED notes from 2011 were set aside for training the MTERMS allergy module. The notes were then processed by the NLP tool and the output reviewed by four reviewers (LZ, FG, DS, and JL). We estimated that 15-25% of our sample notes contain allergy information and sample size of 187 was calculated using Stata[37]. We used a larger test set than the calculated sample size, which consisted of 400 randomly selected ED notes from 2012, in order to incorporate a variety of allergy cases and statements. The test set was created by a non-study staff member and all developers were blinded to it. This study was approved by Partners Institutional Review Board (IRB).

Gold Standard Annotation Schema

We used Protégé to create an annotation schema for identifying and classifying allergy information. The annotation schema included the type of allergy (food, drug, contrast media or environmental) as well as "no known allergies" (NKA). We also captured reactions to allergy statements even if the allergen was not specified.

Verification of System-Extracted Allergy Concepts

Using the testing set of 400 randomly selected ED notes, FG and DS manually reviewed and annotated each note using the annotation tool Knowtator[38]. To create the gold standard, the annotations were merged, reconciled and Inter-annotator agreement (IAA) was determined using the Knowtator IAA calculator. We then processed the testing corpus of clinical notes with MTERMS and compared its XML output to our

gold standard. The capability of MTERMS to detect different types of allergens as well as reactions was evaluated.

Our initial evaluation was on the ability of MTERMS to detect the correct reactions identified to a true positive allergen within the clinical text. Second, we evaluated if the reaction was correctly associated with the allergen identified. False positive reactions as a result of a false positive allergen were not considered in this study.

Results

MTERMs overall F-Measure for encoding allergen names and no known allergies was 87.6% with a recall of 91.0% and precision of 84.4%. The inter-annotator-agreement on the testing sample of 400 notes was 98%. Results by allergen type (Table 1) demonstrate that different types have varying system performance. Recall was highest for the drug allergens (92.5%) and lowest for environmental allergens (68.8%). Precision was highest among food, environmental, and NKA. Overall F-measure was highest among NKA and lowest for drug allergens.

Table 1. MTERMS System Performance on Processing ED Notes for Allergen Names and No Known Allergies

Allergen Type	**Total #**	**Recall (%)**	**Precision (%)**	**F-Measure (%)**
Drug	86	92.5	71.1	80.4
Drug Class	34	87.2	87.2	87.2
Environmental	11	69	100	82
Food	10	74	100	85
Contrast Media	6	86	100	92
Subtotal	147	87.2	78.6	82.7
No Known Allergies	70	100	100	100
Total	**217**	**91.0**	**84.4**	**87.6**

For reactions corresponding to a true positive allergen, we assessed the recall, precision, and F-measure (Table 2). We achieved an F-measure of 90% for identifying true reactions in each allergy statement where true allergens were also identified. When taking into account allergen/reaction association (e.g., reaction correctly paired to allergen), the recall increased to 85% but the precision decreased to 58% with a lower F-measure of 69%. Additionally, MTERMS correctly identified two instances of a true positive reaction, "urticaria," where no allergen (true negative) was identified.

Table 2. Reactions for True Positive Allergens*

Reaction	**No association**	**Association****
Total #	11	13
Recall	82%	85%
Precision	100%	58%
F-Measure	90%	69%

* Reactions associated with true positive allergens identified by MTERMS when allergen is present.

** Evaluates the number of associations present and the ability of MTERMS to correctly associate the reaction to the identified allergen (Note that there are cases in which a single reaction is linked to multiple allergens).

Common issues identified within MTERMS are shown in table 3. These included lexical variants that were not accounted for in our lexicon (e.g., lexicon contained NSAIDs but not NSAID and fish-dietary but not fish). These were common across all types of allergens (food, drug, contrast media and environmental). There were also instances where concepts were combined, resulting in only one of the two potential allergens being identified (e.g., MRI and CT contrast – missed MRI contrast).

Reasons for false positive allergens captured by MTERMS have been summarized in Table 4. There was one incidence of a medication falsely tagged as an allergen because it was written in the allergy section

(e.g. Motrin); however, the context suggests it might be a treatment rather than allergy. Outside of the allergy section, medications were most often falsely tagged as allergens when the medication was associated with treatment of a condition (where the condition can also be regarded as an allergic reaction by the NLP tool). For example, Coumadin is often used for the treatment of patients who have atrial fibrillation. There were also several incidences of medications falsely tagged as allergens near unrelated conditions or symptoms (e.g. antibiotics and hyperglycemia).

Table 3. Issues with MTERMS: False Negatives Allergens & Reactions

Issue	**Examples**
Not in lexicon (Drug)	"nitro" | "skin wipe"
Not in lexicon (Class)	"NSAID" | "statin" | "ace inhibitor" | "known drug allergies" | "multiple drug allergies"
Not in lexicon (Food)	"cherries" | "fish"
Not in lexicon (Environmental)	"seasonal allergies" | "bee stings" | "chemical exposure of unclear etiology" | "cold air" | "dry weather"
Not in lexicon (Reaction)	"GI upset" | "tight throat"
Disambiguation between allergy and medication	"Benadryl"
Combination of Terms	"MRI and CT contrast" misses MRI contrast
Pattern not supported	"She is noted to have a heparin allergy where she has itchy rash when she had been injected in the past"

Table 4. Issues with MTERMS: False Positives Allergens

Issue	**Examples**
Medication in allergy section header	"ALLERGIES: None. She has had some musculoskeletal pain relief with the Motrin she has been taking previously"
Medication associated with condition	"Coumadin/atrial fibrillation | antibiotics/fever | albuterol/asthma"
Conditions or symptoms near medications	"Given his comorbidities with his hyperglycemia, I had multiple conversations with strongly recommending observation stay for IV antibiotics" False Positive: Antibiotics | hyperglycemia

Discussion

The MTERMs allergy module achieved an overall F-measure of 87.6% for identifying allergens within ED clinical notes. Recall for medication names (92.5%) was comparable to Zhou et al[33] for free-text notes (90.6%). However precision was lower (71.1% vs. 90.3%) as the tool picked up medications associated with conditions that are not immune-mediated (Table 4). For reactions to true positive allergens, MTERMS had an overall F-measure of 90%. The F-measure dropped to 69% when analyzing if the reaction was correctly clinically paired with the allergen. This drop reflects the increase of false positive associations and lower precision.

Recent efforts on use of NLP to detect allergies have been undertaken by Epstein et al[6] who developed an algorithm to identify ingredient to which patients had allergies in a perioperative information management system. Epstein's system splits multi-allergen entries into multiple entries for processing and mapped allergies to RxNorm. Epstein's system noted a high precision (99.83%) but lower recall (72.82%) when evaluating unique strings from free-text and non-standard drop list allergy entries. The high precision is expected as the corpus essentially corresponded to only the allergy entries. In contrast, MTERMS had a lower precision (84.4%) but higher recall (91.0%), which is what we would expect given the need to find free-text allergy information from clinical notes and disambiguate an allergy from a treatment or home medication. Additionally, MTERMS attempts to encode and associate allergens with reactions.

MTERMS showed improvement in comparison to previous ADE studies[14, 18-21]. On a clinical level, MTERMS is able to identify true reactions when true allergens are also identified (e.g., fentanyl patch/rash, codeine/itching). Additionally, the correct reaction would be associated with the allergen (fentanyl

patch/rash); however, additional reactions (e.g., itching) from other allergens found (e.g., codeine) would also be listed, making it difficult to discern the correct allergen and reaction association. Disambiguation is a challenge because allergies and reactions are not always written in a one-to-one statement: multiple reactions can be listed for a single allergen or a single reaction can be listed for multiple allergens. MTERMS currently associates reactions based on proximity which allows it to correctly associate multiple reactions to a single allergen. However, MTERMS lacks sufficient logic to handle allergy relationships in close proximity or sentences with multiple sets of allergens and reactions. Identifying association indicators and a better understanding of the spatial relationship of allergens to reactions may help to create better association logic. Additionally, there was large variation in how clinicians document the allergy and while it is expected that only pertinent information related to allergies exists in the allergy section, we found some instances where other types of information was entered in the allergies (e.g., Motrin for musculoskeletal pain). Disambiguating false positive allergens associated with a medication used for treatment of a condition but contained in the allergy section or symptoms closely associated with medication (e.g., antibiotics/hyperglycemia) will also require further refinements to the context analyzer.

The lexicon for MTERMS is based on a subset of terms from standard terminologies that were compiled by expert review. MTERMS missed some common singular or pleural concepts (e.g., NSAID, statin, Seasonal Allergies). Common foods (e.g., fish) were also not identified as not all organism concepts from SNOMED CT were incorporated into the lexicon nor included in the PHIN allergy value set from the CDC (which includes the term "fish – dietary"). Misspelling errors continue to be an issue. Further pre-processing capabilities and lexicon development will be necessary to account these findings.

With the persistence of free-text entry, efficient NLP tools and/or machine learning methods are needed to bridge the gap between free-text and standard terminologies. Pick lists can help facilitate the entry of structured data but the preferred and most efficient method of data entry by clinicians is still often free-text. The benefits of structured data entry are clear with respect to drug-allergy and drug-drug checking and other forms of clinical decision support. Challenges ahead include further refining the allergy module of MTERMS and exploring methods of integration of NLP tools within an EHR and clinicians workflow.

One main limitation of our study is the testing sample consisted of ED notes from a single institution's EHR system, thus our results may not be generalize-able to other departments or institutions. For example, primary care physicians may encounter patients with more mild allergic reactions that are not immediately life threatening compared to patients in the ED. We expect documentation by allergy specialists to include much more detailed descriptions of the allergen and reactions, additional allergy elements regarding desensitization status, and certainty whether the allergic reaction was immune mediated through testing. Conclusions on the tool's ability to handle reactions and the challenges of disambiguating between treatments and allergies are also limited. While our sample size was chosen to provide sufficient examples of true allergies, allergic reactions may not have been well represented. We observed even fewer incidences of a true positive allergic reaction where the allergen was unknown or absent. Further analysis of PEAR, a large allergy repository, may improve our contextual understanding the associations between allergens and their reactions.

Conclusion

We present our preliminary findings for an allergy module for the Medical Text Extraction, Reasoning, and Mapping System (MTERMS) natural language processing (NLP) tool, and evaluated it's performance using free-text ED clinical notes. We found it is feasible to extract and encode allergy information from clinical notes using standard terminologies and contextual analysis.

Acknowledgements

This study was funded by the Agency for HealthCare Research and Quality (AHRQ) grant 1R01HS022728-01.

.

References

1. Gaeta TJ, Clark S, Pelletier AJ, *et al.* National study of US emergency department visits for acute allergic reactions, 1993 to 2004. Annals of allergy, asthma & immunology : official publication of the American College of Allergy, Asthma, & Immunology. 2007;98:360-5.
2. Shehab N, Patel PR, Srinivasan A, *et al.* Emergency department visits for antibiotic-associated adverse events. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 2008;47:735-43.
3. Clark S, Espinola J, Rudders SA, *et al.* Frequency of US emergency department visits for food-related acute allergic reactions. The Journal of allergy and clinical immunology. 2011;127:682-3.
4. Bates DW, Cullen DJ, Laird N, *et al.* Incidence of adverse drug events and potential adverse drug events. Implications for prevention. ADE Prevention Study Group. Jama. 1995;274:29-34.
5. Gandhi TK, Burstin HR, Cook EF, *et al.* Drug complications in outpatients. Journal of General Internal Medicine. 2000;15:149-54.
6. Epstein RH, St Jacques P, Stockin M, *et al.* Automated identification of drug and food allergies entered using non-standard terminology. Journal of the American Medical Informatics Association. 2013;20:962-8.
7. Meystre SM, Haug PJ. Comparing natural language processing tools to extract medical problems from narrative text. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2005:525-9.
8. Xu H, Stenner SP, Doan S, *et al.* MedEx: a medication information extraction system for clinical narratives. Journal of the American Medical Informatics Association. 2010;17:19-24.
9. Hazlehurst B, Mullooly J, Naleway A, *et al.* Detecting possible vaccination reactions in clinical notes. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2005:306-10.
10. Haerian K, Salmasian H, Friedman C. Methods for identifying suicide or suicidal ideation in EHRs. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2012;2012:1244-53.
11. Savova GK, Ogren PV, Duffy PH, *et al.* Mayo clinic NLP system for patient smoking status identification. Journal of the American Medical Informatics Association : JAMIA. 2008;15:25-8.
12. Bates DW, Evans RS, Murff H, *et al.* Detecting adverse events using information technology. J Am Med Inform Assoc. 2003;10:115-28.
13. Savova GK, Masanz JJ, Ogren PV, *et al.* Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010;17:507-13.
14. Cao H, Stetson P, Hripcsak G. Assessing explicit error reporting in the narrative electronic medical record using keyword searching. Journal of Biomedical Informatics. 2003;36:99-105.
15. Melton GB, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. J Am Med Inform Assoc. 2005;12:448-57.
16. Friedman C. A broad-coverage natural language processing system. Proceedings / AMIA Annual Symposium. 2000:270-4.
17. Friedman C, Alderson PO, Austin JH, *et al.* A general natural-language text processor for clinical radiology. J Am Med Inform Assoc. 1994;1:161-74.
18. Murff HJ, Forster AJ, Peterson JF, *et al.* Electronically screening discharge summaries for adverse medical events. J Am Med Inform Assoc. 2003;10:339-50.
19. Forster AJ, Murff HJ, Peterson JF, *et al.* Adverse drug events occurring following hospital discharge. J Gen Intern Med. 2005;20:317-23.
20. Honigman B, Lee J, Rothschild J, *et al.* Using computerized data to identify adverse drug events in outpatients. J Am Med Inform Assoc. 2001;8:254-66.
21. Field TS, Gurwitz JH, Harrold LR, *et al.* Strategies for detecting adverse drug events among older persons in the ambulatory setting. J Am Med Inform Assoc. 2004;11:492-8.
22. Skentzos S, Shubina M, Plutzky J, *et al.* Structured vs. unstructured: factors affecting adverse drug reaction documentation in an EMR repository. AMIA Annual Symposium proceedings / AMIA Symposium. 2011:1270-9.
23. Kuperman GJ, Marston E, Paterno M, *et al.* Creating an enterprise-wide allergy repository at Partners HealthCare System. AMIA Annu Symp Proc. 2003:376-80.

24. Goss FR, Zhou L, Plasek JM, *et al.* Evaluating standard terminologies for encoding allergy information. J Am Med Inform Assoc. 2013;20:969-79.
25. SNOMED-CT (Systematized Nomenclature of Medicine-Clinical Terms). [cited 2011 Oct 28th]. Available from: http://www.ihtsdo.org/ snomed- ct/.
26. RxNorm. [cited 2011 Aug 30]. Available from: http://www.nlm.nih.gov/research/umls/rxnorm/.
27. NDF-RT. National Drug File – Reference Terminology (NDF-RT) Documentation: U.S. Department of Veterans Affairs, Veterans Health Administration; 2010 [cited 2011 Sept 5]. Available from: http://evs.nci.nih.gov/ftp1/NDF-RT/NDF-RT Documentation.pdf.
28. Substance Registration System-Unique Ingredient Identifier (UNII). [cited 2011 Oct 27th]. Available from: http://www.fda.gov/ForIndustry/DataStandards/SubstanceRegistrationSystem-UniqueIngredientIdentifierUNII/default.htm.
29. Health Information Technology Standards Panel C80. HITSP/C80 Clinical Document and Message Terminology Component v 2.0, Section 2.2.3.3.9 "Medication Drug Class" and Section 2.2.3.1.1 "Problem" [updated January 25 2010; cited 2011 Sept 6]. Available from: http://www.hitsp.org/ConstructSet_Details.aspx?&PrefixAlpha=4&PrefixNumeric=80.
30. Health Information Technology Standards Panel C83. HITSP/C83 CDA Content Modules Component v 2.0, 2.2.2.6 Allergy/Drug Sensitivity [updated January 25 2010; cited 2011 Sept 6]. Available from: http://www.hitsp.org/ConstructSet_Details.aspx?&PrefixAlpha=4&PrefixNumeric=83.
31. National Council for Prescription Drug Programs. Interoperable Medication Allergy Vocabulary Recommendations 2011 [updated May 24, 2011; cited 2011 Sept 5]. Available from: healthit.hhs.gov/...pt/.../ncpdp-recommendation-06-24-11.pdf.
32. PHIN VOCABULARY ACCESS AND DISTRIBUTION SYSTEM (VADS). Allergy/Adverse Event Type Value Set. 07/31/2013 (revised) ed: Centers for Disease Control.
33. Zhou L, Plasek JM, Mahoney LM, *et al.* Using Medical Text Extraction, Reasoning and Mapping System (MTERMS) to process medication information in outpatient clinical notes. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2011;2011:1639-48.
34. Zhou L, Plasek JM, Mahoney LM, *et al.* Mapping Partners Master Drug Dictionary to RxNorm using an NLP-based approach. Journal of Biomedical Informatics. 2011.
35. U.S. Department of Health and Human Services CfMMS. International Classification of Diseases, 10th Edition, Clinical Modification. Centers for Medicare & Medicaid Services; 2012.
36. HL7 Working Group. Allergy and Intolerance 2012 [updated March 20, 2012]. Available from: http://wiki.hl7.org/index.php?title=Allergy_%26_Intolerance.
37. StataCorp. Stata Statistical Software. 12 ed. College Station, TX: StataCorp LP; 2011.
38. Ogren PV. Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. New York, New York: Association for Computational Linguistics; 2006.

Extracting Concepts Related to Homelessness from the Free Text of VA Electronic Medical Records

Adi V. Gundlapalli, MD, PhD, MS[1,2] , Marjorie E. Carter, MSPH[1,2], Guy Divita, MS[1,2], Shuying Shen, MStat[1,2], Miland Palmer, MPH[2], Brett South, MS[1,2], B.S. Begum Durgahee, MS[1,2], Andrew Redd, PhD[1,2], Matthew Samore, MD[1,2]
[1]VA Salt Lake City Health Care System and [2]University of Utah School of Medicine, Salt Lake City, UT

Abstract

Mining the free text of electronic medical records (EMR) using natural language processing (NLP) is an effective method of extracting information not always captured in administrative data. We sought to determine if concepts related to homelessness, a non-medical condition, were amenable to extraction from the EMR of Veterans Affairs (VA) medical records. As there were no off-the-shelf products, a lexicon of terms related to homelessness was created. A corpus of free text documents from outpatient encounters was reviewed to create the reference standard for NLP training and testing. V3NLP Framework was used to detect instances of lexical terms and was compared to the reference standard. With a positive predictive value of 77% for extracting relevant concepts, this study demonstrates the feasibility of extracting positively asserted concepts related to homelessness from the free text of medical records.

Introduction

Homelessness, especially among Veterans, is a matter of national importance to the United States. The US Department of Veterans Affairs (VA) has committed to ending homelessness among Veterans[1]. Providing appropriate services for those who are experiencing homeless and, of equal importance, early interventions to those at risk are key components of this initiative. An essential element of preventing homelessness among Veterans is a systematic approach to identifying those at risk.

Current methods of identification of Veterans with any of these risk factors from within the VA consist of mining administrative databases for ICD-9-CM codes associated with these diagnoses [2, 3]. The reliability and validity of using administrative data is domain specific and has been shown to be useful in several domains [4-8]. However, this method is unlikely to be either timely or complete enough to be useful for planning interventions. Additionally, many factors of a social and behavioral nature are not captured completely by ICD-9-CM codes and so these may offer only a limited view of a patient's needs or risk factors.

References to risk factors as well as evidence for homelessness are often found only in the free text of medical records written by VA providers and possibly precede the formal identification of Veterans as being homeless[9]. Homelessness also serves as an ideal use case for identifying a wide range of psychosocial risk factors that may be documented in the free text during visits to the health care system [10]. The value of these text data has been shown in several clinical and biomedical domains including bio-surveillance, adverse event detection, and quality improvement [11-14].

Mining the free text of electronic medical records requires informatics-based methods such as natural language processing (NLP) to reliably extract appropriate and relevant information of interest, often referred to as concepts [12, 14]. These concepts can be single words, phrases, or larger spans of text. In general, NLP systems function by parsing sentences, identifying words or phrases, and mapping those to standardized vocabularies or ontologies such as the Unified Medical Language System (UMLS) Metathesaurus. In the absence of a formal ontology or in a situation where it is not known how domain-specific concepts are represented in clinical documents, an intermediate step would be to develop a lexicon or vocabulary of relevant concepts in the domain of interest. The lexicon could then be used as a dictionary in an NLP system for identification and extraction of appropriate concepts.

With no existing literature, reference standard, or lexicon to refer to in the specialized domain of homelessness, the goal of this study was to develop an NLP pipeline to extract positively asserted concepts related to homelessness

from the free text of VA medical records. The primary objective was to describe the positive predictive value (PPV) of the extraction process. We first describe the development of a lexicon of concepts related to homelessness through an iterative process. This lexicon was then used in an NLP pipeline to extract the concepts of interest. The performance of the extraction process was evaluated against a reference standard of free text clinical documents that were classified at the document level by human review to have evidence of homelessness along with a control group of documents with no evidence of homelessness. The prevalence of different categories of concepts in this reference standard is described using the NLP pipeline.

Methods

Setting and data sources

This study was performed using Veterans Information and Computing Infrastructure (VINCI), a central repository of VA-wide databases available for research including administrative, pharmacy and laboratory data, and free-text clinical documents [2]. For this study, a cohort of Veterans was identified who had at least one encounter within the VA health care system during the calendar year 2009 in two of four VA national regions. A corpus of free text clinical documents from outpatient encounters for this cohort of Veterans was extracted from VINCI databases and made available for this concept extraction project.

Developing a lexicon for concepts related to homelessness

A review of available concepts and variant lexical terms covering risk factors for homelessness and homelessness status was determined to be sparse and incomplete within available and frequently used controlled clinical vocabularies. It was determined that the traditional NLP extraction using UMLS based terminology needed to be augmented with additional concepts and terms that are missing.

There were several stages involved in lexicon development. The initial list of terms and phrases was constructed from a literature review for known risk factors for homelessness in general and Veterans in particular [15]. Next, a think-a-loud session was held, where the members of the research team, many of whom have experience with health care for the homeless, reviewed charts from known homeless Veterans and enriched the lexicon with concepts identified from the free text notes written by various VA providers including medical, mental health and social work providers.

The lexicon was refined using an iterative process. A panel of domain experts reviewed documents pre-annotated by NLP to determine if relevant concepts were identified, make modifications to annotations, add missing annotations, or reject annotations found to be incorrect or irrelevant. Concepts found within the notes that were relevant but not currently in the lexicon were added to it. An additional 154 variant forms of the terms were added through the use of NLM's lexical variant generation tools[16] coupled with additional human curation. The final list was adjudicated by a group of clinicians, social workers as well as homeless shelter providers to ensure generalizability in both the VA and external settings. A further refinement of the lexicon was performed by manual review of the documents for false negatives after processing by the NLP pipeline.

The resulting set of 356 terms were categorized into eight broad categories: direct evidence, "doubling up", mentions of mental health diagnoses, behavioral factors, social stressors, sexual and other trauma, medical comorbidities, and other risk factors.

Establishing a human-reviewed reference standard of text documents for homelessness

The premise for choosing text documents was based on discussions with VA clinicians and service providers who care for Veterans experiencing homelessness. The rationale was that barring a self-declaration of homeless status by the Veteran, the most direct reference to a Veteran experiencing homelessness is likely to be in the free text notes of VA providers.

From the text documents generated from VA facility encounters during the calendar year 2009, a random sample of 500 documents was extracted without regard to VA facility or any Veteran characteristic. Another random 500 documents were extracted with the word 'homeless' in the note title. This step was taken to enrich the text document

corpus for homelessness as a random sample of documents is likely to have a low prevalence of homelessness and concepts of interest.

Using the official US Government definition for homelessness [17], a written guideline was developed for human review of text documents. Reviewers were instructed to read a document and then classify the document as either (1) having *positive* documentation for evidence of current or past homelessness or risk factors associated with homelessness (ever homeless/at risk) or (2) *no* evidence of homelessness. Three reviewers were recruited to perform the classification. As part of the training to establish and achieve high agreement, each reviewer was instructed to classify 4 sets of 20 documents (total 80) with review and discussion among reviewers of concordantly and discordantly classified documents. Once the set of 80 'human training' documents were completed, reviewers were assigned sets of 20 documents till the corpus was exhausted (total corpus of 862 documents). The classification task was performed using a general purpose annotation tool called eHOST, in which documents were displayed in human-readable format similar to the VA electronic medical record [18]. The file format for the documents was based on Knowtator, a widely used annotation tool [19].

Researching available electronic resources to map concepts related to homelessness

In an effort to map the terms in our lexicon to existing ontologies, two steps were taken. First, a search for concepts related to homelessness was conducted using the NCI Metathesaurus browser [20]. This resource is described as a wide-ranging biomedical terminology database that covers most terminologies used by NCI for clinical care, translational and basic research, and public information and administrative activities. It is an extensive dictionary of concepts from over 75 sources including the Unified Medical Language System (UMLS) Metathesaurus.

In addition, an effort was made to map the lexical concepts to established ontologies by utilizing BioPortal (http://bioportal.bioontology.org) to ensure uniformity and to avoid duplication of efforts in converting unstructured free text to structured data. BioPortal is an open source Web repository enabling access to terms from about 372 biomedical ontologies and data sources [21, 22]. However, in addition to clinically-related ontologies, BioPortal includes genomic-associated ontologies, such as cell line ontology, as well as other organism ontologies, such as mouse genome. The 372 ontologies as of year 2013 were manually curated in order to limit our lexical search to clinically-related data sources. A total of 98 ontologies and terminologies were considered for concepts mapping and two searches were performed on the lexicon created for this project: an exact match term search and an "at least one" term search (wildcard search strategy). The results of both search types were retrieved and manually curated by three people, with one adjudicating the review results of the other two. UMLS Concept identifiers were also generated for each term in the lexicon via term to concept mapping tools [23]and human judgment for those that did not map. UMLS concept ids are provided as attributes for the 84 terms that were mapped. These efforts provide an additional semantic locality entry point that could be used for classification from different perspectives.

Natural language processing (NLP) pipeline: V3NLP

The basic premise of an NLP pipeline is to take unstructured free text and convert the text to machine-readable 'structured' elements. The V3NLP Framework, used for this project, is a UIMA [24] based set of tools, annotation label guidelines, annotators, readers and writers designed to aid VINCI NLP developers to build out applications. The V3NLP Framework evolved initially from HITEx [25], CTAKEs [26], and MetaMap[23], which are NLP systems that have been applied across a wide variety of use cases.

The pipeline created for this task involved term lookup utilizing the lexicon developed for this project along with negation detection. Additional modules were developed and employed for this task to address boilerplated text such as Slot: Value, check boxes, and question structures. The slot:value pipeline component and the question pipeline component were employed to identify concepts that fall within these structures so that the concept assertion status is updated [27]. Concept assertion semantics are different for each of these boilerplated structures. For example, checkbox structures of the form *Homeless: Y N* was observed within Homelessness Survey records. It is necessary to recognize that this is a check box structure, and that the assertion status for the *homelessness* concept rests on the dependent content, to the left of the delimiter. If only *Y* appears, or any positive variant, the concept is asserted. Any other dependent content value would cause the concept to be tagged with negated. A tail end module was created to turn the homelessness terms into annotations in 8 named categories as described above.

Error analysis for training NLP pipeline

False positive error analyses were performed by two human reviewers on a subset of 50 documents that were classified by the human reference standard as "positive" evidence of homelessness and 50 documents with "no" evidence of homelessness. The task was to review all the concepts extracted by the NLP pipeline (true and false positives) and classify the concept as either positively asserted or not, thus establishing the true positive rate which is the positive predictive value, PPV). This was performed using eHost and reviewing text documents highlighted with concepts identified by the NLP pipeline [18]. The false positive analysis was used to identify problem areas in the text such as templates (such as question/answer formats) and refine the NLP pipeline in an iterative manner. As an initial review of false positives were noted to be due to negation and boilerplated text such as check boxes and questions, modules to address these issues were added to the NLP pipeline (Negex for negation and slot:value and question for boilerplated text).

Further, false negative analyses were performed by reviewing the text of the documents to identify additional concepts related to homelessness that were not already identified by the NLP algorithm. These words, terms or phrases were then included in the lexicon to be used in the next iteration of the NLP algorithm.

Positive predictive value of the NLP pipeline for extracting positively asserted concepts

An agile review of 50 documents from the homeless and non-homeless category was used to improve the overall PPV of concept extraction in an iterative manner. The final, refined iteration of the NLP pipeline was used to determine the instances of positively asserted concepts in eight categories in the 862 documents of the reference standard. After excluding the 100 documents used above for the error analysis and training of the pipeline, the remainder 762 documents were reviewed by two human reviewers to determine the true positivity or PPV of the concepts identified by the NLP pipeline.

The rationale for determining the PPV or precision alone as opposed to precision and recall (sensitivity) as is often reported in the informatics literature is due to the low prevalence of concepts of interest and the resulting high number of documents that are expected to be negative with regard to concepts.

Data extraction and analyses was performed using SAS (Version 9, SAS Institute Inc., Cary, NC) and R software (Version: 2.15.0, The R Foundation for Statistical Computing, Vienna, Austria). This study was approved by the Institutional Review Board of the University of Utah and Research Review Committee of the VA Salt Lake City Health Care System. The research protocol was granted waiver of authorization and waiver of consent to access existing electronic medical records in VA research databases.

Results

Lexicon of Concepts Related to Homelessness

The final version of the manually created lexicon contained 202 high-level psychosocial and homelessness related concepts, divided in to the eight categories already mentioned. Table 1 lists the categories and examples of the terms and concepts included in each one.

Reference standard of free text documents for concept extraction

Starting with a corpus of 1000 documents, our annotators reviewed a total of 862 documents for the reference standard after removing 138 documents due to training and for logistic reasons. Of these, 424 were classified as having evidence of homelessness at the document level (Homeless) and 438 as having no evidence of homelessness (Not Homeless). Reviewers had 98% overall inter-rater agreement in classifying the documents.

The top 5 document types in terms of frequency as determined by the note title in the "Homeless" group were: HCHV Healthcare for Homeless Veterans, Homeless Program Initial Assessment (X), Homeless Program/OPC/SOAP/Social Work, Homeless Program Intake, and Healthcare for Homeless Veterans. The top 5 most frequent document types in the "Not Homeless" group were: Addendum, Discharge Summary, Primary Care, Ambulatory Outpatient Care Note, and Primary Care Clinic Notes.

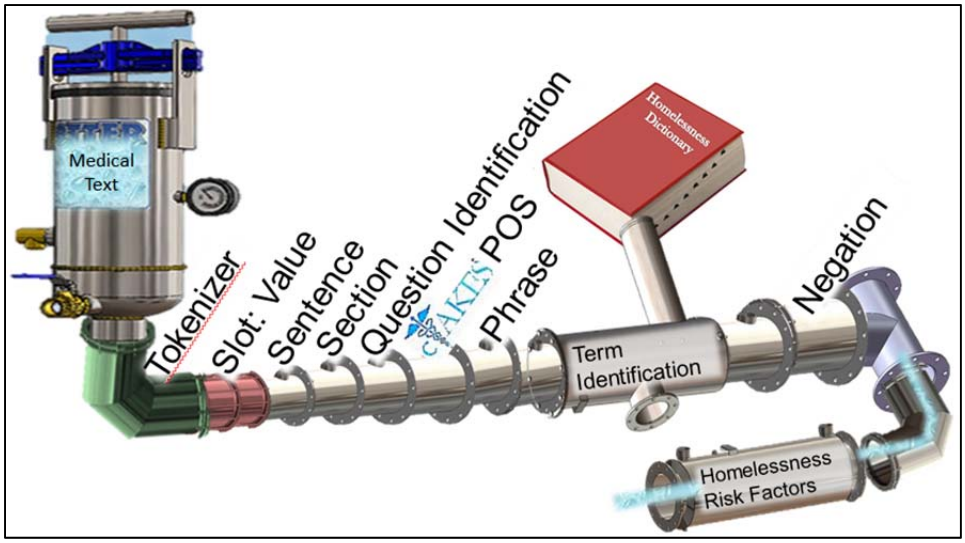


Figure 1: Homelessness psychosocial risk factors NLP pipeline

Table 1. Lexicon concept categories and examples of terms.

Concept Categories	Examples
Behavioral Health	Addiction, alcohol/drug abuse or dependence, detox, DUI, pathological gambling, violent behavior
Direct Evidence	Homeless, living on streets, sleeping outdoors, lack of housing, V60.0
Doubling up	Doubled up, couch surfing, lives with parents/significant other/sibling, crashing at friend's house
Medical Comorbidities	Hepatitis C, frostbite, HIV, gangrene
Mental Health	Axis II, poor coping skills, social isolation, paranoia, schizophrenia, depression, anger
Other Risk Factors	Emergency medical services, at risk ethnic or racial groups, no ID, needs ID
Sexual & Other Trauma	Sexual abuse/trauma, childhood abuse/trauma, domestic violence, military sexual trauma (MST)
Social Stressors	No or poor family support, recent divorce, death of close family member, job loss, legal issues

UMLS and BioPortal searches

The search for 'homeless' and 'homelessness' on the NCI Metathesaurus revealed one unique concept of homelessness with several synonyms (concept unique identifier, CUI C0237154). A review of the relationships of this concept to others in the form or parent, child and sibling relationships (representing hierarchies of relationships between associated concepts) reveals mapping to several risk factors related to homelessness. For example, parent concepts include social problems; child concepts include temporary shelter arrangements; sibling concepts include social behavior and substance abuse disorders. The challenges in directly applying NLP algorithms that map free text to unique concepts in the Metathesaurus is the problem of one term mapping to several different related concepts and the lack of knowledge of how these concepts are represented in the electronic medical record by diverse types of providers in a healthcare system.

The Bioportal search resulted into 648 terms or phrases that were mapped to the initial 201 high-level concepts. This included similar terms from mostly SNOMED, LOINC and Medical Subjects Headings. The results were manually reviewed by two reviewers and ultimately, 185 concept unique identifiers (CUIs) that are psychosocial and homelessness related were retrieved. However, about 62% of the initial concepts did not result into any mapping from Bioportal.

Prevalence of concepts in document corpus

The iterative development of the NLP pipeline based on an agile review of false positives in 50 documents each of homeless and non-homeless resulted in improvement of the positive predictive value (PPV) of concept extraction. (Table 2). While several iterations resulted in improvement in PPV of extracted concepts, some iterations resulted in degradation of PPV in certain concept categories. Table 2 shows three representative steps of the iterative process with corresponding numbers of hits from NLP and associated PPV.

The final iteration of the NLP pipeline (version 10) was then applied to the entire reference standard corpus (Table 3). A total of 4251 concepts were identified from the 862 documents. Among the documents with 'homeless' in the note title, 403 out of 417 (97%) contained at least one concept. Based on a human review of the false positive concepts, the overall positive predictive value (PPV) for extracting homeless-related concepts from this set of documents was 76%. Among the random sample of documents, only 150 out of 455 (33%) contained at least one concept, with an overall PPV of 77%.

Table 2. True positive (Positive Predictive Value) during NLP pipeline iterations based on error analysis

		Documents with evidence of homelessness (N=50 documents)				**Documents with no evidence of homelessness (N=50 documents)**			
		Concept Categories Number of True + False Positives (PPV%)				Concept Categories Number of True + False Positives (PPV%)			
	Error Analysis Comments	Direct	Mental Health And Behavioral	Social Stressors	All others	Direct	Mental Health And Behavioral	Social Stressors	All others
NLP Version 6 Key word; Negation with phrasal span	False positives noted in templated questions; acceptable PPV	1323 (85)	881 (84)	558 (87)	57 (72)	80 (73)	109 (85)	46 (80)	20 (65)
NLP Version 8 Templated question/answer sections ignored	Decrease in overall positivity due to ignoring template question/answer; decrease in PPV	850 (72)	344 (72)	208 (82)	16 (63)	36 (44)	50 (68)	20 (70)	10 (40)
NLP Version 10 Recognize template question/answer; also template with colon (:) and[]	total all 3690	1557 (76)	1005 (61)	681 (91)	66 (79)	14 (71)	239 (80)	79 (95)	49 (69)

In addition, the "homelessness" documents had an average of 8 concepts per document, as compared to 2 concepts per document in the random sample. Table 2 contains additional information related to the prevalence of concepts, including concepts per category, PPV by category, and most frequent concepts from each category.

False positive analysis

The human review of documents for false positive and false negative analyses was conducted using eHOST (Figure 2).
The most common reason for false positivity was that the concept was part of a templated question, such as "How long have you been *homeless*?' In addition, the context of the word or alternate meanings of a word or abbreviation lead to several false positives. For example, the abbreviation SA is often used to refer to substance abuse and was included in the lexicon. However, in the context of a medication list, SA TAB refers to sustained action. The phrase *rule out* is commonly used in medical records to literally indicate that the providers were trying to rule out a particular condition or diagnosis. This had to be added as a term to insure the negation module picked it up as one unit within one phrase to appropriately negate the rest of the phrase.

Table 3. Prevalence of concepts related to homelessness in reference standard document corpus (Total N=862 documents, Total positively asserted concepts = 4251)

	Documents with evidence of Homelessness (N=403 documents)* Total concepts = 3309 Average 8 concepts per document, range 1 to 34, median 8			Documents with no evidence of Homelessness (N=150 documents)** Total concepts = 320 Average 2 concepts per document, range 1 to 13, median 1		
	Number of Concepts (Positive Predictive Value %)	**Most Frequent Concepts**	**Most Frequent Document Types**	**Number of Concepts (Positive Predictive Value %)**	**Most Frequent Concepts**	**Most Frequent Document Types**
Direct Evidence	1557 (76)	shelter, homelessness, hchv, homeless program, rehabilitation	homeless, social work, mental health program notes	**20 (70)**	shelter, rehabilitation, halfway house, housing program	psychiatry nursing assessment, mental health
Mental Health and Behavioral Factors	1016 (62)	detox, heroin, drug use, etoh, alcohol abuse	homeless, social work, discharge summary and mental health program notes	**195 (77)**	etoh, alcohol use, withdrawal, alcohol abuse, substance abuse,	discharge summaries, general mental health note, follow-up primary care
Social Stressors	681 (91)	divorced, unemployed, separated, prison, jail	homeless, social work, discharge summary and mental health program	**66 (95)**	divorced, unemployed, lives alone, prison, jail	mental health, , discharge summaries, history & physical
Other	55 (75)	trauma, hep c, mst	homeless, medical progress and social work program	**39 (67)**	mst, trauma, hep c	primary care, medical clinic, mental health note

* **14 of 417 (3%) documents with evidence of homelessness had 0 concepts**

** **305 of 455 (67%) documents with no evidence of homelessness had 0 concepts**

hchv = health care for homeless veterans; etoh = alcohol; hep c = hepatitis c; mst = military sexual trauma

False negative analysis

In the first round of review, the most common terms that represented concepts related to homelessness that were found in the free text and were missing from the lexicon were terms related to alcohol use. For example, a template used in screening for alcohol use, the phrase used was 'how many drinks containing alcohol do you consume each day?'; similarly several creative variations on the theme of heavy drinking were noted in the free text including the terms 'problem drinker'. Other examples of missed concepts were due to misspelling and the use of different cases (capital vs. lower case) in a non-standard fashion. For example, typically the usage for the Housing and Urban Development, VA Assisted Housing is represented as HUD-VASH; in some instances, the provider used a non-standard mix-up of upper and lower case letters such as Hud-Vash which was not recognized by the NLP program as a variant (even though NLP programs have a case insensitive algorithm built in).

Discussion

Using the VA electronic record and informatics methods, this study demonstrates the feasibility of extracting positively asserted concepts related to homelessness from the free text of medical records. Sub-optimal mapping of concepts to existing ontologies demonstrated the necessity for developing a customized lexicon for concepts related to homelessness. Homelessness may be considered a 'non-medical' condition and as such this represents an interesting use of NLP to extract non-medical concepts from medical records. It is important to note that the established risk factors for homelessness (as denoted by the different categories in Table 1) are very much part of the medical domain and are expected to be found in medical records.

The potential applications of the lexicon and NLP algorithm are to study the prevalence of these concepts among those who have evidence of homelessness versus those who are not. These analyses may also play a role in

determining associations of concepts with the onset of homelessness and in the ideal situation, provide a means of identifying early warning indicators of risk of homelessness, prior to the formal 'diagnoses' through ICD-9-CM codes for homelessness or known risk factors. This could be achieved by applying the algorithm on longitudinal electronic medical records of Veterans. With the use of available ontologies and vocabularies, this work could be extended to other medical domains.

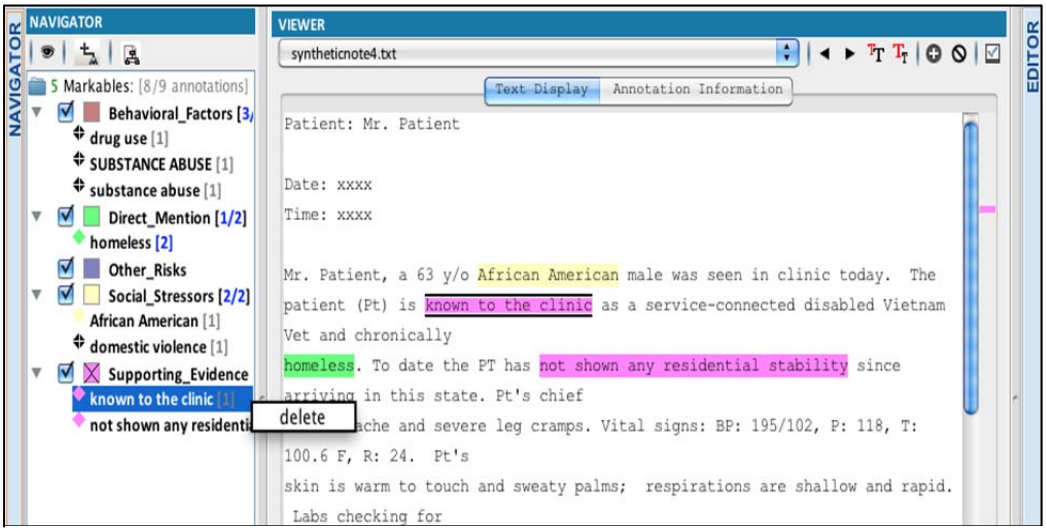


Figure 2. Synthetic record containing mentions of homelessness as displayed in eHost, the annotation tool used.

The issue of templates in VA medical records has been shown to be a major factor in poor performance of information extraction algorithms[27], as demonstrated by the use of a version of this pipeline in a study of extracting psychosocial factors from large electronic medical record corpora [10]. This study demonstrates the feasibility of successfully processing question/answer and other formats of templates to extract positively asserted concepts to increase yield (hit rate of concepts) and PPV. This has broad applicability in the field of NLP as templated medical records are used in many settings to document patient-related information both by providers (forms and checklists in the EMR) and by patients (in-clinic, through kiosks and on-line surveys and questionnaires).

The iterative process by which the lexicon was developed and made available for use by the NLP algorithm reinforces the challenges of 'teaching' computers to read human language. The myriad variations possible in terms of spelling, meaning and abbreviations, the variation noted in text written by different types of providers and regional variations lead one to conclude that this likely represents a sub-language of its own.

We note several limitations. Based on our experience, the lexicon for homelessness concepts is neither exhaustive nor complete. Thus we would consider the current lexicon to be a living document. We did not attempt to add to the lexicon using automated or 'self-learning' methods. It is likely that there are variations in use of concepts related to homelessness based on sub-language differences among various clinical providers or based on geography (by state or region of the US). Thus, it would be important to study such variations to look at iso-semantic concepts. A more detailed mapping of our lexical terms to available vocabularies and ontologies is necessary and ongoing as it is desirable to convert all NLP-derived phenotypes to structured data. We applied the NLP pipeline to a small subset of VA records in this study. A version of the pipeline has been applied to a large corpus with acceptable performance [10]. While this pipeline has been developed using VA medical records, the principles are generalizable and adaptable to the EMR of other large health care systems.

Concept assertion status including asserted, negated, hypothetical, historical, and not relating to the patient are ever present sources of failures within the NLP process. NegEX2 [28]was used for this study to determine asserted/negated status. Future iterations will incorporate a refined version of ConTEXT to determine assertion status with more granularity and precision[29].

Concepts of interest extracted from the NLP pipeline are not sufficient to make a homelessness classification for a given patient. On-going work involves using concepts of interest extracted from reference documents as features to train a machine learning predictive model.

Conclusion

Extracting information related to a non-medical condition from the free text of medical records can be a challenge. In this use case involving concepts related to homelessness, it required the development of a specialized lexicon, as the concepts of interest were poorly covered by existing ontologies. This method can be applied to other areas of interest and relevance while mining the EMR.

Acknowledgements

This work is supported by VA HSR&D Merit Review Award # HIR 10-002 (PI: AVG). We would like to express our gratitude to the administration and staff of the VA Informatics and Computing Infrastructure (VINCI) for their support of our project. The project benefited immensely from an active advisory role by the VA SLC Health Care System's homeless service coordinator, Mr. Aldo Hernandez. We are deeply appreciative of the expertise and active participation in this project by our homeless service community partners in Salt Lake City, Utah: Fourth Street Clinic (Mr. Monte Hanks); The Road Home (Ms. Michelle Flynn and Ms. Michelle Vasquez), Volunteers of America (Ms. Jessica Fleming, Ms. Jamie Jones) and the State of Utah (Ms. Kathleen Moore). We would like to thank our colleagues at the National Center on Homelessness Among Veterans (Drs. Dennis Culhane, Steven Metraux and Jamison Fargo) for their discussions and advice. We gratefully acknowledge the support of our research team members: Mr. Thomas Ginter, Ms. Sarah Craig and Ms. Natalie Kelly. We also acknowledge other resources and facilities provided by the VA Salt Lake City Health Care System.

The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the United States government.

References

1. United States Interagency Council on Homelessness, *Opening Doors: Federal Strategic Plan to Prevent and End Homelessness*, United States Interagency Council on Homelessness, Editor. 2010, United States Interagency Council on Homelessness Washington, DC.
2. US Department of Veterans Affairs. *VA Information Resource Center*. 2012 [cited 2012; Available from: http://www.virec.research.va.gov/DataSourcesName/DataNames.htm.
3. U.S. Department of Veterans Affairs Office of Inspector General, *Homeless Incidence and Risk Factors for Becoming Homeless in Veterans*, VA Office of Inspector General, Editor. 2012, VA Office of Inspector General: Washington DC.
4. Kashner, T.M., *Agreement between administrative files and written medical records: a case of the Department of Veterans Affairs.* Med Care, 1998. **36**(9): p. 1324-36.
5. Schneeweiss, S., et al., *Veteran's affairs hospital discharge databases coded serious bacterial infections accurately.* J Clin Epidemiol, 2007. **60**(4): p. 397-409.
6. Roumie, C.L., et al., *Validation of ICD-9 codes with a high positive predictive value for incident strokes resulting in hospitalization using Medicaid health data.* Pharmacoepidemiol Drug Saf, 2008. **17**(1): p. 20-6.
7. Banerjea, R., et al., *Co-occurring medical and mental illness and substance use disorders among veteran clinic users with spinal cord injury patients with complexities.* Spinal Cord, 2009. **47**(11): p. 789-95.
8. Tracy, L.A., et al., *Predictive ability of positive clinical culture results and International Classification of Diseases, Ninth Revision, to identify and classify noninvasive Staphylococcus aureus infections: a validation study.* Infect Control Hosp Epidemiol, 2010. **31**(7): p. 694-700.
9. Redd, A., et al., *Detecting earlier indicators of homelessness in the free text of medical records.* Stud Health Technol Inform, 2014. **202**: p. 153-6.
10. Gundlapalli, A.V., et al., *Validating a strategy for psychosocial phenotyping using a large corpus of clinical text.* J Am Med Inform Assoc, 2013. **20**(e2): p. e355-64.

11. Meystre, S.M., et al., *Extracting information from textual documents in the electronic health record: a review of recent research.* Yearb Med Inform, 2008: p. 128-44.
12. Chapman, W.W., *Closing the gap between NLP research and clinical practice.* Methods Inf Med, 2010. **49**(4): p. 317-9.
13. Jha, A.K., *The promise of electronic records: around the corner or down the road?* JAMA, 2011. **306**(8): p. 880-1.
14. Nadkarni, P.M., L. Ohno-Machado, and W.W. Chapman, *Natural language processing: an introduction.* J Am Med Inform Assoc, 2011. **18**(5): p. 544-51.
15. Balshem, H., et al., *A Critical Review of the Literature Regarding Homelessness Among Veterans*, in *A Critical Review of the Literature Regarding Homelessness Among Veterans*, US Department of Veterans Affairs, Editor. 2011: Washington (DC).
16. McCray, A.T., S. Srinivasan, and A.C. Browne, *Lexical methods for managing variation in biomedical terminologies.* Proc Annu Symp Comput Appl Med Care, 1994: p. 235-9.
17. U.S. Department of Housing and Urban Development. *Federal Definition of Homelessness*. 2011 [cited 2011 December 3, 2011]; Available from: http://portal.hud.gov/hudportal/HUD?src=/topics/homelessness/definition.
18. South, B., et al. *A Prototype Tool Set to Support Machine-Assisted Annotation.* in *BioNLP 2012*. 2012. Montreal, Canada.
19. Ogren, P.V. *Knowtator: A Protégé plug-in for annotated corpus construction.* . in *Proceedings of the Human Language Technology Conference of the NAACL*. 2006.
20. National Cancer Institute. *NCImetathesaurus*. 2012 [cited 2012; Available from: http://ncim.nci.nih.gov/ncimbrowser/.
21. Noy, N.F., et al., *BioPortal: ontologies and integrated data resources at the click of a mouse.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W170-3.
22. Whetzel, P., et al., *BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications.* . Nucleic Acids Res. , 2011. **Jul** (39(Web Server issue)): p. W541-5. .
23. Aronson, A.R., *Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.* Proc AMIA Symp, 2001: p. 17-21.
24. Ferrucci, D. and A. Lally, *UIMA: an architectural approach to unstructured information processing in the corporate research environment.* Nat. Lang. Eng., 2004. **10**(3-4): p. 327-348.
25. Zeng, Q.T., et al., *Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system.* BMC Med Inform Decis Mak, 2006. **6**: p. 30.
26. Savova, G.K., et al., *Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications.* J Am Med Inform Assoc, 2010. **17**(5): p. 507-13.
27. Divita, G., et al., *Recognizing Questions and Answers in EMR Templates Using Natural Language Processing.* Stud Health Technol Inform, 2014. **202**: p. 149-52.
28. Chapman, W.W., et al., *A simple algorithm for identifying negated findings and diseases in discharge summaries.* J Biomed Inform, 2001. **34**(5): p. 301-10.
29. Chapman, W.W., D. Chu, and J.N. Dowling, *ConText: an algorithm for identifying contextual features from clinical text*, in *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*. 2007, Association for Computational Linguistics: Prague, Czech Republic. p. 81-88.

Integrating Public Data Sets for Analysis of Maternal Airborne Environmental Exposures and Stillbirth

Eric S. Hall, PhD[1,2], Natalia Connolly, PhD[2], David E. Jones, MA MPH[3], and Emily A. DeFranco, DO MS[1,4]
[1]Perinatal Institute, Cincinnati Children's Hospital Medical Center, [2]Biomedical Informatics, Cincinnati Children's Hospital Medical Center, [3]Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center, [4]Maternal-Fetal Medicine, University of Cincinnati
Cincinnati, Ohio

Abstract

Efforts to study relationships between maternal airborne pollutant exposures and poor pregnancy outcomes have been frustrated by data limitations. Our objective was to report the proportion of Ohio women in 2006-2010 experiencing stillbirth whose pregnancy exposure to six criteria airborne pollutants could be approximated by applying a geospatial approach to vital records and Environmental Protection Agency air monitoring data. In addition, we characterized clinical and socio-demographic differences among women who lived within 10 km of monitoring stations compared to women who did not live within proximity of monitoring stations. For women who experienced stillbirth, 10.8% listed a residence within 10 km of each type of monitoring station. Maternal race, education, and marital status were significantly different ($p<0.0001$) comparing those within proximity to monitoring stations to those outside of monitoring range. No significant differences were identified in maternal age, ethnicity, smoking status, hypertension, or diabetes between groups.

Background

Despite a steady decline over the previous two decades, the United States stillbirth rate of 6.0 per 1,000 births in 2006 remained substantially higher than that of other high income countries.[1-3] The national stillbirth rate (defined as fetal deaths occurring at or beyond 20 weeks of gestation) also remained 50% above the Healthy People 2010 goal of 4.1 per 1,000 leading to a revised goal for Healthy People 2020 of 5.6 per 1,000.[4, 5] While, obesity, maternal age, and smoking are significant modifiable risk factors for stillbirth, other clinical, genetic, socioeconomic, and environmental factors are also important contributors to a woman's risk of stillbirth.[6, 7] Although methodological limitations in previously reported studies have yielded mixed and inconclusive results, identifying the effects of maternal exposure to airborne pollutants on fetal development and birth outcomes remains a research priority.[8, 9]

The Environmental Protection Agency (EPA) lists six criteria pollutants representing agents harmful to public health.[10] These include carbon monoxide (CO), lead (Pb), nitrogen dioxide (NO_2), ozone (O_3), particle pollution, and sulfur dioxide (SO_2). Pollutants are monitored at regular intervals by stations at fixed locations, though exposures of individuals residing nearby may be approximated. Exposure estimates rely on linking data representing air quality with clinical and socio-demographic measures of individuals. Previous analyses of associations between the criteria pollutants and stillbirth have been limited in analytic design as a consequence of relevant data residing in incompatible systems.[11-17] Former study limitations include inadequate temporal and geographic granularity and insufficient controlling for socio-demographic risk factors. Additionally, previous efforts have typically focused on single pollutant exposures rather than considering exposures to multiple pollutants or the combined effects of multiple pollutant exposures.

A recent study utilizing a geospatial approach identified an association between stillbirth and maternal exposure to pollutants by linking vital records to data from the EPA Air Quality System.[18] However, this type of analysis may also be limited by proximity of mothers to EPA monitoring stations, particularly when conducting analysis of multiple exposure types. The fixed number of monitoring stations as well as the placement of stations near urban settings may limit the coverage and representativeness of population data obtained from the EPA.

The objective of this Ohio statewide study was to describe the coverage and utility of existing air monitoring data to support investigations of maternal pollutant exposures when integrated with vital statistics. This was achieved by 1) identifying among women who experienced a live or stillbirth, the proportion who resided within proximity to an EPA air monitoring station 2) characterizing clinical and socio-demographic differences among women who lived within range of all types of monitoring stations compared to women who did not live within proximity of any air

monitoring stations. This study builds upon previous efforts in developing maternal-child health data resources by integrating disparate data sets including vital statistics, hospital discharge data, geospatially-based measures from the American Communities Survey, and community-based program data.[19, 20] Integration of air quality data provides an additional dimension for investigating the complex interplay of factors contributing to stillbirth and other perinatal outcomes.

Methods

The geospatial population-based cohort study incorporated data from two primary sources, vital records and air monitoring measures. Vital records including fetal death records as well as live birth records were obtained from the Ohio Department of Health representing 4,622 stillbirths and 751,123 live births occurring in Ohio during 2006-2010. The study set was reduced to singleton births with documented gestational age between 20 and 42 weeks and with documented birth weight of at least 350 grams. Additionally, infants with confirmed chromosomal anomalies were not included in the final study set. Lastly, births occurring within Ohio corresponding to mothers reporting residence outside of the state were excluded resulting in a final study set of 3,090 stillbirths and 707,110 live births (Table 1). Basic demographic and clinical measures including maternal race and ethnicity, age, marital status, level of education, smoking status, and diagnosis of hypertension or diabetes were also obtained from vital records data.

TABLE 1. Study exclusion criteria and final sample

Exclusion Criteria	Stillbirths	Live births
Records of birth events within Ohio, 2006-2010	4,622	751,123
Missing gestational age, gestational sge < 20 weeks, or gestational age > 42 weeks	52	2,910
Birth weight <350 grams	954	366
Chromosomal anomalies	128	685
Multiple gestation	86	26,995
Residence outside of Ohio	113	13,057
Final study set	3,090	707,110

Measures of criteria pollutants were obtained from the EPA representing CO, Pb, NO_2, O_3, SO_2 and fine particulate matter measuring less than 2.5 μm in aerodynamic diameter ($PM_{2.5}$).[21] Data were obtained from Ohio air monitoring stations representing pollutant level measures at time periods ranging from hourly to every few days. Data were obtained for 2005 through 2010 to correspond with the pregnancy timeframes represented by the vital records data set. Each pollutant was measured by a unique set of stations, with a range of six to 59 locations in Ohio. Although some stations measured a single pollutant, other stations obtained multiple measures. In all, measures of pollutant levels for the six criteria pollutants were captured at 126 unique geospatial coordinates across the state. Using all measures captured for each pollutant at each location, monthly average measures were calculated for every calendar month of the study period.

Linkage between vital records and air pollution data was performed using ArcGIS 10.1 (ESRI, Redlands, CA). Station locations were geocoded with a latitude-longitude coordinate pair as were maternal residences reported at the time of birth. When specific address information was missing, residence was approximated using the centroid of the residence zip code. For each live or stillbirth record, the closest monitoring station was identified for each pollutant. Additionally, the distance to the nearest monitoring station of each pollutant type was calculated.

The number of birth events listing maternal residence within five, 10, and 15 km of a monitoring station was calculated for each pollutant. We also determined the set of women residing within proximity to monitoring stations for each of the six criteria pollutants as well as the set of women who were not within the proximity of five, 10, or 15 km of any criteria pollutant monitoring stations. Additional analyses were conducted for the women within proximity to all six monitor types compared to women outside of the coverage range of any monitoring station. Consistent with methods used in previous analyses, proximity for the additional analysis was defined as 10 km.[18, 22]

Statistical analysis and calculations were conducted using SAS version 9.3 (SAS Institute Inc., Cary, NC) software. Bivariate χ2- or *t*-tests analyses were performed to identify demographic and clinical differences between populations within proximity to all six pollutant monitors and those outside the proximity of any monitoring station. Because multiple tests were performed, a Bonferroni-Holm correction was implemented to determine significance.[23]

Next, estimated exposures to each pollutant were calculated for each pregnancy trimester and for the entire duration of pregnancy. The initiation of the pregnancy was estimated by subtracting the clinical estimate of gestational age from the documented date of delivery. The first trimester was defined as weeks 0 through 12, the second trimester as weeks 13 through 27 and the third trimester as weeks 28 through 42. Exposures were estimated by first identifying

the average monitoring station measure for the calendar month corresponding to each pregnancy month. Whole pregnancy and per trimester exposure estimates were calculated using monthly monitor measures for which the mother was pregnant for at least half of the calendar month, or for which the pregnancy trimester spanned at least half of the calendar month. For each pregnancy stage, mean population exposures for stillbirth and live birth outcomes were then calculated using exposure estimate calculations for individual mothers. The Ohio Department of Health and University of Cincinnati Institutional Review Boards approved this study.

Results

For both stillbirth and live birth sets, the mean and median distance from each residence to the nearest station monitoring each of the criteria pollutants is recorded in Table 2. For all six of the criteria pollutants, the average distances between residences and stations were inversely related to the number of monitoring stations.

TABLE 2. Number of monitoring stations for each pollutant and the average distance between stations and maternal residence at the time of live or stillbirth event.

Pollutant	Number of stations	Distance from stillbirth residence to nearest station (km) (mean, median)		Distance from live birth residence to nearest station (km) (mean, median)	
CO	17	36.1,	17.6	40.7,	19.8
Pb	20	24.8,	16.5	29.5,	17.7
NO_2	6	68.1,	49.1	72.3,	66.2
O_3	59	14.4,	14.8	18.3,	9.9
$PM_{2.5}$	57	16.3,	8.6	20.7,	10.0
SO_2	39	20.5,	14.4	24.4,	15.9

Table 3 lists the number of stillbirth and live birth events listing a residence within a 5, 10, and 15 km radius from each monitoring station. The proportion of the 3,090 stillbirths and 707,110 live births covered by each radius is also reported. Additionally, the number and proportion of births with residence within range of all six criteria pollutant monitoring station types is listed. Ozone and $PM_{2.5}$ monitoring stations had the best coverage with over 55% of stillbirths and nearly 50% of live births having a corresponding maternal residence within 10 km from a station. Just 10.8% of the stillbirth and 8.1% of the live birth population listed a residence within 10 km of each of the six monitoring station types. On the other hand, 33.6% of the stillbirth and 39.1% of the live birth population listed a residence that was more than 10 km from any of the six monitoring station types (Figure 1).

TABLE 3. Number and percentage of 3,090 stillbirths and 707,110 live births within range of pollution monitoring stations.

Pollutant	Stillbirths within 5 km		Stillbirths within 10 km		Stillbirths within 15 km		Live births within 5 km		Live births within 10 km		Live births within 15 km	
CO	611	19.8%	1,061	34.3%	1,428	46.2%	93,658	13.2%	200,399	28.3%	288,023	40.7%
Pb	527	17.1%	1,107	35.8%	1,451	47.0%	81,601	11.5%	202,609	28.7%	299,932	42.4%
NO_2	237	7.7%	446	14.4%	574	18.6%	34,487	4.9%	80,334	11.4%	114,228	16.2%
O_3	728	23.6%	1,741	56.3%	2,198	71.1%	137,521	19.4%	350,055	49.5%	474,048	67.0%
$PM_{2.5}$	1,004	32.5%	1,782	57.7%	2,095	67.8%	173,391	24.5%	345,782	48.9%	447,803	63.3%
SO_2	516	16.7%	1,104	35.7%	1,583	51.2%	88,784	12.6%	219,416	31.0%	329,355	46.6%
All	104	3.4%	335	10.8%	505	16.3%	14,632	2.1%	57,271	8.1%	96,676	13.7%
None	1,809	58.5%	1,037	33.6%	738	23.9%	457,085	64.6%	276,543	39.1%	182,622	25.8%

Characteristics of the stillbirth and live birth populations are listed in Table 4. Also in the table, clinical and demographic characteristics are further stratified into categories of women who reported a residence within 10 km of each of the six types of monitoring stations and women whose residence was more than 10 km from any station type. Among women who experienced a stillbirth, race, education, and marital status had significant differences between groups. However, no significant differences were identified in age, ethnicity, smoking status, hypertension, or diabetes between groups. Among the large live birth set, all comparisons yielded significance in comparisons between groups.

Exposures to each of the six pollutants are presented in Table 5 for both stillbirth and live birth data sets. Mean exposures are listed for each pregnancy trimester and for the pregnancy duration representing all estimated exposures having a residence within 10 km of the relevant station type. Only exposures to $PM_{2.5}$ exceeded EPA standards; however the high exposure rates affected both stillbirth and live birth populations.

Discussion

The linkage approach described in this manuscript has potential to enable powerful analyses of pollution effects on stillbirth as well as other pregnancy related outcomes. The final linked study data set joined measures from the EPA and vital record sets providing a data set containing socio-demographic measures at the individual person level along with pollution exposure approximations customized both temporally and spatially to each individual pregnancy. Exposure approximations for each trimester of pregnancy enable additional analyses about the stages of fetal development that may be most susceptible to high levels of pollutant exposure. This approach has already provided insight into the impact of high $PM_{2.5}$ exposure levels on stillbirth. In a previous analysis we found a 42% increase in stillbirth risk associated with increased $PM_{2.5}$ exposure during the third trimester of pregnancy.[24] Further, we estimated that up to 2.2 per 1,000 Ohio stillbirths are potentially attributable to high levels of $PM_{2.5}$ exposure.

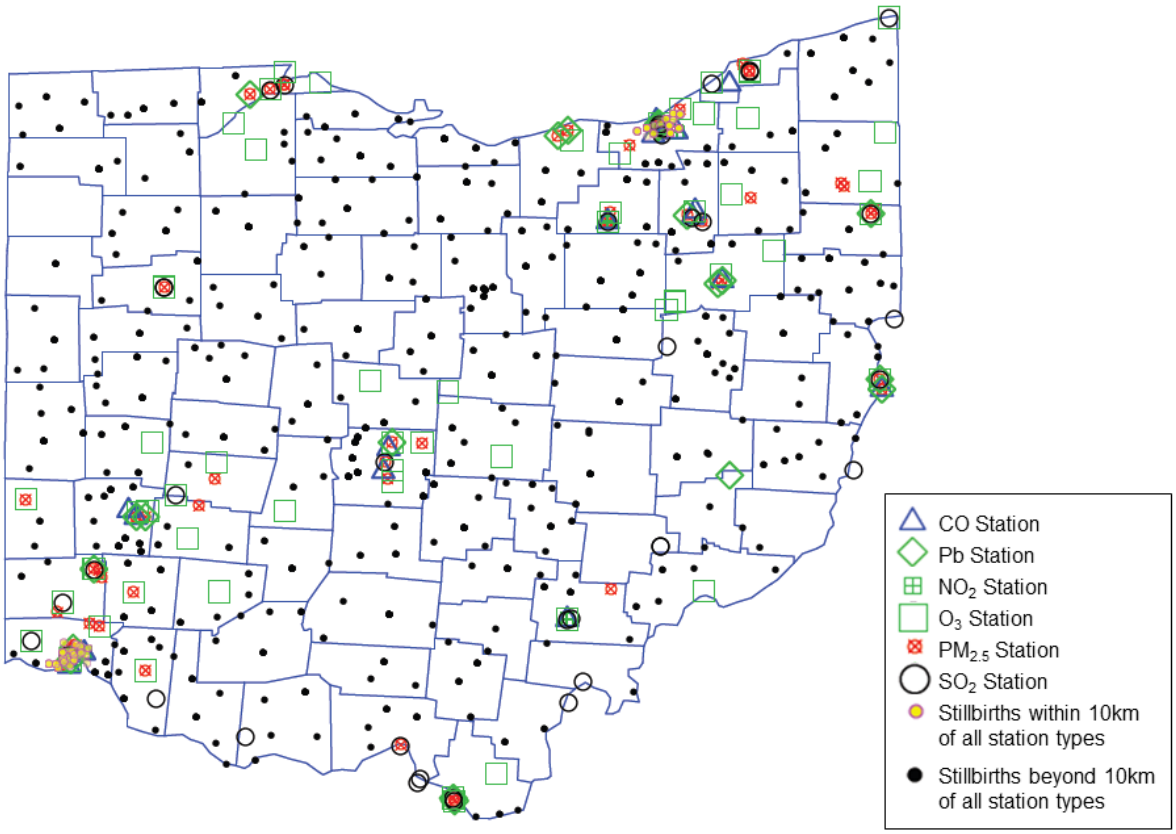


FIGURE 1. Locations of 126 monitoring station locations throughout Ohio along with points indicating stillbirths with residence within 10 km of all six criteria pollutants monitoring stations and with residence beyond 10 km from any monitoring station.

Efforts to use this approach to estimate pregnancy exposures to all six criteria pollutants using a maximum of 10 km distance between residence and monitoring stations would be limited to less than 11% of the Ohio stillbirth and about 8% of the live birth populations. Also the analysis would be restricted to residents of the urban Cincinnati and Cleveland regions. Although this geospatial approach for approximating exposure may be sufficient for single particulate analysis, reliance of residence in close proximity to air monitoring stations is a considerable limitation to the analysis of the effects of multiple pollutants. Having just six monitoring stations in the state of Ohio, NO_2 approximations are the most sparsely represented in the data set. Although it may be difficult to conduct representative analyses of simultaneous exposures to all six criteria pollutants, this geospatial approach could potentially support meaningful analyses of a subset of the criteria pollutants.

TABLE 4. . Clinical and demographic characteristics of the stillbirth and live birth study populations including a comparison of women with a residence within 10 km of all six monitoring station types and women with a residence more than 10 km from every monitoring station type.

	Stillbirths (N=3,090)				Live births (N=707,110)			
	All stillbirths (N=3,090)	Residence near 0 station types (n=1,037)	Residence near all 6 stations types (n=335)	p-value	All live births (N=707,110)	Residence near 0 station types (n=276,543)	Residence near all 6 stations types (n=57,271)	p-value
Age (mean)	27.1	27.1	26.4	0.2	26.7	27.1	25.5	< 0.0001*
Race (%)				< 0.0001*				< 0.0001*
White	66.2	89.7	22.4		79.7	93.9	41.8	
Black	29.6	6.0	74.0		17.2	4.1	55.7	
Other	4.2	4.3	3.6		3.2	2.0	2.5	
Hispanic ethnicity (%)	5.8	4.5	6.6	0.1	4.6	3.4	7.7	< 0.0001*
Maternal education (%)				< 0.0001*				< 0.0001*
No high school degree	20.4	16.8	27.9		17.6	15.1	30.7	
High school degree	40.3	43.6	43.5		26.6	28.4	27.5	
Any college	39.3	39.6	28.6		55.8	56.5	41.8	
Married (%)	48.0	63.8	27.6	< 0.0001*	57.3	65.5	30.3	< 0.0001*
Smoker (%)	23.7	25.1	22.4	0.3	19.7	20.6	19.5	< 0.0001*
Hypertension (%)	9.5	9.3	13.7	0.08	6.2	6.2	6.9	< 0.0001*
Diabetes (%)	7.3	6.9	8.4	0.7	5.8	5.7	5.1	< 0.0001*

*statistical significance

TABLE 5. Calculated trimester and pregnancy exposure estimates of within 10 km of a pollution monitoring station compared to Environmental Protection Agency (EPA) standard.

		Stillbirths					Live births				
Pollutant	EPA Standard[10]	Stillbirths within 10 km	First trimester exposure (mean±sd)	Second trimester exposure (mean±sd)	Third trimester exposure (mean±sd)	Whole pregnancy exposure (mean±sd)	Live births within 10 km	First trimester exposure (mean±sd)	Second trimester exposure (mean±sd)	Third trimester exposure (mean±sd)	Whole pregnancy exposure (mean±sd)
CO (ppm)	9 (8-hour average)	1,061	0.38±0.12	0.37±0.12	0.37±0.13	0.37±0.11	200,399	0.39±0.13	0.39±0.13	0.38±0.14	0.39±0.11
Pb (µg/m3)	0.15 (rolling 3 month average)	1,107	0.1±0.0	0.1±0.0	0.0±0.0	0.1±0.0	202,609	0.1±0.0	0.1±0.0	0.1±0.0	0.1±0.0
NO_2 (ppb)	100 (1-hour average)	446	16.5±2.9	16.5±3.6	16.0±3.0	16.4±2.7	80,334	16.9±3.5	16.6±3.6	16.3±3.6	16.7±3.3
O_3(ppm)	0.075 (8-hour average)	1,741	0.03±0.01	0.03±0.01	0.03±0.01	0.03±0.01	350,055	0.03±0.01	0.03±0.01	0.03±0.01	0.03±0.01
$PM_{2.5}$ (µg/m^3)	12 (annual average)	1,782	13.6±2.8	13.2±2.3	13.3±2.8	13.3±1.8	345,782	13.7±2.9	13.3±2.4	13.0±2.3	13.3±1.8
SO_2 (ppb)	75 (1-hour average)	1,104	3.5±1.9	3.4±2.0	3.5±2.3	3.4±1.8	219,416	3.8±2.1	3.6±2.1	3.5±2.1	3.6±1.9

sd= standard deviation
ppm = parts per million
ppb = parts per billion

In the current analysis, although we found that coverage for estimating exposures to all six criteria pollutants was limited to urban populations, there were no significant difference in the prevalence of important stillbirth risk factors including maternal age, smoking status, hypertension, and diabetes between the rural and urban groups. While the distribution of clinical risk factors was consistent among those in close proximity to all monitoring stations as well as those outside the proximity of any monitoring station, there were significant differences in the socio-demographic makeup of the two groups. Women residing near all six types of monitoring stations were more likely to be black, unmarried, and to have attained a lower level of education than women listing a residence beyond the range of any monitoring station. Because many of these characteristics are associated with the stillbirth outcome, any study using our data management approach would need to make proper adjustment for these coexisting risk factors when conducting analyses and in reporting results.

The described approach has generalizability beyond analyses related to stillbirth. Other studies have focused on the effects of maternal exposure to pollutants on preterm birth[25-27] and autism.[28] Our approach would provide the same improvements to temporal and spatial granularity in the analyses of these and other clinical outcomes associated with maternal exposures. A separate investigation using our approach has identified an association between high levels of $PM_{2.5}$ exposure and preterm birth.[29] As all pregnant women are potentially at risk for negative outcomes resulting from pollution exposures, our methods have broad applicability and relevance in better understanding the true risks to fetal and infant health associated with in-utero exposures.

Although the use of maternal residence enables additional spatial granularity over cruder county- or state-wide exposure approximations, our approach does not capture spatial mobility throughout the course of pregnancy. Our approach utilized a residence provided at the time of birth; however, a more detailed residence history would enable further customization of individual exposure approximations throughout the pregnancy period. In addition, our approach was unable to adjust to variations in pollutant levels within the 10 km radius of each monitoring station. The described variations introduce the limitation of underestimated risk attributed to maternal exposure.

Conclusions

Using a spatial approach to linking measures of air quality to vital records, we were able to identify characteristics of women residing within the coverage range of EPA monitoring stations. Although women residing within range of stations were significantly different in regards to socio-demographic characteristics than women residing outside the coverage range of monitoring stations, groups expressed similar clinical risk factors for stillbirth. While the spatial approach supports analysis of single particulate exposures, the inclusion of additional pollutants in an analysis of combined exposure effects may be prohibitively restrictive of eligible populations. Although, researchers should consider potential limitations associated with populations in close proximity to air monitoring stations, the described approach has generalizability to support the study of the effects of pollutant exposures on various pregnancy and birth related outcomes.

Acknowledgements

This study included data provided by the Ohio Department of Health, which should not be considered an endorsement of this study or its conclusions.

References

1. MacDorman MF, Kirmeyer S, Wilson EC. Fetal and Perinatal Mortality, United States, 2006. National Vital Statistics Reports. August 28, 2012;60(8).
2. Lawn JE, Gravett MG, Nunes TM, Rubens CE, Stanton C. Global report on preterm birth and stillbirth (1 of 7): definitions, description of the burden and opportunities to improve data. BMC Pregnancy Childbirth. 2010;10 Suppl 1:S1.
3. Stanton C, Lawn JE, Rahman H, Wilczynska-Ketende K, Hill K. Stillbirth rates: delivering estimates in 190 countries. Lancet. 2006 May 6;367(9521):1487-94.
4. U.S. Department of Health and Human Services. Healthy People 2010. Washington, DC 2000.
5. Healthy People 2020. Healthy People 2020 Summary of Objectives. Accessed January 16, 2014; http://www.healthypeople.gov/2020/topicsobjectives/objectiveslist.aspx?topicId=26].
6. Flenady V, Koopmans L, Middleton P, Froen JF, Smith GC, Gibbons K, et al. Major risk factors for stillbirth in high-income countries: a systematic review and meta-analysis. Lancet. 2011 Apr 16;377(9774):1331-40.
7. Fretts RC. Etiology and prevention of stillbirth. Am J Obstet Gynecol. 2005 Dec;193(6):1923-35.

8. Maisonet M, Correa A, Misra D, Jaakkola JJ. A review of the literature on the effects of ambient air pollution on fetal growth. Environ Res. 2004 May;95(1):106-15.
9. Shah PS, Balkhair T. Air pollution and birth outcomes: a systematic review. Environ Int. 2011 Feb;37(2):498-516.
10. United States Environmental Protection Agency. National Ambient Air Quality Standards (NAAQS). Accessed November 21, 2013; http://www.epa.gov/air/criteria.html].
11. Bobak M, Leon DA. Pregnancy outcomes and outdoor air pollution: an ecological study in districts of the Czech Republic 1986-8. Occup Environ Med. 1999 Aug;56(8):539-43.
12. Glinianaia SV, Rankin J, Bell R, Pless-Mulloli T, Howel D. Particulate air pollution and fetal health: a systematic review of the epidemiologic evidence. Epidemiology. 2004 Jan;15(1):36-45.
13. Landgren O. Environmental pollution and delivery outcome in southern Sweden: a study with central registries. Acta paediatrica. 1996 Nov;85(11):1361-4.
14. Pereira LA, Loomis D, Conceicao GM, Braga AL, Arcas RM, Kishi HS, et al. Association between air pollution and intrauterine mortality in Sao Paulo, Brazil. Environ Health Perspect. 1998 Jun;106(6):325-9.
15. Vassilev ZP, Robson MG, Klotz JB. Outdoor exposure to airborne polycyclic organic matter and adverse reproductive outcomes: a pilot study. Am J Ind Med. 2001 Sep;40(3):255-62.
16. Hwang BF, Lee YL, Jaakkola JJ. Air pollution and stillbirth: a population-based case-control study in Taiwan. Environ Health Perspect. 2011 Sep;119(9):1345-9.
17. Pearce MS, Glinianaia SV, Rankin J, Rushton S, Charlton M, Parker L, et al. No association between ambient particulate matter exposure during pregnancy and stillbirth risk in the north of England, 1962-1992. Environ Res. 2010 Jan;110(1):118-22.
18. Faiz AS, Rhoads GG, Demissie K, Kruse L, Lin Y, Rich DQ. Ambient air pollution and the risk of stillbirth. American journal of epidemiology. 2012 Aug 15;176(4):308-16.
19. Hall ES, Goyal NK, Ammerman RT, Miller MM, Jones DE, Short JA, et al. Development of a linked perinatal data resource from state administrative and community-based program data. Matern Child Health J. 2014 Jan;18(1):316-25.
20. South AP, Jones DE, Hall ES, Huo S, Meinzen-Derr J, Liu L, et al. Spatial analysis of preterm birth demonstrates opportunities for targeted intervention. Maternal and Child Health Journal. 2012 Feb;16(2):470-8.
21. United States Environmental Protection Agency. AirData. [January 15, 2013]; Available from: http://www.epa.gov/airdata/.
22. Faiz AS, Rhoads GG, Demissie K, Lin Y, Kruse L, Rich DQ. Does ambient air pollution trigger stillbirth? Epidemiology. 2013 Jul;24(4):538-44.
23. Holm S. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics. 1979;6:65-70.
24. DeFranco E, Hall E, Hossain M, Chen A, Haynes E, Jones D, et al. Air pollution and stillbirth risk: exposure to airborne particulate matter during pregnancy is associated with fetal death. Unpublished results. 2013.
25. Pereira G, Belanger K, Ebisu K, Bell ML. Fine Particulate Matter and Risk of Preterm Birth in Connecticut in 2000-2006: A Longitudinal Study. American journal of epidemiology. 2013 Sep 25.
26. Ritz B, Yu F, Chapa G, Fruin S. Effect of air pollution on preterm birth among children born in Southern California between 1989 and 1993. Epidemiology. 2000 Sep;11(5):502-11.
27. Sram RJ, Binkova B, Dejmek J, Bobak M. Ambient air pollution and pregnancy outcomes: a review of the literature. Environ Health Perspect. 2005 Apr;113(4):375-82.
28. Becerra TA, Wilhelm M, Olsen J, Cockburn M, Ritz B. Ambient air pollution and autism in Los Angeles county, California. Environ Health Perspect. 2013 Mar;121(3):380-6.
29. DeFranco E, Chen A, Xu F, Hall E, Hossain M, Haynes E, et al. Air pollution and risk of prematurity: exposure to airborne particulate matter during pregnancy is associated with preterm birth risk. Am J Obstet Gynecol. 2014;210(1):S346.

Using Anchors to Estimate Clinical State without Labeled Data

Yoni Halpern[1], Youngduck Choi[1], Steven Horng MD MMSc[2], David Sontag PhD[1]
[1]New York University, New York, NY
[2]Beth Israel Deaconess Medical Center, Boston, MA

Abstract

We present a novel framework for learning to estimate and predict clinical state variables without labeled data. The resulting models can used for electronic phenotyping, triggering clinical decision support, and cohort selection. The framework relies on key observations which we characterize and term "anchor variables". By specifying anchor variables, an expert encodes a certain amount of domain knowledge about the problem while the rest of learning proceeds in an unsupervised manner. The ability to build anchors upon standardized ontologies and the framework's ability to learn from unlabeled data promote generalizability across institutions. We additionally develop a user interface to enable experts to choose anchor variables in an informed manner. The framework is applied to electronic medical record-based phenotyping to enable real-time decision support in the emergency department. We validate the learned models using a prospectively gathered set of gold-standard responses from emergency physicians for nine clinically relevant variables.

1 Introduction

Health information technology is an essential part of modern health care, providing health care professionals with critical information about patients and allowing them to make maximally informed decisions about a patient's care.

In order to accelerate the development of advanced clinical decision support tools, we seek to create a new middleware application layer consisting of hundreds of clinical state variables that summarize a patient's past and current state. These clinical state variables collectively form a patient phenotype that is continuously estimated throughout a patient's stay and can be used by decision support applications to better inform, guide, and expedite the workflows of clinicians. We define clinical decision support for this paper very broadly to include any functionality that helps a clinician to be better, whether this means more complete or efficient documentation, adherence to clinical guidelines, disease specific order sets, alerts and reminders, visualizations that summarize patient information, or contextual information retrieval, extraction, and summarization.

For example, nursing home patients have different clinical care needs and treatments than other patients in the emergency department. They are more likely to have resistant organisms to standard antibiotic therapy, and therefore must be empirically treated with broad spectrum antibiotics. They are also more likely to fall and therefore require special handling to minimize fall risk. Decision support tools can be used to remind clinicians to order broad spectrum antibiotics and warn them when they do not. They can also be used to alert ancillary staff such as patient transporters and radiology technicians to take appropriate fall precautions, a precaution that may not always be obvious. Although whether a patient is from a nursing home could be collected manually as structured data, there are hundreds of clinical state variables that would be valuable for decision support. Collecting all of these variables for all patients is not feasible. Previous systems that use this type of approach often fail to get the support of clinical users and are systematically not used.

Standard ontologies and knowledge representations are necessary but not sufficient to catalyze the development of advanced decision support and enable transferability of models across institutions. In particular, the patient's state may not be directly observable. Machine learning allows us to reason about patients in these settings. Previous approaches such as logistic regression, support vector machines, decision trees, and neural networks require domain experts to label a fairly large number of positive and negative examples which is time consuming (e.g. [1, 2, 3]). Even after this labeling work has been done, these learned classifiers often do not generalize well across institutions since the learned classifiers are highly dependent on the representation they are trained on. Retraining for each site can require repeating the labeling process, modifying the representation and adjusting rules (e.g. [4, 5]). A review of recent approaches to automated patient phenotyping from electronic records can be found in [6].

In this paper, we describe a methodology for learning to estimate a patient phenotype, consisting of hundreds of different clinical state variables, based on information available in the Electronic Medical Record (EMR). We

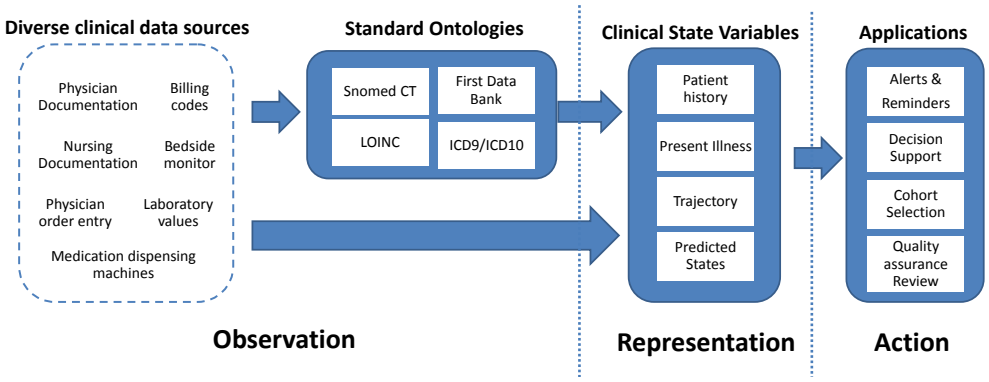


Figure 1: Schematic diagram of a middleware layer, reading inputs from the EMR and acting as an intermediate representation for applications. To improve generalization across institutions, we make no assumptions about the raw clinical data sources. Rather, we make use of existing standardized ontologies to specify anchors for each clinical state variable, and then use the anchors within each institution to learn classifiers to estimate the variables using previously collected, unlabeled, raw clinical data.

use a combination of domain expertise and vast amounts of unlabeled data, rather than relying solely on domain expertise or requiring labor-intensive manual labeling. To learn the models, we require only that certain highly informative variables that we call *anchor variables* be identified by an expert, and the rest is learned from large amounts of unlabeled data in an unsupervised manner. To promote transferability between institutions, we assume only that these anchor variables remain stable between institutions while the rest of the underlying observation model can change. Figure 1 presents a schematic view of such a system.

The main contributions of this paper are as follows:

1. We introduce the concept of anchor variables.
2. We show how to use anchors within an unsupervised machine learning algorithm to estimate each clinical state variable without any human labeling.
3. We describe a novel user interface developed to help with choosing a good set of anchors for each clinical state variable and for performing interactive cohort selection.
4. We evaluate our algorithm's performance using a prospectively gathered set of gold-standard responses from emergency physicians on nine different clinically relevant patient phenotyping tasks.

More broadly, this paper presents a novel approach to harnessing unstructured and structured data found in electronic medical records to estimate clinical state variables that can be used in a wide variety of settings, including both retrospective and prospective clinical care, research, administration, and quality improvement.

2 Anchor variable framework

2.1 Observed and latent variables in the EMR

We formalize anchor variables within the context of latent variable models. In our framework, variables are categorized as either *latent* or *observed*. In a model with m latent variables and n observed variables, each patient is described with a collection of latent variables $\mathcal{Y} = \{Y_1, Y_2, ...Y_m\}$ and a collection of observations $\mathcal{X} = \{X_1, X_2, .., X_n\}$. Here we focus on binary variables but the work can be readily extended to consider categorical variables as well.

Observed variables represent quantities that can be observed directly from the EMR. These include structured variables like age and sex, but also queries that can be computed on semi-structured or free-text data, such as *Does any of the free text contain the phrase "chest pain"?* or *Does the medication list include "GSN_004380" (aspirin)?* Here we assume a very general form of the EMR without making assumptions of which fields are present and what structure they take. These observations are represented in Figure 1 as the leftmost section of the schematic.

Observations from the EMR are inherently noisy and may miss important information. For example, if the note contains the short-form "cp" instead of chest pain, the chest pain observation would be negative even though

the patient may actually have the symptom. Our model addresses this by explicitly treating observations as noisy evidence about the clinical state variables of interest. Additionally, as we will see later on in more detail, we allow certain key observations to be trusted as true when they are *positive* but not when they are negative.

Latent variables represent quantities that **cannot** be observed directly from the EMR or computed as a simple query, but for which the EMR can provide information that would be useful to answer the question. These variables could be the answers to higher level questions such as *Does the patient have* $\{$*an infection, altered mental status, a history of alcoholism*$\}$*?* The answers to these higher level clinical state questions form a useful representation of the patient, shown in the middle section of Figure 1. In many cases, the answers can be found in the EMR, though they are usually found in free-text sections of the record. Formulating queries to extract this information directly from the free-text is notoriously difficult due to the vast number of ways each fact can be expressed [7]. In other cases, the answers may simply not be documented at all, either because they are not known or due to a failure to document. In all of these cases, even though it is impossible to observe the answers directly, it may still be possible to infer them based on all of the other data observable in the record. In our framework, the goal will be to learn a classifier to predict the value of the latent variables $\mathcal{Y}$ with only access to the observations $\mathcal{X}$.

2.2 Anchor variables

It is important to recognize that many different EMR systems exist and the set of observable variables in one system may not map directly to the set of observable variables in another [8]. Promulgation of standards in the form of ontologies to normalize the representation of observations as well as knowledge representation standards such as the Continuity of Care Document will help to reduce some of this effect. However, different vendors will continue to innovate beyond these standardizations and there will always be a structural differences in data representations. In our framework we assume that the set of observable variables can change from system to system as long as a few key informative observations are conserved. These observations are *anchor* observations with the following property:

Definition 1. *An observation X_j is an* anchor *for a latent variable Y_i if X_j is conditionally independent of all other observations, X_k $\forall k \neq j$, conditioned on the value of Y_i.*

In other words, anchor observations provide a direct, albeit noisy, view of the underlying latent variable we wish to predict. The key characteristic of an anchor is the conditional independence property which states once the value of the latent variable is known, no other observables provide additional information about the anchor variable. While anchors tend to provide strong evidence towards the value of the latent variable, it would not make sense to answer the questions simply based on the values of the anchors alone. However, our insight is that we can explicitly treat anchors as noisy labels and use them within a learning algorithm. Learning with noisy labels is a subject that has been studied extensively in machine learning literature (e.g. [9, 10]) and we leverage that work here. By using domain knowledge to identify these noisy labels in the data itself, we no longer require manual labeling of the data before it can be fed to a machine learning algorithm. The resulting method is extremely portable. As long as anchors can be shared between institutions, no new labeling work is required to train the classifiers for a new institution's data.

In the next section we describe an unsupervised method to learn decision rules to predict the values of latent variables, $\mathcal{Y}$, using all of the the observed variables, $\mathcal{X}$, when anchors for each latent variable have been specified. When transferring to a new system with a potentially different set of observed variables we can perform the same unsupervised training on the data available in the new system with no additional human input as long as the anchors are preserved or can be mapped into the new data system. In order to ensure that anchors are not particular to an institution, we can use standard ontologies when specifying the anchors.

2.3 Learning decision rules with positive anchors

Let $A \in \mathcal{X}$ be an anchor variable for Y_i. The collection $\tilde{\mathcal{X}} = \mathcal{X} \setminus A$ represents the observations $\mathcal{X}$ with the anchor removed. We describe a procedure for learning a single decision rule, trying to predict the value of a single latent variable Y_i from observed values X using a special type of anchor that we call *positive* anchors.

Definition 2. *An observation A is a* positive anchor *for a latent variable Y_i if it is an anchor for Y_i and $P(Y_i = 1|A = 1) = 1$.*

Intuitively, observing the anchor to be positive unambiguously reveals the state of the latent variable to be positive, while observing it to be negative does not reveal the state of the latent variable. This setting has previously been studied under the name *positive-only labels* [9] and a procedure for learning with positive-only labels is as follows (see [9] for a detailed derivation):

1. Learn a calibrated classifier (e.g. logistic regression) to predict $P(A = 1|\tilde{\mathcal{X}})$.
2. Using a validate set, compute $C = \frac{1}{|\mathcal{P}|} \sum_{k \in \mathcal{P}} P(A = 1|\tilde{\mathcal{X}}^{(k)})$ where $\mathcal{P}$ is the set of data in the validate set where $A = 1$.
3. For a previously unseen patient t, predict $\begin{cases} P(A = 1|\tilde{\mathcal{X}}^{(t)})/C & \text{if } A^{(t)} = 0 \\ 1 & \text{if } A^{(t)} = 1 \end{cases}$

Positive anchors have some appealing properties. First, the algorithm to learn with them is extremely simple and requires only the capability of learning logistic regression models, which are standard in most statistical packages. Second, positive anchors have an intuitive interpretation for the human providing the anchors. For positive anchors, the anchor should be a quantity that *can only be caused* by the latent variable being "on". Finally, the positivity of an anchor is easy to verify. If an expert is presented with examples of patients with a positive anchor, they can confirm that in fact all (or almost all) of the presented patients are positive for the latent variable of interest. Conditional independence is more difficult to verify and requires that an expert verify that the proposed anchor is truly completely explained by the clinical state of interest.

Anchors like this exist in medicine. For example, a positive rapid antigen test for Group A streptococci is very specific for strep throat and can serve as a positive anchor. The absence of a positive test result can be uninformative either due to the low sensitivity of the test, or due to the possibility that the test was never performed because the diagnosis was obvious and the patient was treated without testing [11]. In addition to carefully choosing anchors, data preprocessing can be performed to increase the amount of conditional independence between anchors and other observations. For example, in the Methods section we describe how common bigrams are represented in order to avoid obvious violations of conditional independence. In real use cases, no anchors will ever perfectly meet the criteria in Definitions 1 and 2, and missing data will not be completely at random. Nonetheless, approximate anchors can perform well on real data, as we demonstrate with experimental results for a range of clinical variables. The above definitions are still useful as they give theoretical principles by which to choose good anchors.

If multiple anchors are specified, they can be combined in a number of different ways. The simplest way is to create a composite anchor out of the union of all of the individual anchors. For example, if the diagnosis code ICD9-288.00 (neutropenia, unspecified) and the word "immunocompromised" are both anchors for the latent variable `isImmunosuppressed`, then we can create a single composite anchor which is present if *either* of these two observations is present. If the original anchors are positive as in Definition 2, then the new composite anchor is also a positive anchor. Using the composite anchor is advantageous compared to choosing a single anchor because it occurs more frequently, providing more positive examples for training.

3 Specifying anchors with an interactive display

In practice, specifying anchors can be challenging. In order to specify anchors, one must have sufficient domain knowledge to evaluate whether a variable fits the definition of an anchor. To ease the process of eliciting anchors from domain experts, we built an interactive interface to allow domain experts to specify anchor variables and visualize the resulting model learned with those anchors. Figure 2 shows a screenshot of the tool being used to specify anchors to identify HIV positive patients.

The interface is a general tool for specifying anchors and viewing the learned classifier. A user can add latent variables and specify anchors for them. After adding an anchor, the user can, in real time, update the learned model and view a ranked list of patients at the bottom of the screen. The ranking is generated according to the predicted likelihood of the latent variable being positive according to the model built with the current set of anchors. For each patient, a short summary is presented for easy viewing, and selected patients can be viewed in more detail in the middle pane. Patients can be filtered according to three different criteria: view only patients with anchors (to judge whether the anchors are catching the correct subset of patients), view patients that have the most recently added anchor (to judge the incremental effect) and view patients without anchors (looking at a ranked list of these patients provides an idea of how well the learning algorithm has *generalized* beyond simply looking for patients that have the anchors).

After learning a model, the tool additionally *suggests* new anchors by showing the observations ranked by weights of a linear classifier learned with a penalty on the L1 norm of the weight vector. The L1 penalty encourages the learned classifier to use a minimal number of variables, effectively selecting highly informative observations. The user then uses clinical judgment to decide whether or not each suggestion would make a good anchor, e.g. by including the new observation and seeing the incremental effect on the ranking, or by viewing the newly anchored patients. The result is a simplified active learning workflow with a human-in-the-loop. In this work we focus on

the basic task of learning classifiers with anchors, leaving more advanced active learning techniques like asking questions about specific patients in a maximally informative manner [12] and providing detailed performance feedback for the user to future work.

The ranking and filtering mechanisms provide feedback to the user, giving information about whether the model is being built reasonably or not. Figure 3 in Section 5 shows a user trace of a clinician using our interface to specify anchors to identify patients with a cardiac etiology. The feedback coming from patient rankings, viewing recently anchored patients and looking at suggested anchors is sufficient to allow him to incrementally build better models by specifying new anchors.

Anchors can be specified as words or phrases, and they are interpreted as queries on the free text portions of the medical record. Additionally, the interface allows for incorporation of anchors according to standardized hierarchical ontologies. For example, medications are grouped by families according to First Databank's Enhanced Therapeutic Classification (ETC) hierarchy, and diagnosis codes are grouped according to the ICD9 hierarchy. For these hierarchical structures, including a parent as an anchor automatically adds all of its children as well.

In addition to specifying anchors, the tool is useful for performing fast interactive cohort selection, allowing the user to quickly learn classifiers to find members of a target population using the anchor approach. The learned classifiers can be exported as well for use in real-time decision applications. The tool is freely available for download at `http://sontaglab.cs.nyu.edu/`.

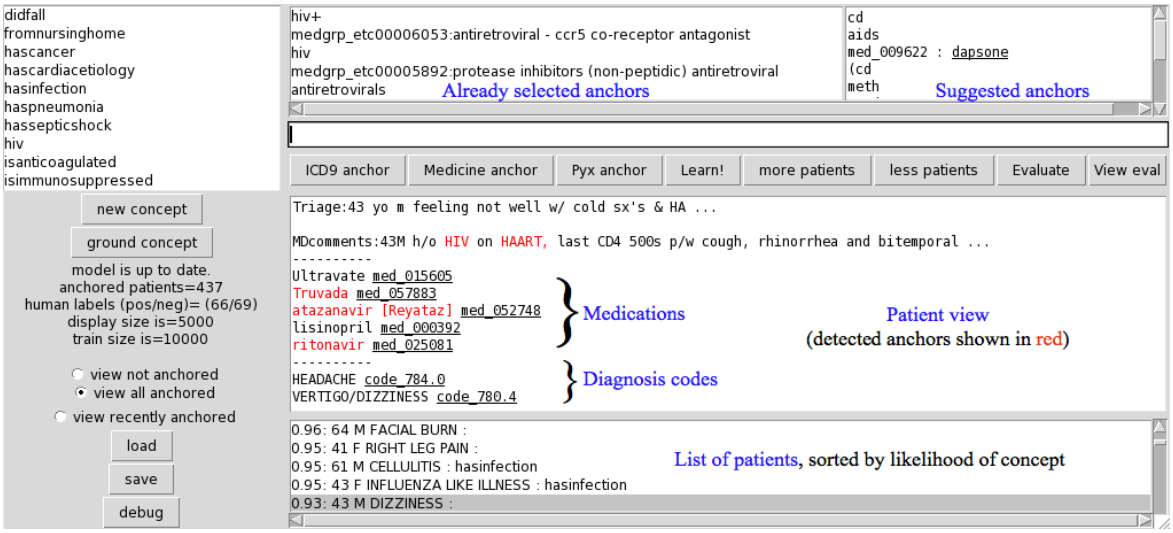


Figure 2: A screenshot of the anchor elicitation tool using deidentified patient information. Conditional highlighting emphasizes the presence of previously specified anchors in a note so that the physician can determine quickly whether its usage is as expected.

4 Methods

4.1 Data description

For training and evaluation we use a collection of 273,174 emergency department patient records collected from a 55,000 patient/year Level 1 trauma center and tertiary academic teaching hospital between 2008 and 2013. Each record represents a single patient visit. All consecutive ED patient visits were included in the data set. No visits were excluded. This study was approved by our institutional review board.

In order to evaluate the utility of the latent variable predictors learned, we collected gold standard labels from the primary clinical provider caring for a patient at the time of disposition from the emergency department, (admission, discharge, or transfer). As part of the routine clinical workflow of disposition, clinicians were asked a series of 2-3 questions chosen randomly from a rotating pool of questions. The answers serve as gold standard labels. These questions were entirely voluntary, but received an approximately 85% completion rate. As part of our quality assurance process, we routinely review random samples of cases to confirm the quality of data obtained using this technique and confirm that missing data occurs completely at random. Since only 2-3 questions were asked at any given time, it took approximately 2 years to collect the labels for this paper, in addition to labels for other unrelated research projects not presented here. Table 1 shows the clinical state variables and the associated disposition questions that we use for evaluation in this paper. Responses were collected for all consecutive patients

except for the `didFall` question which was only asked for patients who had a head CT scan. Clinicians reported responses using a five point scale with 1 being most negative and 5 being most positive. The `didFall` question used a three-point scale of "No", "Uncertain" or "Yes". When converting the labels to binary values we take 4 or above as positive. Responses labeled as "Uncertain" are treated as unlabeled.

Latent Variable	Disposition Question	Additional Information	Labels Collected	Fraction Positive
didFall	Did the patient fall from standing or lesser height?	None	9,831	0.118
hasCardiacEtiology	In the workup of this patient, was a cardiac etiology suspected?	None	17,258	0.068
hasInfection	Do you think this patient has an infection?	Suspected or proven viral, fungal, protozoal, or bacterial infection	62,589	0.213
fromNursingHome	Is the patient from a nursing home or similar facility?	Interpret as if you would be giving broad spectrum antibiotics	36,256	0.045
hasCancer	Does the patient have an active malignancy?	Malignancy not in remission; and recent enough to change clinical thinking	4,091	0.042
hasPneumonia	Do you think the patient has pneumonia?	None	9,934	0.073
isAnticoagulated	Prior to this visit, was the patient on anticoagulation?	Excluding antiplatelet agents like aspirin or plavix	1,082	0.047
isImmunosuppressed	Is the patient currently immunocompromised?	None	12,857	0.040
hasSepticShock	Is the patient in septic shock?	None	6,867	0.020

Table 1: Questions asked of physicians at disposition time to obtain a gold standard set of labels. Additional information was displayed in a clickthrough screen with a link from the main question page.

4.2 Representation and preprocessing

Patient records are represented as containing six distinct types of observable variables which come from semi-structured sections of the EMR: 1. ICD9 diagnosis codes (from billing information), 2. current medications recorded during medication reconciliation, 3. medications dispensed during the ED course as reported by medication dispensing machines (Pyxis), 4. free text sections formed by a concatenation of chief complaint, triage assessment and physician's comments, 5. Age, 6. Sex.

Deidentified free text was preprocessed using a modified version of NegEx [13, 14] and negated words were replaced by a new token (i.e. if the token "fever" was within the scope of a negation, it was transformed to a new token, "negfever"). A second step of preprocessing collected 1,500 significant bigrams and appended them to the text (i.e. the phrase "chest pain" was augmented to be "chest pain chest-pain" with an extra token representing the bigram). When learning with anchors, we remove the component words (i.e. "chest pain" is replaced by a single token "chest-pain"). We do this in order to increase the amount of conditional independence between anchors which are bigrams and the rest of the text. If the token "chest-pain" is chosen as an anchor, it will not be conditionally independent of the tokens "chest" and "pain" without the removal step. For training linear classifiers, the first representation is strictly more general, so it should not hurt the performance of our baseline algorithms.

Medications are represented by generic sequence number (GSN) and diagnosis codes by ICD9 codes. Age was discretized by decade with a binary indicator for each decade. Patients are represented as a binary feature vector representing the presence or absence of each distinct diagnosis code, current medication, dispensed medication, word, discretized age value and sex. Observations that occur in fewer than 50 patients in the entire dataset were discarded, leaving a final binary feature vector of size 20,334.

4.3 Anchor specification

A single emergency physician specified anchors for each clinical state variable using our custom anchor elicitation tool with access to a database of 20,000 unlabeled patients chosen at random from the full patient set. Figure 3 shows the evolution of the performance of the learned classifier as the physician specified anchors. Unless otherwise noted, results are reported using the final set of anchors specified by the physician. Note that the physician was not provided with explicit feedback about the performance of the model on the ground truth labels, but was able to use the interface to determine which anchors were useful and make progress. In order to accurately assess

the effort involved in specifying anchors for a new classification task and to avoid overfitting, each anchor-based predictive model was only built once with the exception of `hasCancer` which was used as an example for development purposes.

4.4 Machine learning

We compare the classifiers learned using anchors to a simple rule-based baseline and a supervised machine learning baseline which uses a subset of the collected gold standard labels for training. Evaluation is reported on nine separate estimation tasks, one for each clinical state variable in Table 1. We emphasize that unlike the supervised baseline, the anchor algorithm does not require ground truth labels for training.

The rule-based baseline simply predicts positively when at least one anchor is present and negatively otherwise. This approach also requires no training, and is evaluated on the entire labeled set.

Evaluation of the supervised baseline is reported using 4-fold cross validation in order to fully utilize the limited number of gold standard labels we have for some of our clinical state variables. In each experiment, the labeled patients are divided into four equal-sized test sets. For each test set, a classifier is trained using a portion of the 75% of patients which are not in the test set ("training patients") and then used to predict for the 25% of patients designated as test. The results are averaged across the four test sets, giving an estimate of the performance on the entire labeled dataset. Each classifier of the supervised baseline is learned using at most 3000 training patients, representing approximately 3 weeks-worth of patients at our institution.

The supervised baseline was learned with logistic regression using the scikit-learn package [15] in Python. We use 5-fold cross validation within the train set to choose parameters, trying all combinations of the regularization constant (options are $\{10^{-6}, 10^{-5}, ..., 10^{6}\}$) and the norm used in regularization. Choices for the norm are L1 (encourages the learned classifier to use a minimal number of features by penalizing the sum of absolute values of the regression weights) or L2 (avoids overly emphasizing any one feature by penalizing the sum of squares of the regression weights). For the supervised results for 100 and 200 labels, 5-fold cross validation is not viable, due to the low number of positively labeled samples for each disposition question. Hence, for these numbers of labels the parameters are chosen to be the default values in scikit-learn. Lastly, we reduce regularization of the bias parameter by setting the "intercept_scaling" parameter to 1000.

The anchor method is trained using the specified anchors and 200,000 examples chosen randomly from the unlabeled dataset, and tested on the entire labeled set. Scikit-learn is used to fit logistic regression models as in the supervised setting, but holding the regularization norm fixed as L2 and doing cross-validation over the regularization parameter. Since the logistic regression models learned in the anchor method are meant to predict the presence or absence of the *anchor* (as described in Section 2.3), the cross-validation technique to choose parameters also uses the presence or absence of the anchor to measure performance, requiring no ground truth labels.

We measure performance using area under the ROC curve (AUC), a measure of the overall quality of a ranking predictor. Estimating the constant C in step 2 of the anchor algorithm is not necessary to obtain a ranking, so we omit that step. In the rule-based approach, ties are broken by counting the number of distinct anchors present in the patient record. In the anchor approach, ties among patients with anchors are broken according to the predicted probability of the latent variable ignoring the presence of the anchors.

4.5 Real-time decision support evaluation

We evaluate a real-time decision support scenario where the estimation tasks are performed without access to diagnosis codes (i.e., diagnosis codes are excluded from the feature vector), as these would usually be assigned after the patient leaves the emergency department. The supervised baseline and rule-based approach simply ignore all ICD9 codes because they cannot incorporate extra information that is not available at test time. However, since the algorithms are meant to be trained on previous patients, it is reasonable to assume that diagnosis codes would be available at the time of training. Thus, the anchor algorithm uses ICD9 codes as *anchors* during training, even though the resulting prediction rules do not use them as features.

We also tested a retrospective setting where the algorithms have access to diagnosis codes, both at train and test time. This setting is meaningful since our algorithm can also be used in retrospective settings for tasks such as cohort selection, information retrieval and quality control. The results were qualitatively similar, so we only present the real-time setting here.

5 Results

In this section we present results comparing the performance of the anchor-based learning method to the baseline methods on the task of predicting the nine clinical state variables listed previously in Table 1.

First, we show the utility of the anchor-specification interface described in Section 3. Figure 3 shows the learning path for the `hasCardiacEtiology` clinical state variable, describing the changes to AUC as the clinician added and subtracted anchors from the model. It is noteworthy that using our interface, the quality of the model tends to increase as time progresses. We observed this trend for all of the clinical state variables. Table 2 shows some of the final anchors specified by the clinician who used the interface. Using our interactive tool, the total time to specify anchors for all nine models was approximately 5 hours.

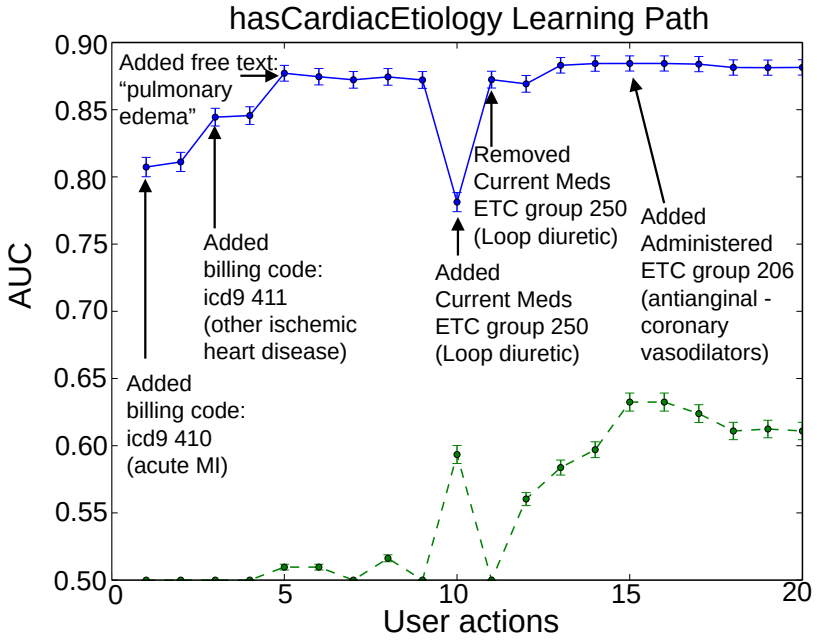


Figure 3: A learning path for one of the learned models (`hasCardiacEtiology`) as the interface is being used to build this model. On the x-axis are user actions, either additions to or deletions from the current set of anchors. On the y-axis is AUC evaluated on the gold standard collected labels. The dotted line shows the progress of the rules baseline which simply predicts 1 when an anchor is present and 0 otherwise. The rules baseline initially has an AUC of 0.5 since the earliest anchors are ICD9 codes and we consider the real-time setting where ICD9 codes are not available for the test set.

Clinical State Variable	Selected Anchors
didFall	Billing code [Accidental Falls] (ICD9 E880-E888), "slipped", "s/p-fall', "slip-fall", "fell down", "mechanical_fall", "witnessed-fall", ...
hasSepticShock	Billing code [septic shock] (ICD9 785.52) , Administered meds from [cardiac sympathomimetics] (ETC:00003064)
isimmunosuppressed	Current meds from [antineoplastic - antimetabolite - folic acid analogs] (ETC:00002881), "immunocompromised", ...
hascancer	Billing code [neoplasms] (ICD9 group 2), "breast-ca", "mets", "oncology", ...

Table 2: A selection of anchors specified by a clinical collaborator. Anchors that utilize structured variables such as diagnosis codes or medications (current or administered) are mapped to a relevant structured ontology.

Table 3 shows a comparison to supervised learning for the real-time setting described in Section 4.5. For many of the clinical state variables, the anchor algorithm outperforms learning with 3K labels, suggesting that it would take a non-trivial amount of time and human effort to collect the data necessary to train in a supervised setting with comparable accuracy. The `didFall` task was evaluated on a biased population since the question was only asked for patients with a head CT scan. The supervised method was trained on a similarly biased population, so it makes sense that the performance of the anchor algorithm underperforms in this setting.

To provide a comparison to a semi-supervised algorithm, we experimented with training using SVMlight in a transductive setting [16], using up to 20,000 visit instances (500 labeled, 19500 unlabeled) for the `hasInfection`

variable. Samples were reweighted to account for the class imbalance. However, we found that this semi-supervised approach did not improve AUC over the supervised baseline with 500 labeled examples.

Variables	Rules	Supervised						Anchors
		100	200	500	1K	2K	3K (min, max)	
didFall	0.725 ± 0.008	0.814	0.852	0.900	0.914	0.920	**0.924** (0.917,0.934)	0.883 ± 0.006
hasCardiacEtiology	0.611 ± 0.006	0.772	0.827	0.824	0.875	0.900	**0.906** (0.891, 0.920)	0.881 ± 0.006
hasInfection	0.723 ± 0.002	0.728	0.767	0.804	0.830	0.861	0.883 (0.881, 0.886)	**0.903± 0.001**
fromNursingHome	0.620 ± 0.005	0.725	0.792	0.822	0.869	0.894	0.891 (0.873, 0.906)	**0.918 ± 0.004**
hasCancer	0.822 ± 0.018	0.635	0.673	0.693	0.810	0.882	0.902 (0.880, 0.930)	**0.945 ± 0.01**
hasPneumonia	-	0.856	0.907	0.933	0.947	0.956	0.963 (0.954, 0.972)	**0.971 ± 0.003**
isAnticoagulated	0.849 ± 0.03	-	-	-	-	-	-	**0.930 ± 0.02**
isImmunosuppressed	0.650 ± 0.01	0.584	0.659	0.740	0.814	0.842	**0.862** (0.840, 0.877)	0.840 ± 0.009
hasSepticShock	0.738 ± 0.02	-	0.760	0.773	0.863	0.920	0.952 (0.928, 0.967)	**0.967 ± 0.008**

Table 3: Comparing AUC in the real-time setting. The supervised method is trained using logistic regression with a small number of gold standard labels. When the anchors are composed entirely of diagnosis codes, the rules approach cannot be meaningfully evaluated on the test set (in the real-time setting, diagnosis codes are not available at test time). When we had insufficient data to train, the supervised approach could not be evaluated. Best methods in each row are bolded. The anchor approach uses 200K *unlabeled* examples in training. Standard errors of the AUC for Rules and Anchors are computed using 1000 bootstrap samples of the test set. Min and max values for the 3K supervised baseline are from the 4-fold cross validation.

6 Discussion

Across the nine clinical state variables considered in our evaluation, our anchor-based unsupervised learning algorithm obtains prediction accuracy comparable to and in many cases better than a supervised prediction algorithm. It is important to note that the clinical states we are interested in are precisely those for which ground truth labels cannot be easily derived from diagnosis codes or from natural language processing on the clinical notes. Labeling data would be expensive, time consuming, and in many cases institution-specific since the learned predictors may not generalize.

Surprisingly, despite the physician not receiving explicit feedback about the performance of the model on the ground truth label, using our user interface he was able to determine when adding an anchor helped or hurt overall performance (as seen by a generally monotonic increase in AUC with each additional anchor specified). Our initial user interface also allowed the physician to label individual patients as positive or negative for a given clinical state variable. However, despite trying various ways of integrating this feedback into our learning algorithm, we found negligible gains in accuracy compared to only using the anchors.

There are several interesting directions for future work. Our current approach predicts each clinical state variable independently, but it would also be interesting to jointly model the clinical states. Doing so may provide a solution for the vexing problem of how to efficiently provide the learning algorithm *negative* feedback, which could be addressed by introducing additional variables (and anchors for them). We could then use negative correlations between clinical state variables to disambiguate commonly confused clinical states.

Although our algorithm can provide useful predictions without any labeled data, its use within a clinical setting presents opportunities to gather additional data to improve its performance. One direction for future work would be to develop a feedback mechanism (implicit or explicit) to learn from the algorithm's use. We also plan to integrate our predictions into an active learning algorithm to be used with the current system that asks questions prospectively at the time of disposition from the ED. We currently rotate through the questions asked so as to not overburden the clinician. Instead, we can carefully select which questions to ask for the specific patient at hand so as to best improve our prediction algorithms for a range of clinical state variables.

The most important next step will be to test the generalization of anchor-based learning in departments other than the ED and in other institutions. We are also continuing to use the user interface to specify anchors for dozens of additional clinical state variables, using these to enable new contextual user interfaces and to trigger decision support.

Acknowledgments

This work is partially supported by a Google Faculty Research Award, grant UL1 TR000038 from NCATS, NIH and CIMIT Award No. 12-1262 under U.S. Army Medical Research Acquisition Activity Cooperative Agreement W81XWH-09-2-0001. Yoni Halpern was supported by an NSERC Postgraduate Scholarship. The information contained herein does not necessarily reflect the position or policy of the Government, and no official endorsement should be inferred.

References

[1] Nachimuthu SK, Haug PJ. Early detection of sepsis in the emergency department using dynamic Bayesian networks. In: AMIA Annual Symposium Proceedings. vol. 2012. American Medical Informatics Association; 2012. p. 653.

[2] Kawaler E, Cobian A, Peissig P, Cross D, Yale S, Craven M. Learning to predict post-hospitalization VTE risk from EHR data. In: AMIA Annual Symposium Proceedings. vol. 2012. American Medical Informatics Association; 2012. p. 436.

[3] DeLisle S, South B, Anthony JA, Kalp E, Gundlapallli A, Curriero FC, et al. Combining free text and structured electronic medical record entries to detect acute respiratory infections. PloS one. 2010;5(10):e13377.

[4] Liu M, Shah A, Jiang M, Peterson NB, Dai Q, Aldrich MC, et al. A study of transportability of an existing smoking status detection module across institutions. In: AMIA Annual Symposium Proceedings. vol. 2012. American Medical Informatics Association; 2012. p. 577.

[5] Carroll RJ, Thompson WK, Eyler AE, Mandelin AM, Cai T, Zink RM, et al. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. Journal of the American Medical Informatics Association. 2012;19(e1):e162–e169.

[6] Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. Journal of the American Medical Informatics Association. 2013;.

[7] Friedman C, Elhadad N. Natural language processing in health care and biomedicine. In: Biomedical Informatics. Springer; 2014. p. 255–284.

[8] Smith SW, Koppel R. Healthcare information technology's relativity problems: a typology of how patients' physical reality, clinicians' mental models, and healthcare information technology differ. Journal of the American Medical Informatics Association. 2013;.

[9] Elkan C, Noto K. Learning classifiers from only positive and unlabeled data. In: KDD; 2008. p. 213–220.

[10] Natarajan N, Dhillon I, Ravikumar P, Tewari A. Learning with noisy labels. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, editors. Advances in Neural Information Processing Systems 26; 2013. p. 1196–1204.

[11] Tanz RR, Gerber MA, Kabat W, Rippe J, Seshadri R, Shulman ST. Performance of a rapid antigen-detection test and throat culture in community pediatric offices: implications for management of pharyngitis. Pediatrics. 2009;123(2):437–44.

[12] Dasgupta S. Two faces of active learning. Theoretical Computer Science. 2011;412(19):1767 – 1781. Algorithmic Learning Theory (ALT 2009).

[13] Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal of biomedical informatics. 2001;34(5):301–310.

[14] Jernite Y, Halpern Y, Horng S, Sontag D. Predicting chief complaints at triage time in the emergency department. NIPS 2013 Workshop on Machine Learning for Clinical Data Analysis and Healthcare. 2013;.

[15] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al.. Scikit-learn: machine learning in python; 2011. `http://scikit-learn.org/0.14/`.

[16] Joachims T. Transductive inference for text classification using support vector machines. In: International Conference on Machine Learning (ICML). Bled, Slowenien; 1999. p. 200–209.

What Is Asked in Clinical Data Request Forms? A Multi-site Thematic Analysis of Forms Towards Better Data Access Support

***David A Hanauer[1,2], MD, MS, *Gregory W. Hruby[3], MA, Daniel G. Fort[3], MPH, Luke V. Rasmussen[4], BS, Eneida A. Mendonça[5,6], MD, PhD, Chunhua Weng[3,], PhD, MS**
(*equal contribution first-author)

[1]Dept. of Pediatrics, [2]School of Information, University of Michigan, Ann Arbor, MI; [3]Dept. of Biomedical Informatics, Columbia University, New York, NY; [4]Dept. of Preventive Medicine, Northwestern University, Chicago, IL; [5]Dept. Pediatrics, [6]Dept. of Biostatistics & Medical Informatics, University of Wisconsin, Madison, WI

Abstract

Many academic medical centers have aggregated data from multiple clinical systems into centralized repositories. These repositories can then be queried by skilled data analysts who act as intermediaries between the data stores and the research teams. To obtain data, researchers are often expected to complete a data request form. Such forms are meant to support record-keeping and, most importantly, provide a means for conveying complex data needs in a clear and understandable manner. Yet little is known about how data request forms are constructed and how effective they are likely to be. We conducted a content analysis of ten data request forms from CTSA-supported institutions. We found that most of the forms over-emphasized the collection of metadata that were not considered germane to the actual data needs. Based on our findings, we provide recommendations to improve the quality of data request forms in support of clinical and translational research.

Introduction

Clinical and translational research is a growing priority of the United States (US) National Institutes of Health (NIH). To encourage greater advancements in this area the NIH has supported over 60 research institutions through the Clinical and Translational Science Awards (CTSA).[1] At the same time there has been a substantial increase in the adoption of electronic health records (EHRs), with nearly half of US hospitals now with one or more EHRs in place.[2] This concomitant investment in both the research enterprise and in health information systems that capture data electronically has presented unprecedented opportunities for advancing clinical and translational science, and is a necessary precursor for building the foundations of a broad-scale 'learning health system'.[3, 4]

Research tasks that were previously impractical, if not impossible, to perform with paper-based health records have now become achievable due to the large volumes of data stored in a 'readily accessible' electronic format. However, in addition to privacy and security constraints, numerous difficulties remain with respect to access and use of the data.[5] Compared to paper records, EHR data should be much easier to aggregate across large numbers of patients, but the complexity of the underlying systems, including the heterogeneity in metadata, data structures, and even the data itself, often hinders their computational reuse by a broad range of stakeholders.[6-8] Further, prior work has shown that it is not uncommon for a hospital to have hundreds of different IT systems.[9] Data from multiple health information systems are thus often aggregated into databases commonly referred to as data or information warehouses, data repositories, data marts, or data networks.[10-16] A major theme of the NIH roadmap has been the idea of "Re-Engineering the Clinical Research Enterprise",[17] and one of the recognized challenges has been providing the means to "facilitate access to research…resources by scientists, clinicians" and others.[18] However, two major barriers exist with respect to data access.

First, to help meet the needs of clinical and translational research, 'self-service' tools have been developed to provide a means for data access as well as analysis and visualization.[19-26] Many of these tools have been widely implemented and have achieved a good level of adoption. While self-service tools have been demonstrated to work well for various scenarios,[27] by nature of their intended simplicity for a broad user base, these systems often cannot handle all of the complex data needs that are required by biomedical research teams.[20, 28] Researchers usually do not have the database knowledge nor do they understand what is involved in data retrieval. In contrast, data managers or query analysts usually do not know how to ask questions to elicit data needs using non-technical language understandable by researchers.[29] Data need negotiation usually involves several "trial-and-error" iterations. As a

result, many institutions have recognized the need to invest in informatics or IT experts (often called data analysts or report writers) to serve as an intermediary between the complex data sources and the biomedical researchers, the latter of whom have significant domain expertise but often lack training in data access approaches such as the use of structured query languages (SQL). [29-31]

Second, for liability considerations and HIPAA or regulatory compliance, data owners need to carefully check the credentials and qualifications of data requesters, which usually involves lengthy review processes involving multiple institutional review offices. An important artifact, the *data request form*, is the nexus linking all the stakeholders in the process of providing data access for researchers. Such forms are generally meant to serve documentation and communication needs for multiple stakeholders, including researchers, query analysts, data owners, and regulatory officers.[32] They can provide a means for researchers to list their credentials and specify their needs through a formal request process. They also help data stewards verify if the appropriate regulatory approvals are in place and to help with other administrative bookkeeping. Importantly, data request forms are also meant to provide a means for research teams to communicate complex data needs in a manner that can be understood by a data analyst and converted into executable database queries for data retrieval.[33] It follows, then, that the manner in which these forms clearly, and unambiguously, define the data needs and sources can have major downstream consequences for the subsequent research on which the request is based. Yet there are no published standards for designing EHR data request forms, or even best practices about which an institution can turn to in constructing a form. It is therefore up to each institution to develop their own form with the hope that the right questions are being asked of data requestors in order to ensure that data needs are being met accurately and efficiently.

Therefore, the data request form plays an indispensable role in facilitating data access for researchers in many institutions. Motivated to provide better data access to the broad clinical and translational research community, we aim to understand (1) how the current forms efficiently collect information needed by data owners and effectively communicate data needs of researchers and (2) if they collect necessary and relevant information that cannot be extracted for reuse from existing institutional information systems. In this study we conducted a formal content analysis of data request forms from multiple academic institutions affiliated with a CTSA award. Our goal was to develop a deeper understanding of what questions are typically asked on the forms, and to help provide insights regarding whether current data request forms provide adequate coverage of salient details to capture data needs effectively. To achieve these goals, first, we obtained ten data request forms from CTSA-supported academic medical centers in the US. Then we developed a form annotation schema based on the consensus of two annotators and used this coding book to annotate the forms. On this basis, we conducted a detailed content analysis of the forms and identified information deficiencies as well as unnecessary workload imposed on researchers that exist across many forms in use today. Finally, we provided insights and recommendations from our analysis that could be used to improve the content of data request forms and, ultimately, improve the process for obtaining complex data from institutional repositories in support of clinical and translational research.

Methods

A. Collection of data request forms

Ten data request forms were obtained for this study. All forms were in use at CTSA-supported academic medical centers around the US as of February 2014. Four of the forms were obtained through personal contacts by the authors, whereas the remaining six were identified through an online search with the Google search engine using the strings “EHR data request” and “medical research data request.” The ten CTSA-supported institutions from which these forms were actively in use were: Boston University, Columbia University, Northwestern University, University of California - San Diego, University of California - San Francisco, University of Colorado Denver, University of Kansas, University of Michigan, University of Wisconsin, and Vanderbilt University. Note that these institutions are listed here in alphabetical order, which does not match the order in which they are presented in the results section, wherein only a letter is used to identify each form.

B. Development of a codebook

Five of the ten data request forms were randomly selected for developing the coding schema for the content analysis. Two reviewers (GH and DH) independently evaluated the five forms and developed a list of themes derived from the forms. These themes were based only on the actual questions asked on each of the forms, although several

Table 1. Form elements comprising the codebook for the content analysis of the data request forms, including examples of each type of element. When an element could be coded as Simple [S] or Extensive [E], an example of each is provided. Basic elements were only coded as Simple [S] if present; thus no Extensive example is provided.

Code	Name	Description	Example(s)
1.0	*Requester Metadata*	*Any form elements that describe the user requesting data*	*not a coding element*
1.1	Name	This element may include the name, and/or contact data of the requester	[S] Requester Name [E] Requester Name, Department, Email
1.2	PI/Supervisor/ Department Head	This element may include the name, and/or contact data of the requester's PI, supervisor and/or department head	[S] Supervisor Name [E] Supervisor Name, Department, Email
1.3	Billing/ Administrative	This element may include the name, and/or contact data of the requester's administrator or other billing information	[S] Administrative Name [E] Administrative Name, Department, Email
1.4	Other	Any other attributes associated with the requester, and not associated with the content of the request	[S] Are you a part of the CTSA?
2.0	*Request Metadata*	*Any form elements that describe the actual request*	*not a coding element*
2.1	Study Title/Request	This is a brief summation of the request.	[S] Project Title; Research Question
2.2	Existing/ New Request	This element specifies if the request is new or a modification to an existing request	[S] Is this a new request or a modification to an existing report
2.3	Funding Source	This element is asking who is financially supporting the use of this data.	[S] What are your funding sources? [E] Will funds be used to pay subcontractors; do funding sources have restrictions on the use of the data collected for this project?
2.4	Request Purpose	Concerns the use of the data being request. For example will it facilitate an internal administrative report, research or preparatory for research, cohort/Clinical trial recruitment?	[S] Will the requested data be applied to any of the following areas? Non-research, Patient Care, Operations, Research, etc.
2.5	Request Type	This element specifies the degree of data access the user requires.	[S] Multiple Choice: Self-service, Super user [E] Study Design Consultation, Research Navigator
2.6	Data Sources	Any element that asks the user to specify the source of data, for example this maybe a particular database, or a particular clinical site where the user thinks the data may originate.	[S] Sources of data? (Text Box) [E] Sources of data? (Multiple Choice)
2.7	Data Element Specification	This element refers to any description of the medical data elements the requester is after.	[S] Describe the data you need. [E] What is your selection criteria, From what time period… What data fields do you need?
2.8	Recurring Requests	This element is specific to the frequency of data delivery. A clinical trial that submits a request to aid recruitment may wish to receive a weekly dump of potential matches.	[S] Is this a one-time request or recurring?
3.0	*Compliance*	*Form elemenst related to a compliance attribute, such as IRB, PHI, internal regulations, or documentation requirements*	*not a coding element*
3.1	Institutional review board (IRB)	If the request is research, this element request details on the IRB number or if the protocol is IRB exempt.	[S] IRB number
3.2	IRB Proof	Elements that require IRB proof	[S] Please upload your approved IRB protocol
3.3	Protected Health Information (PHI)	Regardless of request purpose, this element specifies HIPAA compliance and asks to what level of identified data (if any at all) are needed.	[S] Will the data be identified or de-identified [E] Please select the type of data you will need: identified, de-identified, limited decedent, aggregate counts…
3.4	Compliance Other	This element concerns any type of compliance attribute, whether it be IRB, PHI, internal regulations, or documentation requirements that could not be classified elsewhere	[S] Provide your consent (or waiver of consent)
4.0	*Data Use*	*Refers to how the requester is going to use or share the data*	*not a coding element*
4.1	Internal Data Sharing	This element represents how the user is sharing the data within their team, where the data is going to be stored, how the data is to be delivered, or the format of the data.	[S] Please describe data storage and use plan [E] Who will have access to the data, where the data is to be stored, data delivery & format
4.2	External collab-orators data use agreement (DUA)	If the requester is sharing the information with an external collaborator, is there a formal data use agreement	[S] Is there a DUA? [E] Name non-affiliated project team members that will have access to the data; upload DUA.
4.3	Public Sharing of Original Dataset	This elements refers to the intent of the requester to publish the original dataset	[S] Will data be made publically available? [S] Do you plan on making this data publically available, how so?
4.4	Terms and conditions of use	This element refers to any mention of terms and conditions the requester must agree to for the release of the data to them.	[S] Please read/agree to these terms and conditions for the use of this data.
4.5	Data Use Other	This element includes items that were not specifically covered in the other data use categories	[S] Who is your intended audience for data reporting?
5.0	*Miscellaneous*	*Form element that cannot be categorized elsewhere*	*not a coding element*
5.1	Elements not classified elsewhere	Items that did not fit into other categories.	[S] Is this an emergency request due to a grant deadline [S] Will you be contacting patients?

research questions helped guide the analysis. These included (1) what high-level organizational categories can data request form elements be assigned to? (2) what percentage of metadata could potentially be obtained from source systems without asking research teams to copy it to a form? (3) how are request form items distributed between administrative data (i.e., 'bookkeeping') and actual data requests? (4) how much detail does each element on a form seek to obtain from a user completing the form?

The theme lists from both reviewers were then compared, discussed, and consolidated into a single list. A third reviewer (CW) evaluated the merged list and refined it further. Finally, two reviewers (GH and DF) compared two randomly selected forms to finalize the codebook and address additional gaps in code coverage. Similar themes were then grouped into logical categories (e.g., "Compliance", "Data Use") and numbered. This final list served as our codebook, which is shown in Table 1.

We also utilized a 'comprehensiveness' measure to indicate the breadth of each element: Simple (S) or Extensive (E). Simple elements were related to a very focused, narrow question on a form (e.g., "Your Name"), whereas Extensive elements had a much broader scope. For example, an Extensive element asked the requestor to "indicate all identifiers (PHI) that may be included in the study research record", followed by a list of all 18 HIPAA identifiers with a checkbox next to each. Examples of Simple and Extensive elements with respect to the codebook are also shown in Table 1. Note that some elements (e.g., codes 1.4, 2.1, 2.2 in Table 1) were judged by the team to only be coded using a Simple 'comprehensiveness' measure; others could be either Simple or Extensive.

C. Form annotation by two annotators

Each data request form was divided into individual, granular form elements based on the questions asked on each form. For example, one of the forms had a single numbered question comprised of two sub-questions, (1) "describe the data security procedures" and (2) "who will have access to the data". These were split into two distinct elements for coding. Each data request form element was then entered into the Coding Analysis Toolkit (CAT; Texifter, Amherst, MA). The CAT provided the capability for each element to be shown to an annotator on a computer screen along with the codebook so that all elements could be reviewed and coded efficiently. Using the CAT, two annotators (GH and DF) independently reviewed and coded all of the data elements from each of the ten data request forms, including whether each was Simple or Extensive in terms of comprehensiveness. Inter-rater agreement for each form was assessed with the kappa statistic. Coding disagreements were then discussed between the two coders and code assignment consensus was reached.

D. Content analysis of the ten forms

From the coded elements on each form we estimated the completeness of information about data needs captured in each form. This was done by assigning a numerical score to each element in the code book based on the comprehensiveness measure (Simple=1, Extensive=3) that represented the maximum score each item could be assigned. Forms that had ≥ 3 Simple elements assigned to the same code were considered to have an Extensive comprehensiveness measure of that code by nature of having multiple elements covering the same concept. We then computed the percent coverage of all possible elements by summing the scores per form and dividing by the total number of possible points a theoretical, all-inclusive form would have had. Finally, we assessed the form elements coded with either code element 2.1 and 2.7 for their ability to capture the salient details that would likely be necessary to capture the context and content of data requests in a reliable manner, which may serve as a communication channel between biomedical research teams and data analysts.

Results

The primary results from our analysis are shown in Table 2. There was substantial variation in how much detail each form covered and in the elements that were covered. Based on our metric of coverage, the top three forms (A, C, and J) had coverage of 52%, 48%, and 48%, respectively. Form B was much more sparse with only 11% total coverage. In general, forms that had more overall elements (or individual questions) also had better coverage, but the relationship was not completely linear. For example, Form A with the highest percentage of coverage (52%) only had 15 total elements whereas form F had 19 total elements but only 35% overall coverage. This discrepancy was most often due to either the number of Simple versus Extensive elements used on a form (e.g., fewer elements, but

Table 2. Summary of the coding analysis performed on the ten data request forms. If a cell is shaded it means that the specific code (row) was found to exist in the specific form (columns A-J). Additionally, the comprehensiveness measure of each element is shown with either an S (Simple, light shading) or E (Extensive, dark shading); those with ≥3 Simple elements on a form related to a single code were assigned an 'E' label even if it was not originally coded as being Extensive. The "Max Score" column represents the total number of points a form element could be assigned as a representation of its comprehensiveness. The total coverage of all elements for each form is shown at the bottom of the table as both a sum and percentage. Note that cells with an Extensive comprehensiveness label were given a score of 3 and those with a Simple comprehensiveness label were given a score of 1. The "# Forms with element" column is a sum of the number of distinct forms that had at least one element on the form that had the respective code in it. For example, nine forms contained code 2.1 ("Study Title/Request").

Code	Description	Max Score	Form										# Forms with element
			A	B	C	D	E	F	G	H	I	J	
1.0	*Requester Metadata*												
1.1	Name	3	E				E		E	E		E	5
1.2	PI, supervisor, department head	3	E	S	E	E		E			E	E	7
1.3	Billing/Administrative content	3			E		S	S				E	4
1.4	Other	1		S		S					S		3
2.0	*Request Metadata*												
2.1	Study Title/Request	1	S	S	S		S	S	S	S	S	S	9
2.2	Existing/New request	1	S		S					S			3
2.3	Funding source	3			E			S			S	S	4
2.4	Request purpose	1	S	S	S		S	E*	S		S	S	8
2.5	Request type	3		S		E					S		3
2.6	Data sources	3	E					S		S			3
2.7	Data element specification	3	E					E	S	S	S	S	6
2.8	Recurring requests	1	S		S								2
3.0	*Compliance*												
3.1	IRB	1	S					S			S	S	4
3.2	IRB proof	1	S									S	2
3.3	PHI	3	E				E						2
3.4	Compliance other	3	S				E		S			E	4
4.0	*Data Use*												
4.1	Internal data sharing	3			E			S	E	S			4
4.2	External collaborators DUA	3			E							S	2
4.3	Public sharing of original dataset	1							E				1
4.4	Terms and conditions of use	1	S								S		2
4.5	Data use other	1								S			1
5.0	*Miscellaneous*												
5.1	Elements not classified elsewhere	3	S		E	S	S	E		S	E	E	8
	Total Score	46	24	5	22	8	13	16	13	10	14	22	
	Percent coverage of all possible elements	100%	52%	11%	48%	17%	28%	35%	28%	22%	30%	48%	
	Total number of distinct form elements identified for coding		15	5	25	10	9	19	11	11	21	36	

***** This was labeled Extensive because there were 4 distinct Simple elements related to category 2.4; however, this category was considered to be a Simple category. Thus, in this row it still only counts as 1 towards the total score.

more extensive coverage by each element) or due to many elements disproportionately being related to only a handful of related questions (e.g., one form had four elements dedicated to the funding source).

Nine out of the ten forms asked about the title of the study/request, and this was the most common question asked across the forms. Other questions were less commonly asked. Only two forms (A and J) explicitly requested proof of study approval from an institution review board, and only one form (G) asked if there was a plan to share the original data set publically. At a category level, four forms did not have a single element related to "Compliance" and three did not have a single element related to "Data Use". All forms incorporated at least one element related to the categories of "Requester Metadata" and "Request Metadata", the latter of which is most important for understanding the actual data needs for a request. Within the "Request Metadata", codes 2.1 ("Study Title/Request")

and 2.7 ("Data element specification") were determined to be the most relevant for a data analyst to understand the specific needs of the research team. Therefore, we list the specific elements for codes 2.1 and 2.7 derived from all 10 forms within Table 3. Some forms asked detailed questions (e.g., five distinct elements coded 2.7 on form F) whereas others asked very basic questions (one element coded 2.1 on form C).

During the coding process we also came across form elements that stood out from the rest, based on the unusual or interesting nature of the questions. These are detailed in Table 4. This table also contains descriptions based on our consensus opinion on why those specific elements were noteworthy. Overall, coding the forms was challenging due to the highly variable manner in which questions were worded. For the ten forms in our analysis, the initial Kappa scores measuring the inter-rater agreement were quite variable, ranging from 0.14 to 0.86 (full list for the forms in the order presented in Table 2: 0.83, 0.86, 0.57, 0.14, 0.64, 0.65, 0.52, 0.55, 0.43, 0.76). Thus, some forms required considerable effort to reach consensus on the final coding of each element.

Table 3. Data elements related to codes 2.1 ("Study Title/Request") and 2.7 ("Data element specification"). These two codes were judged to be the most relevant for a data analyst to understand the information needs of the research team. Note that form D did not contain any elements for which these codes could be applied.

Form	Code	Comprehen-siveness	Element Header Excerpt	Element Question	Element Options
A	2.1	S	General Reason for Request	Brief description of intent for use of data and/or associated project	Text Box
A	2.7	E	Research Request Reason	Please included as applicable: Request Information (Please include Request Description and if known) - Data Elements, Date Range/Parameters, Sort Sequence, Included Population (e.g. nursing units, DRG codes), Excluded Population (exceptions to the included population), Associated Form (Eclipsys Use Only)…	Docu-ment Upload
B	2.1	S	Please provide the following information	I need the new report because…	Text Box
C	2.1	S	Data Type	Full Study Title	Text Box
E	2.1	S	If the purpose of your request is for Patient Care, Education, Administrative, Billing/Payment…complete the following	Give a brief description of your project in the space below:	Text Box
F	2.1	S	DATA REQUEST FORM	Study Title/Study Idea	Text Box
F	2.7	E	Data and/or Records Needed for Research Protocol: Include the following…	Selection Criteria (e.g., all patients with a visit with an ICD-9 780.3x and/or 345.x, English speakers whose age > 50 and age <= 75, etc.)	Text Box
F	2.7	E	Data and/or Records Needed for Research Protocol: Include the following…	Counts (if applicable): (e.g., number of patients seen by Firm A, B, C grouped by under 65 and 65 or older)	Text Box
F	2.7	E	Data and/or Records Needed for Research Protocol: Include the following…	Dates of Records: (e.g., January 1, 2004 March 31, 2005)	Text Box
F	2.7	E	Data and/or Records Needed for Research Protocol: Include the following…	Number of Records: (e.g., 2000 patients with specified diagnosis, 10% sample of patients with diagnosis, all patients admitted thru ED)	Text Box
F	2.7	E	Data and/or Records Needed for Research Protocol: Include the following…	List of Data Fields: (e.g., age, race, diagnosis, service area, PCP, etc.)	TextBox
G	2.1	S	complete the following questions	Describe the project for which the data is requested:	Text Box
G	2.1	S	complete the following questions	What is the purpose of the project or study?	Text Box
G	2.7	S	complete the following questions	Describe the data elements needed, such as cancer type (site and histology), geographic location and dates…	Text Box
H	2.1	S	What are the objectives of this project?	What question(s) are you trying to answer?	Text Box
H	2.1	S	What are the objectives of this project?	What problem(s) are you trying to solve?	Text Box
H	2.7	S	What are the data requirements?	How much historical data are needed to meet the targeted reporting scope?	Text Box
H	2.7	S	What are the data requirements?	How current do the data need to be to support the targeted reporting?	Text Box
I	2.1	S	Project Details	Project Title	Text Box
I	2.7	S	Project Details	Please explain below and describe, in detail, the nature of your request to BMI/ICTR. Please do not include any protected health information (PHI)	Text Box
J	2.1	S	General Question	Protocol Title	Text Box
J	2.7	S	General Question	Anticipated Enrollment	Text Box
J	2.7	S	General Question	Is your anticipated enrollment period greater than a year	Y/N/NA

Table 4. Noteworthy atypical form elements grouped from different forms.

Element	Why noteworthy
"I want to write my own SQL queries"	Allows for the possibility of self-service of the complex databases for advanced users. It is unclear what type of guidance or oversight is provided for such requests.
"Please specify what type of Biomedical Informatics Services you are requesting: REDCap, Velos…"	This form combined questions related to data requests and those related to data storage.
""Will you be contacting patients? ____No ____Yes. If yes, please justify the need."	This form seemed to conflate the role of data request fulfillment with that of an institutional review board (IRB). A judgment about the appropriateness of contacting patients is generally handled within the framework of an IRB.
"Principal Investigator: Degree(s):"	It is unclear what the need is for the academic degrees of the principle investigator. It is possible that some institutions limit data access to investigators with a terminal degree.
"What question(s) are you trying to answer" "What problem(s) are you trying to solve"	These questions appear to be aimed at developing a broader perspective about the specific needs and goals of the research term. This information could be useful to help the analyst better understand the context for the data request.

Discussion

Our analysis of research data request forms revealed several interesting findings. Foremost was the substantial variability in the content and comprehensiveness of the forms. This variability suggests that there is no universal or community-based consensus, even among CTSA institutions, about the optimal way in which a data request form should be designed, what the 'right' questions to ask are, and how they should be asked (i.e, expecting simple or extensive answers). This could cause downstream consequences including an inability to meet regulatory requirements (e.g., no record of IRB approval verification) or an inability to track research data use in trustworthy ways, as well as problems developing the right queries to meet the fine-grained needs of research teams.

Our analysis raised the important question about how well overall the forms were designed. Being able to answer this question adequately depends, in part, on how well the forms could capture complex data needs accurately and in a reproducible manner. Some forms were very vague or brief about asking researchers what was needed, whereas others asked about specific elements (Table 3). Yet we did identify one form that contained questions that seemed to be aimed at helping the analyst develop a deeper understanding of what data were being sought (Table 4, row 5) and this may be a useful approach to improve communication.

Because data request forms might serve as the first point of contact between a data management team and a research team, improvement of these forms could provide great benefit. It has been shown that work focused on redesigning pathology test request forms has been beneficial,[34-36] so it may be reasonable to extrapolate that similar benefits could be achieved with redesigned data request forms. The process of developing appropriate data queries from complex user needs can take multiple rounds of refinement,[29] but current forms do not appear to be designed to support this process well. It has been noted in the literature that adequately meeting the data needs of investigators for a single request can take a long time[37] so any efficiencies that can be gained would be welcomed.

Data requests forms have been mentioned in the literature[32, 38] (often as a side note) but little attention has been paid to their role in helping investigators obtain data accurately and efficiently. Relative to other form elements, our analysis indicates elements used to elicit the context and content of the requester's data need are lacking. The utilization of frameworks such as PICO (problem/population, intervention, comparison, and outcome) might prove to be advantageous in this setting.[39, 40] With PICO, requesters are encouraged to structure the information need along each of the four dimensions, which could help convey a more realistic description of the request.

Additionally, the effectiveness of forms could likely be improved by providing additional education to investigators about the nature of the data in the systems while at the same time helping to guide researchers through the request form in a more logical manner to ensure that all the important aspects are covered. It has been observed that familiarity with the database fields by research teams is essential even when working with data analysts[41] but the forms we analyzed did not provide such details. It is possible that some of the forms we reviewed were meant to be accompanied by additional descriptive documents, but we did not come across them in our search. We also did not identify any forms that discussed the issues about data in coded format versus free text narratives, or what types of data are generally found in either of those types of sources.

The forms that comprised our analysis appeared to be constructed to meet the needs of multiple stakeholders (researcher, compliance, IT, etc.). What was surprising, however, is that many forms were unbalanced and placed a greater emphasis on capturing administrative (i.e., bookkeeping) data rather than on the details necessary to execute an effective data query. At the large academic centers generally funded by CTSAs, it is likely that many of these data elements already exist in electronic format in administrative databases and might not even need to be transcribed onto a form. Additionally, asking about the degrees of the principal investigator (Table 4, row 4), for example, may be a reflection of a data governance concerns; that is, trainees or temporary employees without terminal degrees may not be granted access to the data at some institutions.

Future work should seek to understand what are the core set of elements that elicit actionable information related to common data requests. To this end, a careful analysis of actual data requests in order to be able to map the type of data needs to appropriate elements on existing forms, or to create new form elements when needed. Understanding these needs is a first step towards developing solutions to meet those needs.[42] Cimino *et al.* recently described their work related to understanding complex queries to better develop data retrieval capabilities in the self-service tool BTRIS (Biomedical Translational Research Information System) in use at the NIH.[20] Their goal was to better empower users to obtain the needed data rather than having to rely on data analysts to retrieve the data for them. Several of their observations could likely also improve the design of data request forms, specifically the recognition that the requirements from users "included types of data, constraints on data, and data sets formed from inclusion from multiple data sources."[20]

In addition, future work should seek to quantify the time it takes to complete the elements on a data request form, and if there may be a reasonable tradeoff between form length and the subsequent quality and efficiency of the data extraction. Additionally, observing investigators as they fill out the forms could provide insights about what form elements may be confusing or ambiguous.

From our analysis we are able to make several recommendations about future data request form development: (1) more effort should be made to standardize the types of questions being asked across institutions; (2) whenever possible, forms should de-emphasize the collection of administrative metadata and expand the scope of elements related to the request itself; (3) despite decrease administrative metadata, forms should capture enough information to ensure that regulatory requirements about data use, privacy, and human subjects protection are being met; (4) form design should match the data requirements of investigators--since this is not well described, further research will be needed to elucidate these requirements; (5) because data requirements may vary based on the intended use (e.g., research versus administrative), a 'one-size-fits-all' form may not always be ideal, and forms customized to various use cases may be more effective; and (6) forms should provide at least a minimal level of detail to ensure that users understand the selections and options, including details about data sources and data types.

Conclusions

To serve people we must first understand them. A data request form is meant to be a tool to facilitate an understanding between data owners and data requesters, rather than a burden on researchers serving bureaucratic purposes. This analysis of research data requests forms revealed considerable heterogeneity in form content, both in the breadth and depth of the topics covered. Additionally, most forms over-emphasize the collection of administrative metadata and under-emphasize the collection of important details necessary to communicate a complex data request to a reporting team. Future work should focus on better understanding the content and nature of data requests from the perspective of multiple stakeholders to help inform the design of new data requests forms that can better capture the complexity of clinical and translational research teams.

Acknowledgements

This work was supported by National Library of Medicine grants R01LM009886 and R01LM010815, and by National Center for Advancing Translational Sciences grant UL1TR000040. The content is solely the responsibility of the authors and does not necessarily represent the official views of the supporting agencies.

References

1. Zerhouni EA, Alving B. Clinical and translational science awards: a framework for a national research agenda. Transl Res. 2006 Jul;148(1):4-5.
2. DesRoches CM, Charles D, Furukawa MF, et al. Adoption of electronic health records grows rapidly, but fewer than half of US hospitals had at least a basic system in 2012. Health Aff (Millwood). 2013 Aug;32(8):1478-85.
3. Friedman C, Rigby M. Conceptualising and creating a global learning health system. Int J Med Inform. 2013 Apr;82(4):e63-71.
4. Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. Sci Transl Med. 2010 Nov 10;2(57):57cm29.
5. Christensen T, Grimsmo A. Instant availability of patient records, but diminished availability of patient information: a multi-method study of GP's use of electronic patient records. BMC Med Inform Decis Mak. 2008;8:12.
6. Chute CG, Ullman-Cullere M, Wood GM, Lin SM, He M, Pathak J. Some experiences and opportunities for big data in translational research. Genet Med. 2013 Oct;15(10):802-9.
7. Sujansky W. Heterogeneous database integration in biomedicine. J Biomed Inform. 2001 Aug;34(4):285-98.
8 Yu C, Hanauer DA, Athey BD, Jagadish HV, States DJ. Simplifying access to a Clinical Data Repository using schema summarization. AMIA Annu Symp Proc. 2007:1163.
9. Smith SW, Koppel R. Healthcare information technology's relativity problems: a typology of how patients' physical reality, clinicians' mental models, and healthcare information technology differ. J Am Med Inform Assoc. 2014 Jan-Feb;21(1):117-31.
10. PCORnet: The National Patient-Centered Clinical Research Network. Found at http://www.pcori.org/funding-opportunities/pcornet-national-patient-centered-clinical-research-network/. Accessed on March 12, 2014.
11. Chute CG, Beck SA, Fisk TB, Mohr DN. The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. J Am Med Inform Assoc. 2010 Mar-Apr;17(2):131-5.
12. Greim J, Housman D, Turchin A, et al. The quality data warehouse: delivering answers on demand. AMIA Annu Symp Proc. 2006:934.
13. Hruby GW, McKiernan J, Bakken S, Weng C. A centralized research data repository enhances retrospective outcomes research capacity: a case report. J Am Med Inform Assoc. 2013 May 1;20(3):563-7.
14. Kamal J, Liu J, Ostrander M, et al. Information warehouse - a comprehensive informatics platform for business, clinical, and research applications. AMIA Annu Symp Proc. 2010;2010:452-6.
15. Lyman JA, Scully K, Harrison JH, Jr. The development of health care data warehouses to support data mining. Clin Lab Med. 2008 Mar;28(1):55-71, vi.
16. Wiesenauer M, Johner C, Rohrig R. Secondary use of clinical data in healthcare providers - an overview on research, regulatory and ethical requirements. Stud Health Technol Inform. 2012;180:614-8.
17. Zerhouni EA. Translational and clinical science--time for a new vision. N Engl J Med. 2005 Oct 13;353(15):1621-3.
18. Shurin SB. Clinical translational science awards: opportunities and challenges. Clin Transl Sci. 2008 May;1(1):4.
19. Cimino JJ, Ayres EJ. The clinical research data repository of the US National Institutes of Health. Stud Health Technol Inform. 2010;160(Pt 2):1299-303.
20. Cimino JJ, Ayres EJ, Beri A, Freedman R, Oberholtzer E, Rath S. Developing a self-service query interface for re-using de-identified electronic health record data. Stud Health Technol Inform. 2013;192:632-6.
21. Del Rio S, Setzer DR. High yield purification of active transcription factor IIIA expressed in E. coli. Nucleic Acids Res. 1991 Nov 25;19(22):6197-203.
22. Lowe HJ, Ferris TA, Hernandez PM, Weber SC. STRIDE--An integrated standards-based translational research informatics platform. AMIA Annu Symp Proc. 2009;2009:391-5.
23. Murphy SN, Weber G, Mendis M, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010 Mar-Apr;17(2):124-30.
24. Pennington JW, Ruth B, Italia MJ, et al. Harvest: an open platform for developing web-based biomedical data discovery and reporting applications. J Am Med Inform Assoc. 2014 Mar 1;21(2):379-83.
25. Weber GM, Murphy SN, McMurry AJ, et al. The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. J Am Med Inform Assoc. 2009 Sep-Oct;16(5):624-30.

26. Zhang GQ, Siegler T, Saxman P, et al. VISAGE: A Query Interface for Clinical Research. AMIA Summits Transl Sci Proc. 2010;2010:76-80.
27. Danford CP, Horvath MM, Hammond WE, Ferranti JM. Does access modality matter? Evaluation of validity in reusing clinical care data. AMIA Annu Symp Proc. 2013;2013:278-83.
28. Deshmukh VG, Meystre SM, Mitchell JA. Evaluating the informatics for integrating biology and the bedside system for clinical research. BMC Med Res Methodol. 2009;9:70.
29. Hruby GW, Boland MR, Cimino JJ, et al. Characterization of the biomedical query mediation process. AMIA Summits Transl Sci Proc. 2013;2013:89-93.
30. Brown PJ, Warmington V. Data quality probes-exploiting and improving the quality of electronic patient record data and patient care. Int J Med Inform. 2002 Dec 18;68(1-3):91-8.
31. Wakefield DS, Clements K, Wakefield BJ, Burns J, Hahn-Cover K. A framework for analyzing data from the electronic health record: verbal orders as a case in point. Jt Comm J Qual Patient Saf. 2012 Oct;38(10):444-51.
32. Gallagher SA, Smith AB, Matthews JE, et al. Roadmap for the development of the University of North Carolina at Chapel Hill Genitourinary OncoLogy Database--UNC GOLD. Urol Oncol. 2014 Jan;32(1):32 e1-9.
33. Post AR, Sovarel AN, Harrison JH, Jr. Abstraction-based temporal data retrieval for a Clinical Data Repository. AMIA Annu Symp Proc. 2007:603-7.
34. Durand-Zaleski I, Rymer JC, Roudot-Thoraval F, Revuz J, Rosa J. Reducing unnecessary laboratory use with new test request form: example of tumour markers. Lancet. 1993 Jul 17;342(8864):150-3.
35. Durieux P, Ravaud P, Porcher R, Fulla Y, Manet CS, Chaussade S. Long-term impact of a restrictive laboratory test ordering form on tumor marker prescriptions. Int J Technol Assess Health Care. 2003 Winter;19(1):106-13.
36. Henderson AR. The test request form: a neglected route for communication between the physician and the clinical chemist? J Clin Pathol. 1982 Sep;35(9):986-98.
37. Dattani N, Hardelid P, Davey J, Gilbert R. Accessing electronic administrative health data for research takes time. Arch Dis Child. 2013 May;98(5):391-2.
38. Jackson JH, Gutierrez B, Lunacsek OE, Ramachandran S. Better Asthma Management with Advanced Technology: Creation of an Asthma Utilization Rx Analyzer (AURA) Tool. P T. 2009 Feb;34(2):80-5.
39. Huang X, Lin J, Demner-Fushman D. Evaluation of PICO as a knowledge representation for clinical questions. AMIA Annu Symp Proc. 2006:359-63.
40. Schardt C, Adams MB, Owens T, Keitz S, Fontelo P. Utilization of the PICO framework to improve searching PubMed for clinical questions. BMC Med Inform Decis Mak. 2007;7:16.
41. Loke YK. Use of databases for clinical research. Arch Dis Child. 2014 Jan 31.
42. Natarajan K, Sobhani N, Boyer A, Wilcox AB. Analyzing Requests for Clinical Data for Self-Service Penetration. AMIA Annu Symp Proc 2013. p. 1049.

Evaluating health interest profiles extracted from patient-generated data

Andrea L. Hartzler, PhD [1], David W. McDonald, PhD[1], Albert Park, MS[2], Jina Huh, PhD[3], Charles Weaver, MD[4], Wanda Pratt, PhD[1,2]

[1]The Information School, [2]Biomedical and Health Informatics, University of Washington, Seattle, WA; [3]Telecommunication, Information Studies and Media, Michigan State University, East Lansing, MI; [4]CancerConnect, OMNI Health Media, Ketchum, ID

Abstract

Patient-generated health data (PGHD) offers a promising resource for shaping patient care, self-management, population health, and health policy. Although emerging technologies bolster opportunities to extract PGHD and profile the needs and experiences of patients, few efforts examine the validity and use of such profiles from the patient's perspective. To address this gap, we explore health interest profiles built automatically from online community posts. Through a user evaluation with community members, we found that extracted profiles not only align with members' stated health interests, but also expand upon those manually entered interests with little user effort. Community members express positive attitudes toward the use and expansion of profiles to connect with peers for support. Despite this promising approach, findings also point to improvements required of biomedical text processing tools to effectively process PGHD. Findings demonstrate opportunities to leverage the wealth of unstructured PGHD available in emerging technologies that patients regularly use.

Introduction

One of the most promising trends in health informatics is the rise of patient-generated health data (PGHD), ranging from patient-reported outcomes (PRO)[1-2] and observations of daily living[3] to quantified self[4] and qualitative illness narratives[5] collected outside of clinical care. Whereas traditional consumer health technologies focus largely on *pushing* resources to patients, emerging opportunities leverage existing PGHD for better care. Health information technology, including social media, provides a vital source of PGHD used as the basis of automated health profiling for targeted prevention[6] and treatment.[7-8] To promote social support in the context of online health communities, we leverage PGHD to automatically profile the health interests of online community members and then use those profiles to facilitate peer connections and support for cancer.[9]

Although PGHD has long been the prized treasure of online health communities,[10] it has become recognized as a critical tool for improving clinical care and population health by complementing traditional forms of data collected in the clinic.[1,11-12] For example, social support provided through narrative posts on online health communities promotes empowerment by improving psychological adjustment to cancer,[13] increasing social wellbeing, and helping patients feel better informed.[14] Electronic self-reported quality of life collected through structured PRO tools can reduce symptom distress and improve patient-provider communication.[15-16] Patient illness collected through a self-report tool was found to identify respiratory illnesses with greater sensitivity than chief complaint data used by traditional disease surveillance systems.[17] Growing patient engagement in health care highlights the important role that patient experience plays in policy, such as meaningful use criteria.[11,18] In particular, the Office of the National Coordinator for Health Information Technology initiated several activities to advance the application of PGHD in clinical workflows, research and development, and policy.[12]

Emerging technology (e.g., social media, mobile devices, sensors) bolsters the opportunity for using PGHD to profile the needs and experiences of patients, making this an important area for research and policy.[19] Thus, examining the validity of inferences extracted from PGHD is critical. Whereas progress has been made processing structured PGHD, such as PROs,[2] opportunities remain to leverage the wealth of unstructured PGHD in social media and other technologies that patients regularly use. Several studies benchmark the validity of inferences drawn from PGHD against clinical comparisons.[17,20] Yet few efforts examine the validity and use of health profiles extracted from PGHD from the patient's perspective.

Using automated text processing, we extract health-related terms from online community posts to summarize individual members' health-related interests.[21] Our long-term goal is to use the resulting **health interest profiles** to connect members with shared interests for peer support.[9] Our partnership with CancerConnect.com provides a unique opportunity to evaluate this approach with online community members. As a first step, we conducted a user

study to evaluate individualized health interest profiles with users based on their posts. In this work, we address two key research questions:

RQ1. How closely do health interest profiles extracted from PGHD align with members' stated interests?

RQ2. What are members' preferences for using health interest profiles to connect with peers for support?

Profiling users from PGHD in online health communities

Growth in health-related use of social media,[22] including online health communities, helps patients share experience and advice with peers (i.e., patient expertise).[23] Many individuals now use these tools more often to exchange information and advice than to obtain emotional support.[24] Yet, reading numerous posts to identify peers with shared interests takes time and effort. It can be difficult for community members to relate to the health experiences of other users[25] and build relationships that support rich exchange of patient expertise.[13]

Profiles about users and their interests provide a key means for exploring potential relationships in online communities. Most online communities encourage users to create a profile by manually entering a few key details, such as diagnosis or treatment. Detailed user profiles are invaluable for summarizing the experience and expertise of available from peers,[26] yet manual upkeep of detailed profiles takes time and energy away from managing a serious illness. We explore one possible solution that augments user profiles with details extracted automatically from community posts contributed by each user, such as treatments, tests, or other topics of interests.

Members of early online communities without user profiles were often limited to communicating their personal characteristics and interests through "signature line" descriptions at the end of message board posts. Today, user profiles are a fundamental component of modern social media that represent individual community members.[27] By aggregating distinctive features that characterize a user, a user profile provides a synopsis, typically through a combination of manually entered elements (e.g., personal interests) and semi-automatically generated elements (e.g., number of followers or friends). From these elements, users form "thin slice" impressions when establishing online connections.[28] Thus, user profiles help establish social context as conversation starters.[29]

In the health domain, some researchers enrich user profiles by dynamically leveraging PGHD. For example, Nuschke and colleagues[30] designed a community-based diet and exercise journal that dynamically illustrates progress towards health goals and community participation on user profiles. Similarly, profiles on PatientLikeMe.com summarize historical trends in PGHD that users post about their experience with treatments, symptoms, outcomes, and community participation with dynamic icons.[31] Temporal charts and graphs extend profiles to illustrate trends in these metrics over time. Although automatic extraction of PGHD can help users to build detailed profiles, this approach raises a number of questions about how machines might assist users–does automatic extraction produce more content than what users manually enter? How accurate is extracted content? Does automatic extraction capture interests and experiences that users do not otherwise enter?

In our work, we are enriching user profiles with health-related interests automatically extracted from a user's community posts.[21] The resulting health interest profile could efficiently summarize a user's experience through the health terms they discuss in posts as their community participation evolves over time. One could imagine extending such profiles with additional personal characteristics that members wish to share when forging connections with community members, such as demographics, education, livelihood, or connection to cancer as a patient, survivor, caregiver, or other role.[26] Despite the potential value of reducing the effort required for profile creation and maintenance, user perceptions about the accuracy and value of automated profile generation remain unknown.

Extracting Health Interest Profiles in CancerConnect Online Community

Within the context of CancerConnect (http://cancerconnect.com/), our partnering online health community, we examined the automatic extraction of individualized health interest profiles by processing the text of community members' posts. CancerConnect is an award winning resource for web-based cancer resources that facilitates peer support for cancer through forum-style community posts. Health interests profiles provide the basis for our broader effort aimed at peer matching to recommend "mentors" with shared interests.[9]

We developed a text extraction approach to automatically generate health interest profiles from online community text.[21] The profile is a vector of terms representing the health interests of an individual member based on all posts that user has contributed to the community. Our profile extraction pipeline includes MetaMap to support automatic extraction of health-related terms and sematic concepts that populate health interest profiles. MetaMap[31] is a natural language processing tool designed to extract health-related terms from biomedical text that map to concepts in the Uniform Medical Language System (UMLS).[33] The UMLS consists of more than 1.3 million concepts from over 100

biomedical vocabularies.[34] Each concept in the UMLS is classified into one or more semantic types.[35] Together, semantic types make up the UMLS Semantic Network that unifies the vocabularies within the UMLS, and thus, provide a means for grouping semantically similar terms. Because health terminology used by biomedical professionals can differ from the ways many patients express and think about health topics, researchers map patient-friendly terms to UMLS concepts in an effort to develop consumer health vocabularies (CHV).[36] Recent efforts include computer assisted updates that leverage social network data from PatientsLikeMe.com to identify new terms for inclusion in CHV.[37]

To generate a health interest profile for an individual community member, we collect all of the posts that member contributes to the community. We then process those posts to automatically extract health-related terms, which we refer to as "health interests." Since we wish to present those health interests to users, we were faced with a small dilemma. Do we present the user with an uncategorized set of extracted health interests? Or do we attempt to group health interests in some coherent way? We chose to group similar health interests using UMLS semantic types.[35] There are a number of strategies for grouping UMLS semantic types, such as maximizing semantic coherence.[38] Our aim was to present groups of terms relevant to the cancer experience that would be sensible to users who view health interest profiles. To do so, we considered a sample set of terms with their associated UMLS semantic types and created our own five categories that group similar semantic types (Table 1). Through this process, we populate a user's health interest profile with the health interests we extracted across the five categories (see example in Figure 1). This categorization enabled us to generate individualzed profiles that present information about uers in logical chunks, similar to user profiles they might find in any other online community.

Table 1. Health interest profile categories and associated UMLS semantic types

Category	UMLS Semantic type
Health problems	Neoplastic Process Disease or Syndrome Acquired Abnormality Injury or Poisoning Anatomical Abnormality Finding Sign or Symptom Pathologic Function Clinical Attribute Laboratory or Test Result
Treatments	Antibiotic Biomedical or Dental Material Medical Device Pharmacologic Substance Therapeutic or Preventive Procedure Hormone Vitamin
Diagnostics & tests	Diagnostic Procedure Research Activity Laboratory Procedure
Provider care	Health Care Activity
Genetics	Gene or Genome

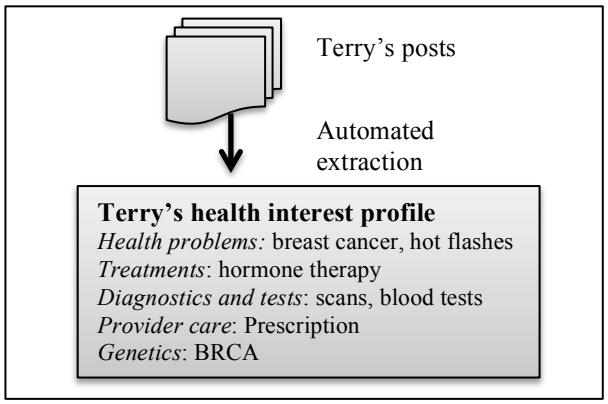


Figure 1. Example health interest profile populated with health interests extracted from community posts contributed by fictitious member "Terry"

We evaluate these individualized health interest profiles through a user study with community members. Our CancerConnect partnership enabled us to examine both the accuracy and perceived value of health interest profiles. We could generate health interest profiles for individual community members, and then ask those members to evaluate their own individualized profile. We were particularly interested in evaluating the accuracy of our automatically extracted health interest profiles compared to the stated health interests that community members could provide directly through manual entry (RQ1), as well as examining community members' perceived value of using those profiles to connect with other members for peer support (RQ2).

Methods

We conducted a web-based user study by recruiting members of CancerConnect to answer our two key research questions about the accuracy and perceived value of extracting health interest profiles from PGHD in online health communities. Before the study, we processed the text from all posts each community member contributed to the CancerConnect community. Using the extracted terms, we constructed individualized health interest profiles that summarize health-related issues each member discussed in their posts across our five categories (Table 1). To be

eligible for participation, members were required to have posted sufficient text to the CancerConnect online community over the past six months to extract a health interest profile with at least ten unique terms.

We used the individualized health interest profiles with participants during the two-part user study. To examine the alignment of health interests extracted from members' posts with members' stated health interests (RQ1), participants first completed a set of recall and recognition tasks during which they manually entered health issues of personal interest. Then, participants completed a set of preference ratings to describe their perceptions about using health profiles to disclose their personal interests and characteristics to other community members (i.e., "peers"), to learn about the interests and characteristics of peers in the community, and to be matched and interact with peers with shared interests and characteristics (RQ2). At the end of the study we collected participant demographics. IRB approval was granted from the University of Washington Human Subjects Division. We configured and administered the user study using Lime Survey (http://www.limesurvey.org/en/).

Part 1. Recall and recognition tasks

Each participant completed a sequential set of recall and recognition tasks to assess the accuracy of their individualized health interest profile. The participant first completed a free recall task in which they were asked to enter terms or phrases that describe the health issues they discuss within the community for as many of the five categories as they wished (i.e., health problems, treatments, diagnostics and tests, provider care, and genetics). Next in the recognition task, we primed the participant by showing the extracted health interests from their profile across the five categories. We asked the participant to add or remove terms and phrases until they were satisfied. Terms and phrases were shown as tags that could be easily added or clicked on to remove (Figure 2).

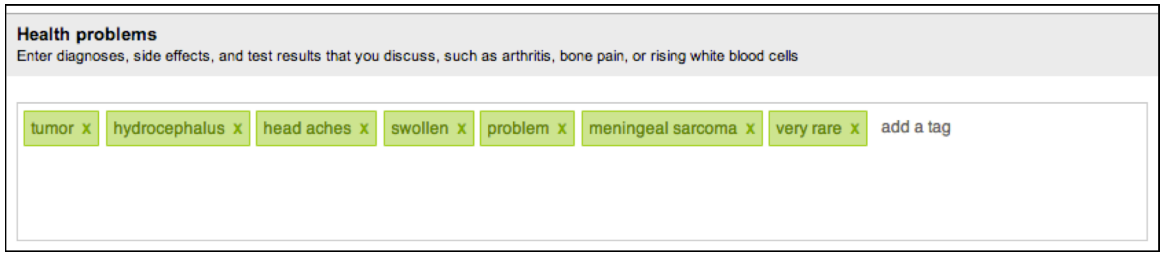


Figure 2. Example of recognition task to add or remove health interests for category "health problems"

At the conclusion of the recall and recognition tasks, we obtained three lists of health interests for each participant: (1) ***extracted terms*** making up the participant's health interest profile that we automatically generated, (2) ***recalled terms*** that the participant entered from memory during the recall task, and (3) ***recognized terms*** that resulted after the participant added and removed terms from their automatically generated health interest profile during the recognition task. We assessed the alignment of those lists of health interests using Wilcoxon sign rank to compare the number of terms between lists, as well as Jaccard index, precision, and recall to compare term similarity between lists. In particular, we examined the alignment between extracted terms from the health interest profile with the member's stated health interests through recall and recognition tasks (RQ1). This analysis enables us to examine how close a set of extracted terms can come to an acceptable set of profile terms. It also allows us to consider the similarity/differences between the health interests we can extract and the health interests users choose to enter.

Part 2. Preference ratings

Following the recall and recognition tasks, the participant provided structured feedback by rating their preference across a range of potential uses of their personal health profile (RQ2). We were particularly interested the perceived value of using profiles to share health interests with community members, as well as expanding profiles with other personal characteristics. Guided by our prior work,[25] we chose 10 personal characteristics which that could be used as the basis for peer matching and interaction including: all health interests, only common health interests, personality, participation level, demographics, education and livelihood, geographic location, common users followed, common groups followed, and common posts responded to. Preference ratings were made on a 5-point Likert scale (1="not at all" to 5="very") and covered four main areas: (1) ***comfort in disclosing*** personal interests and characteristics to other members of the online community on the profile, (2) ***interest in viewing*** personal interests and characteristics of other community members on their profiles, (3) ***importance of matching characteristics*** sought when connecting with online community members (e.g., someone with the same diagnosis or similar age), and (4) ***interest in interaction styles*** for connecting with online community members. Interaction styles included four types: (1) "one shot" anonymous interaction that is one-to-one, short-term, and impersonal, (2) one-to-one interaction that is personal, confidential, and sustained over time with a "buddy," (3) group interaction that

occurs regularly among individuals with a shared issue modeled after a "support group," and (4) group interaction that occurs as needed and focused on a topic of interest in a "group campaign."

We report preferences using descriptive statistics and comparisons based on Friedman chi square (X^2) and Wilcoxon signed rank (V) tests. We also offered participants the option to provide open-ended responses for concerns and suggestions regarding the use of health profiles to connect with community members, which we grouped qualitatively for emergent themes. This analysis enabled us to examine participants' attitudes regarding a range of potential uses of health profiles for sharing their health interests as well as broader personal characteristics for connecting with peers in the online community.

Results

Participants

A total of 34 CancerConnect members participated in the study with one participant completing only the recall and recognition tasks in part 1. The remaining participants ranged in age from 31 to 76 and are mostly female and white (Table 2). Table 3 shows participants' online community participation including average weeks worth of posts, posts per week, words per post, and extracted health interests per week.

Table 2. Demographics of participants

Age mean(sd) range	55(11) 31-76
Sex	85% Female 9% Male 6% na
Education	9% High school graduate 31% Some college 33% College graduate 24% Post graduate 3% na
Race/ethnicity	94% white 6% na
Social network size	24% Extensive 49% Moderate 18% Small 9% na
Geographic location	22% Western United States 15% Midwestern United States 27% South United States 27% Northeastern United States 9% na
Top personal interests/hobbies	Reading, exercise, cooking, gardening, television/movies

Recall and recognition: Profile validation

We report on alignment in the number and similarly of health interest terms we automatically extracted, terms participants freely recalled, and terms resulting after the participant completed the recognition task.

The number of terms that resulted following automated extraction, the recall task, and the recognition task are shown in Table 4. Compared to recalled terms, we extracted significantly more health problems (V=450, p=0.001) and treatments (V=444, p=0.004). Participants entered more terms than we extracted for diagnostics and tests, but that difference was not significant. Compared to what we automatically extracted, participants entered significantly more genetic terms and provider care terms (V=109, p=0.004). Automated extraction was limited for those two categories–we extracted provider care terms for 19 participants and extracted genetics terms for only three.

Table 3. Online community participation

	Mean (sd)	Range
Weeks worth of posts	32 (38.4)	0.1 - 129
Posts/week	3 (4.5)	0.1 - 14
% Initiating posts	20%	0 - 100%
% Replies	80%	0 - 100%
Mean words/post	105(55.0)	36 - 227
Mean extracted terms/week	13 (137)	0.1 - 49

In contrast, the difference in number of terms that resulted following automated extraction and the recognition task was much less striking. Participants removed an average of 2.4 health problems from the health interest profile, resulting in significantly fewer terms on the resulting recognized term list than we extracted (V=205, p=0.05). With the exception of treatments, participants added terms for remaining categories, resulting in more terms on the resulting recognized term list than we extracted for diagnostics and tests (V=21, p=0.02), provider care (V=34, p=0.008), and genetics (V=0, p=0.001). Table 5 shows numbers and examples of added and removed terms.

Table 4. Number of terms that resulted following the text extraction, recall task, and recognition task

	Extracted terms		Recalled terms		Recognized terms	
	mean (sd)	range	mean (sd)	range	mean (sd)	range
Health problems	13.15 (7.16)	3 - 26	6.94 (5.23)	1 - 23	11.74 (6.18)	2 - 26
Treatments	9.79 (7.80)	0 - 25	5.09 (3.82)	0 - 20	9.88 (7.41)	0 - 24
Diagnostics and tests	3.38 (3.23)	0 - 13	4.65 (4.85)	0 - 25	4.38 (3.61)	0 - 17
Provider care	1.82 (2.96)	0 - 13	4.15 (3.86)	0 - 17	3.00 (2.83)	0 - 12
Genetics	0.09 (0.29)	0 - 1	1.71 (1.73)	0 - 6	0.82 (1.14)	0 - 4

Table 5. Terms added and removed during recognition task

	Terms added		Examples of added terms	Terms Removed		Examples of removed terms
	mean (sd)	range		mean (sd)	range	
Health Problems	1.0 (1.8)	0 – 7	no side effects	2.4 (2.8)	0 – 10	alone, pain, HIV
Treatments	0.9 (2.0)	0 – 9	folfiri 5fu, nutrition	0.9 (1.2)	0 – 4	oxygen, procedure
Diagnostics and tests	1.4 (2.1)	0 – 7	ct scan, mri, blood test	0.4 (0.8)	0 – 3	hgb, research, color
Provider care	1.4 (2.1)	0 – 7	2nd opinion, exams	0.3 (0.6)	0 – 3	report, documented
Genetics	0.7 (1.1)	0 – 4	brca, family history	0.0 (0.0)	0 – 0	(none)

Findings on alignment of the number of extracted, recalled, and recognized terms demonstrate that automatic extraction can help users populate their profiles. This machine assistance was most effective for categories in which we extracted more terms (i.e., health problems and treatments). When given the opportunity, participants removed extracted terms they found inappropriate. Some removed health problems, such as "alone" and "clarity", point to limitations of extracting terms with MetaMap that do not seem health-related from the perspective of participants. Other removed terms were more clearly health-related, such as "pain" and "HIV," but participants chose not to keep them on their health interest profile. Categories with fewer extracted terms (i.e., provider care, genetics) required participants to add terms during the recognition task because extraction alone did not sufficiently populate their profile. Health problems that participants added, such as "no side effects", might be impossible to extract unless the user explicitly stated in a post. Thus the size of our categories varied and this had impact on our approach. These findings suggest that machines can help augment profiles through automated extraction, but that users should be provided the opportunity to edit extracted data. These findings also illustrate limitations of applying biomedical text processing tools to unstructured PGHD.

Term similarity between extracted terms and recognized terms was substantially higher than between extracted terms and recalled terms across all 5 categories of health interests (Table 6). With the exception of genetics, where terms were extracted for only 3 participants, overlap between extracted terms was recognized terms was substantial with Jiccard indices ranging from 0.48 to 0.79. When recognized terms served as the gold standard, the precision and recall of extracted terms was also high with worsening performance as categories become sparser moving from provider care to genetics. When we used recalled terms as the gold standard, overlap with extracted terms was much lower, including low precision and recall across all categories.

Table 6. Similarity between extracted terms and recognized terms, recalled terms

	Extracted (test) vs. Recalled (gold standard)		Extracted (test) vs. Recognized (gold standard)	
	mean (sd)	range	mean (sd)	range
Health problems				
Jaccard index	0.04 (0.04)	0.00 - 0.17	0.75 (0.20)	0.36 - 1.00
Precision	0.05 (0.06)	0.00 - 0.20	0.82 (0.19)	0.36 - 1.00
Recall	0.16 (0.22)	0.00 - 1.00	0.91 (0.16)	0.44 - 1.00
Treatments				
Jaccard index	0.08 (0.13)	0.00 - 0.60	0.79 (0.25)	0.00 - 1.00
Precision	0.13 (0.17)	0.00 - 0.75	0.89 (0.18)	0.40 - 1.00
Recall	0.26 (0.31)	0.00 - 1.00	0.89 (0.23)	0.00 - 1.00
Diagnostics & tests				
Jaccard index	0.03 (0.06)	0.00 - 0.20	0.61 (0.42)	0.00 - 1.00
Precision	0.08 (0.20)	0.00 - 1.00	0.80 (0.35)	0.00 - 1.00
Recall	0.04 (0.09)	0.00 - 0.33	0.65 (0.42)	0.00 - 1.00
Provider care				
Jaccard index	0.01 (0.03)	0.00 - 0.14	0.48 (0.44)	0.00 - 1.00
Precision	0.04 (0.13)	0.00 - 0.50	0.83 (0.28)	0.00 - 1.00
Recall	0.01 (0.04)	0.00 - 0.20	0.55 (0.48)	0.00 - 1.00
Genetics				
Jaccard index	0.00 (0.0)	0.00 - 0.00	0.17 (0.36)	0.00 - 1.00
Precision	0.00 (0.00)	0.00 - 0.00	1.00 (0.00)	0.00 - 1.00
Recall	0.00 (0.00)	0.00 - 0.00	0.17 (0.36)	0.00 - 1.00

Findings on term similarity further support our claim that participants appear largely satisfied with the accuracy of individualized health interest profiles we extracted and thus made few changes. Users are beginning to discuss emergent topics, such as genetics, which may be challenging to extract using tools like MetaMap. Such categories may be small with few terms, but that does not mean they are unimportant and users should be solicited to input terms. Further, lack of overlap between extracted terms and recalled terms suggests that users and machines might contribute different kinds of health interests to profiles. To further investigate this possibility, we compared the similarity

between recalled terms and the terms participants added during the recognition task (Table 7). The small term overlap and low precision and recall suggest that when participants are prompted with extracted terms, they add different kinds of terms than those they freely recall. This finding holds despite the lack of a washout period between sequential recall and recognition tasks. Our findings suggest a **valuable role for machines** to assist users not only in augmenting profiles with extracted terms that they are unlikely to recall and enter manually—by showing users the extracted terms, machines can also remind users of additional terms they are unlikely to recall on their own.

Table 7. Similarity between recalled terms & added terms

	Added terms (test) vs. Recalled terms (gold stand.)	
	mean (sd)	range
Health problems		
Jaccard index	0.03 (0.13)	0.00-0.67
Precision	0.03 (0.13)	0.00-0.67
Recall	0.15 (0.33)	0.00-1.00
Treatments		
Jaccard index	0.05 (0.16)	0.00-0.78
Precision	0.07 (0.24)	0.00-1.00
Recall	0.27 (0.38)	0.00-1.00
Diagnostics & tests		
Jaccard index	0.15 (0.29)	0.00-1.00
Precision	0.17 (0.31)	0.00-1.00
Recall	0.48 (0.45)	0.00-1.00
Provider care		
Jaccard index	0.03 (0.07)	0.00-0.33
Precision	0.07 (0.19)	0.00-1.00
Recall	0.15 (0.17)	0.00-0.50
Genetics		
Jaccard index	0.21 (0.36)	0.00-1.00
Precision	0.22 (0.36)	0.00-1.00
Recall	0.47 (0.49)	0.00-1.00

Preference ratings: Attitudes toward profile use

<u>Self-disclosure:</u> On average, participants expressed comfort in disclosing 10 types of personal interests and characteristics by publicly displaying them on their profile (Figure 3), with the greatest mean comfort expressed for disclosing common health interests and the least mean comfort disclosing geographic location. For participants without missing data (n=30), the difference in comfort level among types of personal characteristics was significant (X^2 = 19.7, p=0.02). Pairwise comparison between the highest level of comfort disclosing common health interests and lowest level of comfort disclosing geographic location shows a significant difference (V=93, p=0.005). Most participants (25/33) expressed no concerns about disclosing personal characteristics to other community members. When asked about concerns, 11 left their open-ended response blank and 14 responded with an explicit statement (e.g., no concerns). Concerns expressed by the remaining 8 participants included revealing personal information that could be used in unapproved ways (e.g., targeted advertising) (4/33), desire to choose with whom to share personal information (1/33), desire to keep personal information private (1/33), and creating stress (1/33) or embarrassment (1/33).

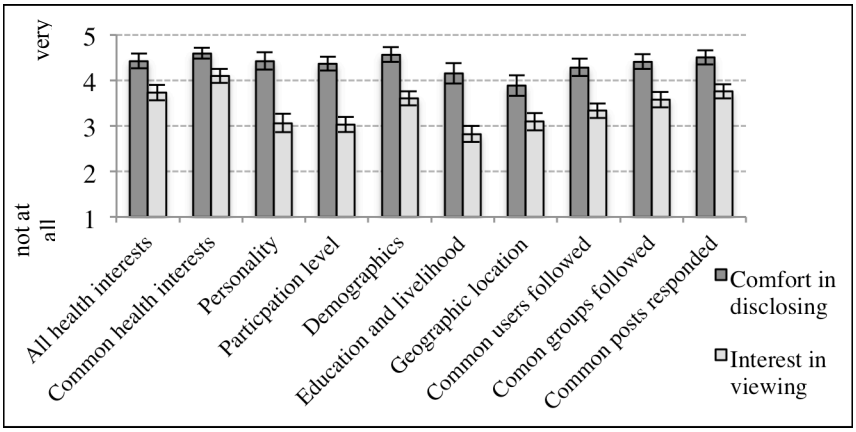


Figure 3. Comfort in disclosing vs. Interest in viewing personal characteristics

<u>*Viewing preferences*</u>: On average, participants expressed interest in viewing the 10 personal characteristics of other members on user profiles (Figure 3), but this interest level varied significantly among the 10 types (X^2=62 p<0.001). Participants expressed the greatest interest in viewing common health interests and the least interest in viewing education and livelihood. Pairwise comparison between the highest interest in viewing common health interests and lowest interest in viewing education and livelihood location shows a significant difference (V=210 p<0.001). Interest was significantly higher for viewing *common health interests* than any other personal characteristic except *common posts responded to*. Interest in viewing *education and livelihood* was significantly lower than other characteristics except *personality*, *participation level*, and *geographic location*. When compared to ratings for comfort disclosing personal characteristics, participants expressed greater comfort disclosing than interest in viewing all personal characteristics (Figure 3).

Matching preferences: When considering connecting with an online peer, participants rated the importance of a range of matching characteristics (e.g., *someone who…*). Figure 4 shows that on average participants rated some matching characteristics more important than others. For participants without missing data (n=19), this difference was significant (X^2=27, p<0.001). Matching characteristics rated highest include *someone who is* trustworthy, has experience with cancer as a patient or caregiver, and has a similar diagnosis. Characteristics rated least important include someone who lives nearby, who I already know, and who is of similar race/ethnicity. Pairwise comparisons between highest and lowest rated characteristics show significant differences at or below the 0.001 level.

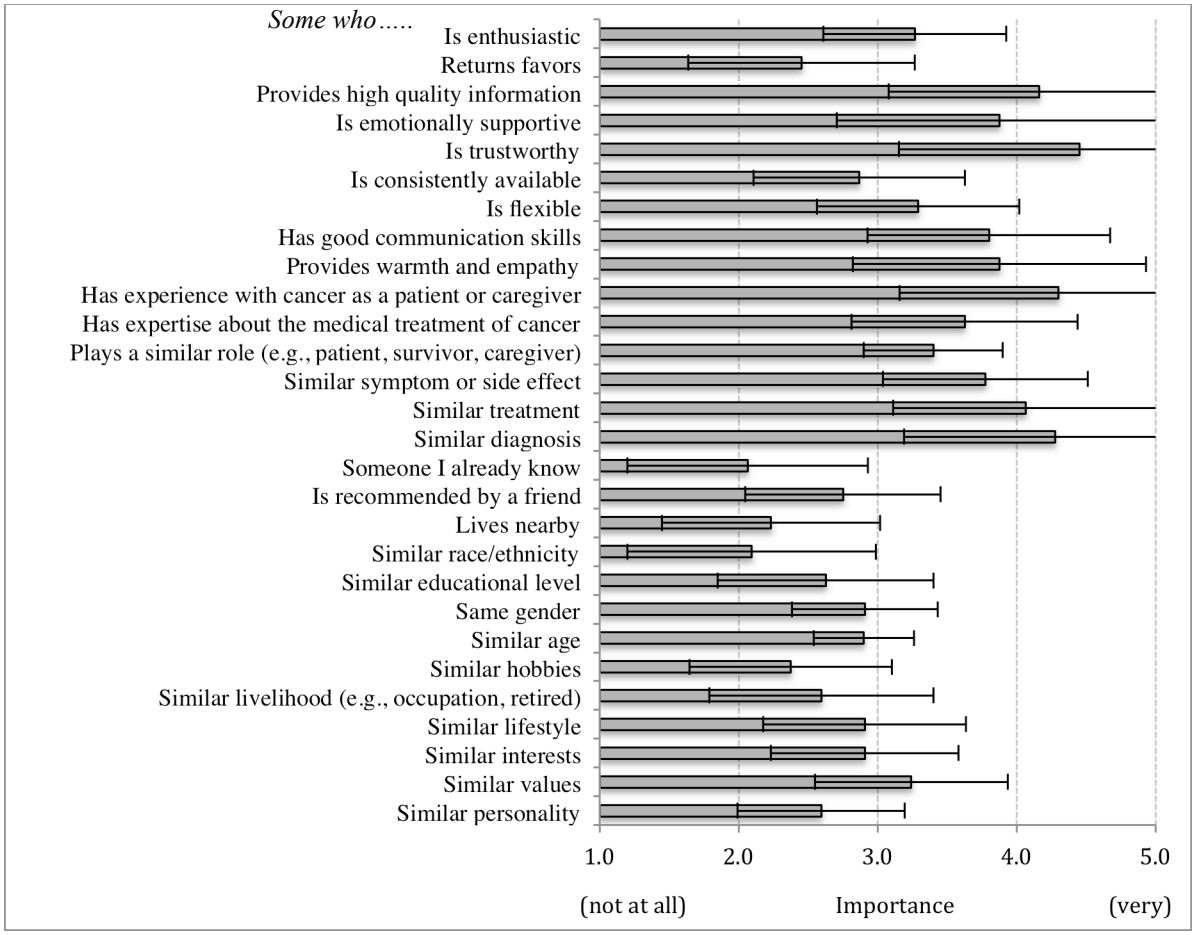


Figure 4. Mean importance of matching characteristics

Interaction preferences: On average, participants expressed interest in interacting with members across a range of styles (Figure 5). There was a significant difference in interest level among the four interaction styles (X^2=14, p<0.003). Pairwise comparisons show significantly less interests in "one shot" style, than other styles, including "buddy" style, "support group" style, and "group campaign" style. When asked about additional ways they wished to interact with online peers, participants reported email (33%), phone (25%), around specific topics (17%), through social media (17%), such as Pinterest or Blogger, and in disease-specific groups (8%).

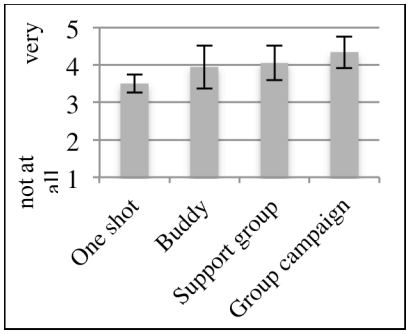


Figure 5. Mean interest level in interaction styles

Discussion and Conclusion

Substantial opportunities exist to leverage the wealth of unstructured PGHD available in emerging technologies that patients regularly use, yet few efforts examine the validity and use of health profiles extracted from PGHD from the patient's perspective. Although findings point to the perceived value of health interest profiles, we will evaluate their actual value in connecting peers in our future work. Findings from our user evaluation of health

interest profiles extracted from online community posts demonstrates value in augmenting detailed health profiles through text extraction applied to unstructured PGHD in multiple ways.

First, health interest profiles not only align closely with members' stated health interests, but expand upon those interests with little user effort. Automated extraction can populate profiles with content that community members accepted—when given the opportunity to add or remove health interests, few changes were made. Further, extracted content appears to overlap little with the content that members manually enter. This finding suggests that automated text extraction captures new and different kinds of interests and experiences than community members generally recall. In addition, extracted content appears to not only encourage members to refine their profile through manual removal of unsuitable content, but also act as a prompt for users to enter additional content they might not otherwise consider. Extraction is necessarily limited to the text users choose to post, which makes manual profile entry an important option for users. Thus machines can assist, but not necessarily replace, the user when processing PGHD.

Second, our findings illustrate positive attitudes of community members toward the use and expansion of health interest profiles with additional personal characteristics to connect with peers for support (e.g., common health interests). Community members expressed comfort in disclosing a number of personal characteristics to community members. They also expressed interest in viewing those characteristics of others. These findings provide support for expanding our automated extraction of detailed user profiles with additional characteristics of interest that can be used to facilitate peer matching and interaction. Important matching characteristics (i.e., someone who is trustworthy, has cancer experience, and has a similar diagnosis) and preferred interaction styles provide insight for future work in which we will use those profiles to match and help connect users for peer support. Although individuals with similar personal characteristics are likely to be attracted to each other (i.e., "birds of a feather flock together"), there may be merit in exploring "difference matching" (e.g., "opposites attract").

Although our findings provide new insight into the value of machine-assisted processing of PGHD, our text extraction was most effective when sufficient content, but not inappropriate content, was extracted. Finding this sweet spot poses a challenge. Whereas users can edit out irrelevant terms extracted by an overzealous machine, the inability of existing text extraction tools to capture important content is problematic. Topics such as provider care and genetics are clearly important to patients, but represent gaps in MetaMap and UMLS. Despite improvements in mapping UMLS to consumer-oriented terms,[36-37] effectively processing PGHD requires enhancements.

Emerging trends in PGHD present significant promise for shaping health care, self-management, population health, and policy. Our findings offer insights that speak to the value of processing PGHD from the patient's perspective. In particular, we illustrate one promising approach to leverage PGHD in the context of online communities. Substantial opportunities exist for capturing a wealth of unstructured PGHD available in emerging technologies that patients regularly use. As patient engagement in health grows and our desire to capture PGHD intensifies, it is critical that we prioritize development of technologies that can effectively process this unique and valuable resource.

Acknowledgements
We thank participants as well as Corey Shaffer, Alana Brody, and Meg Monday for community data and recruitment. This research was supported by National Science Foundation Smart Health & Wellbeing Award #1117187.

References

1. Wu AW. Advances in the use of patient reported outcome measures in electronic health records. Nov 2013. Retrieved Mar 3 2014 from: www.pcori.org/assets/2013/11/PCORI-PRO-Workshop-EHR-Landscape-Review-111913.pdf.
2. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. J Clin Epidemiol. 2010 Nov;63(11):1179-94.
3. Backonja U, Kim K, Casper GR, Patton T, Ramly E, Brennan PF. Observations of daily living: putting the "personal" in personal health records. 2012 Jun; 2012:6. eCollection
4. Choe EK, Lee NB, Lee B, Pratt W, Kietnz JA. Understanding quantified-selfers' practices in collecting and exploring personal data. Proc. SIGCHI Conference on Human Factors in Computing Systems (CHI '14), p.1143-1152.
5. Sharf BF, Vanderford ML. Illness narratives and the social construction of health. In: Thompson TL, Dorsey A, Parrott R, Mille K, editors. Handbook of health communication, Francis & Taylor e-library, 2008: p. 9-34.
6. Swan M. Health 2050: the realization of personalized medicine through crowdsourcing, the Quantified Self, and the participatory biocitizen. J Pers Med. 2012; 2(3): 93-118.
7. Wicks P, Vaughan TE, Massagli MP, Heywood J. Accelerated clinical discovery using self-reported patient data collected online and a patient-matching algorithm. Nat. Biotechnol. 2011 May; 29(5): 411–414.

8. Nakamura C, Bromberg M, Bhargava S, Wicks P, Zeng-Treitler Q. Mining online social network data for biomedical research: a comparison of clinicians' and patients' perceptions about amyotrophic lateral sclerosis treatments. JMIR. 2012;14(3):e90.
9. Hartzler A, McDonald D, Park A, Huh J, Pratt W. Mentor matching in peer health communities. Proc. AMIA 2012, p.1764.
10. Eysenbach G, Powell J, Englesakis M, Rizo C, Stern A. Health related virtual communities and electronic support groups: systematic review of the effects of online peer to peer interactions. BMJ 2004; 328: 1166.
11. Shapiro M, Johnston D, Wald J and Mon D. Patient-generated health data: White paper prepared for the Office of the National Coordinator for Health IT by RTI International. Apr 2012. Retreived Mar 3 2014 from: www.rti.org/pubs/patientgeneratedhealthdata.pdf.
12. Deering MJ. ONC Issue brief: Patient-generated health data and health IT. Office of the National Coordinator for Health Information Technology. Retrieved Mar 3 2014 from: www.healthit.gov/sites/default/files/pghd_brief_final122013.pdf
13. Hoey LM, Ieropoli SC, White VM, Jefford M. Systematic review of peer-support programs for people with cancer. Patient Educ Couns. 2008;70(3):315-37.
14. Van Uden-Kraan CF, Drossaert CH, Taal E, Seydel ER, van de Laar MA. Participation in online patient support groups endorses patients' empowerment. Patient Educ Couns 2009;74(1): 61-69.
15. Berry DL, Blumenstein BA, Halpenny B, Wolpin S, Fann JR, Austin-Seymour M, et al. Enhancing patient-provider communication with the electronic self-report assessment for cancer: A randomized trial. J Clin Oncol. 2011;29(8):1029-35.
16. Berry DL, Hong F, Halpenny B, Patridge AH, Fann JR, et al. Electronic self-report assessment for cancer and self-care support: Results of a multicenter randomized trial. J Clin Oncol. 2014 32(3):199-205.
17. Bourgeois FT, Porter SC, Valim C, Jackson T Cook EF, Mandl KD, et al. The value of patient self-report for disease surveillance. J Am Med Informatics Assoc. 2007 14(6): 765-771.
18. Ralston JD, Coleman K, Reid Robert J, Handley MR, Larson EB. Patient experience should be part of meaningful-use criteria. Health Affairs. 2010 29(4): 607-613.
19. Frost J, Okun S, Vaughan T, Heywood J, Wicks P. Patient-reported outcomes as a source of evidence in off-label prescribing: Analysis of data from PatientsLikeMe. J Med Internet Res 2011;13(1):e6
20. Bove R, Secor E, Healy BC, Musallam A, Vaughan T, Glanz BI, et al. Evaluation of an online platform for multiple sclerosis research: Patient description, validation of severity scale, and exploration of BMI effects on disease course. PloS one 2013 8(3): e59707.
21. Park A, Hartzler A, Huh J, McDonald D, Pratt W. Extracting everyday health interests from online communities. Proc. AMIA 2012, p.1889.
22. Fox S, Jones S. The social life of health information. Pew Internet & American Life (2009). Retreived Mar 3 2014 from: http://www.pewinternet.org/Reports/2009/8-The-Social-Life-of-Health-Information.aspx
23. Hartzler A, Pratt W. Managing the personal side of health: How patient expertise differs from the expertise of clinicians. J Med Internet Res 2011;13(3):e62
24. Sarasohn-Kahn, J. The wisdom of patients: Health care meets online social media. California HealthCare Foundation. 2008. Retireveid Mar 3 2014 from: http://www.chcf.org/documents/chronicdisease/HealthCareSocialMedia.pdf
25. Rimer BK, Lyons EJ, Ribisl KM, Bowling JM, Golin CE, Forlenza MJ, Meier A. How new subscribers use cancer-related online mailing lists. J Med Internet Res. 2005; 7(3): e32.
26. Civan-Hartzler A, McDonald D, Powell C, Skeels M, Mukai M, Pratt W. Bringing the field into focus: User-centered design of a patient expertise locator. Proc. Conference on Hum Fac in Comput Systems (CHI'10) 2010 p.1675-1684.
27. boyd dm, Ellison NB. Social networking sites: Definition, history, and scholarship. J Computer-Mediated Communication 2007;13:210-230.
28. Stecher K, Counts S. Thin slices of online profile attributes. Proc. ICWSM'08, 2008.
29. boyd dm, Heer J. Profiles as conversation: Networked identity performance on Friendster. Proc. HICSS-39 2006, 59c.
30. Nuschke P, Holmes T, Qadah Y. My health, my life: a web-based health monitoring application. In Extended Abstracts on Human Factors in Computing Systems(CHI '06). 2006, 1861-1866.
31. Frost J, Massagli M. Social uses of personal health information within PatientsLikeMe, an online patient community: What can happen when patients have access to one another's data. J Med Internet Res 2008 10(3):e15.
32. Aronson AR, Lang F-M. An overview of MetaMap: Historical perspective and recent advances. JAMIA 2010;17:229–36.
33. Humphreys B, Lindberg D, Schoolman H. The Unified Medical Language System: an informatics research collaboration. J Am Med Informatics Assoc 1998;5(1): 1-11.
34. Chen Y, Perl Y, Geller J, Cimino JJ. Analysis of a study of the users, uses, and future agenda of the UMLS. J Am Med Informatics Assoc 2007;14:221–31.
35. McCray AT, Nelson SJ. The representation of meaning in the UMLS. Methods Inf Med 1995;34:193-201.
36. Zeng QT, Tse T. Exploring and Developing consumer health vocabularies. J Am Med Informatics Assoc 2006; 13: 24–30.
37. Doing-Harris KM, Zeng-Treitler Q. Computer-assisted update of a consumer health vocabulary through mining of social network data. J Med Internet Res. 2011 May 17;13(2):e37. doi: 10.2196/jmir.1636.
38. Bodenreider O, McCray AT. Exploring semantic groups through visual approaches. JBI 2003; 36(6): 414-432.

Developing a Section Labeler for Clinical Documents

Peter J. Haug, MD[1,2]; Xinzi Wu; PhD[1], Jeffery P. Ferraro; PhD[1,2] Guergana K. Savova, PhD[3]; Stanley M. Huff, MD[1,2]; Christopher G Chute, MD[4]

[1]Intermountain Healthcare, Salt Lake City, UT; [2]University of Utah, Salt Lake City, UT; [3]Boston Children's Hospital and Harvard Medical School, Boston, MA; [4]Mayo Clinic, Rochester, MN

Abstract

Natural language processing (NLP) technologies provide an opportunity to extract key patient data from free text documents within the electronic health record (EHR). We are developing a series of components from which to construct NLP pipelines. These pipelines typically begin with a component whose goal is to label sections within medical documents with codes indicating the anticipated semantics of their content. This Clinical Section Labeler prepares the document for further, focused information extraction. Below we describe the evaluation of six algorithms designed for use in a Clinical Section Labeler. These algorithms are trained with N-gram-based feature sets extracted from document sections and the document types. In the evaluation, 6 different Bayesian models were trained and used to assign one of 27 different topics to each section. A tree-augmented Bayesian network using the document type and N-grams derived from section headers proved most accurate in assigning individual sections appropriate section topics.

Introduction

A key focus in Biomedical Informatics is the use of computerized patient data to impact medical care at the bedside. A variety of tools have been developed that consume electronic data and produce information that can be used by clinicians to gain insight and to direct diagnostic and therapeutic interventions. These include systems that organize and summarize clinical information as well as decision support applications that deliver alerts, suggestions, and otherwise support the delivery of consistent, high-quality care.

To be compatible with the tools that mediate these information interventions, data in the electronic health record is best managed in a structured form whose semantics are captured through reference to standardized medical terminologies. Unfortunately, a large subset of the clinical documentation in electronic medical records consists of free-text reports. A recurring challenge in the use of electronic medical data to support clinical care is the need to extract relevant medical facts from these clinical documents. This can be accomplished through natural language processing (NLP) technologies[1]. However, the use of these technologies is made more difficult by the heterogeneous nature of these documents. Not only do different types of documents focus on different clinical information, but documents are typically divided into sections each of which focuses on a different category of medical data.

For these reasons, an initial step in clinical NLP is to identify these sections and to label each with a concept code representing the principal *topic* of that particular section. Labels such as "History of Present Illness", "Family History", "Allergies", and "Discharge Disposition" are used to represent these topics.

Typically, at this stage in processing a document, an initial parse based on structural features results in a collection of document sections each consisting of a section header followed by one or more paragraphs. The text in these paragraphs reflects those clinical facts relevant to a particular medical topic. Unfortunately, while clinicians generally organize clinical documents using semantically similar component sections, they do not necessarily use a standardized collection of section headers to distinguish among these sections. The topics involved must be discovered through an initial processing effort. In the document collection we describe below, sections containing similar concepts are entitled with a surprising number of different headers. Both the text from each section's header and the text from within the section content can be used as input to a system that assigns a topic as a canonical label for each section. This system is called the "Clinical Section Labeler".

The algorithms described below are designed to assign a coded descriptor to each section. This descriptor tells what the section is about (i.e. the section's *topic*). Section topic labeling is the initial step in information extraction for many types of documents. When NLP systems are focused on extracting particular types of information, this information will typically be located in specific sections within the document. If each document is initially processed by a section labeler that can effectively assign a standardized topic to each section, further processing can be restricted to only those sections in which the required information is likely to be found. This can result in more accurate overall output since, for instance, an NLP system extracting current medications will look in the "Medications" section and not in the "Allergies" section, thereby avoiding mistakes and reducing processing effort.

The character of sections and section topics has been, to some extent, formalized in the HL7 Draft Standard for Trial Use (DSTU) describing the Consolidated Clinical Document Architecture (CDA)[2]. This standard describes approximately 60 different section types and their representation through section templates. Identifiers for these sections generally come from the Logical Observation Identifiers Names and Codes (LOINC) system[3].

The challenges and goals described above have been addressed. Several approaches to automatically assigning canonical section topics have been described in the literature. Denny, et al developed a complete system design to automatically identify section boundaries and to label the sections.[4] This system was designed to find sections with section headers as well as "implied" sections where section headers were absent. The algorithm used a combination of NLP techniques. These included spelling correction, Bayesian components (used to score section headers), and terminology-based rules.

In a more focused effort, Ying et al described a tool specifically for identifying sequences of section types[5]. They applied Hidden Markov Models (HMMs) to this task. They used simple bigrams as features and compared the HMM-based approach to models restricted to a naïve Bayesian algorithm. The version of the system based on HMMs proved significantly more accurate than the naïve Bayesian process.

Approach

In order to develop and test a system to assign topic labels to the sections of clinical documents, we began by collecting a group of medical reports. These reports were chosen to support training and testing for a section topic recognition system. This dataset consisted of 3483 clinical reports extracted from Intermountain Healthcare's Enterprise Data Warehouse (EDW). The distribution of the reports used is indicated in table 1.

Table 1: Reports used in Section Topic analysis.

Report Type	Number of Reports
Consultation Report	491
Discharge Summary	499
Operative Report	499
Surgical Pathology Reports	499
History & Physical Report	497
ED Physician/LIP Report	499
XR Chest 2 Views (Frontal/Lateral)	499

We used a combination of automated and manual annotation to break the text in these reports up into four categories. These were 1) section headers, 2) section content, 3) labels, and 4) values. Labels and values generally occurred as pairs such as labeled dates ("Date of Service: 12/5/2012"), although values occasionally appeared independently.

Combinations of section header and section content define the individual sections and were the targets for further analysis. Occasionally, headers did not have complementary sections. This typically occurred when a header was

followed by sub-headers each of which could then have independent content. In these cases, we collapsed all subordinate sub-headers and their content into a single instance of section content.

The sections defined this way were manually labeled with topic identifiers designed to represent the basic documentation goals of each of these report components. These named identifiers were expected to express the principal documentation goal of each section. To accomplish this, identical section headers were grouped together with links back to section content. A physician reviewed the section groupings using the section content as necessary to confirm membership of different headers in a semantic class. This effort provided the annotated dataset.

Section Topic Identifiers

Once we had identified the topics that we would target, we developed and compared a group of algorithms designed to assign an appropriate topic to each combination of section header and section content found. These systems were alike in two aspects: each used features generated through extraction of N-grams (uni-grams, bi-grams, and tri-grams) from the text of the section headers and/or section content and each employed models for identifying topics which were constructed using Bayesian-network-based approaches[6]. These Bayesian networks (BNs) were designed to use both the document type and the generated N-grams in assigning a topic for each section.

The Bayesian models differed in two aspects. First, for two of these models, N-grams were derived exclusively from the text of the section content in each section's header/content pair; for two models, N-grams were derived exclusively from the text of the section header in each section's header/content pair; and for two models, N-grams were derived from a combination of the text in both the header and content of each section.

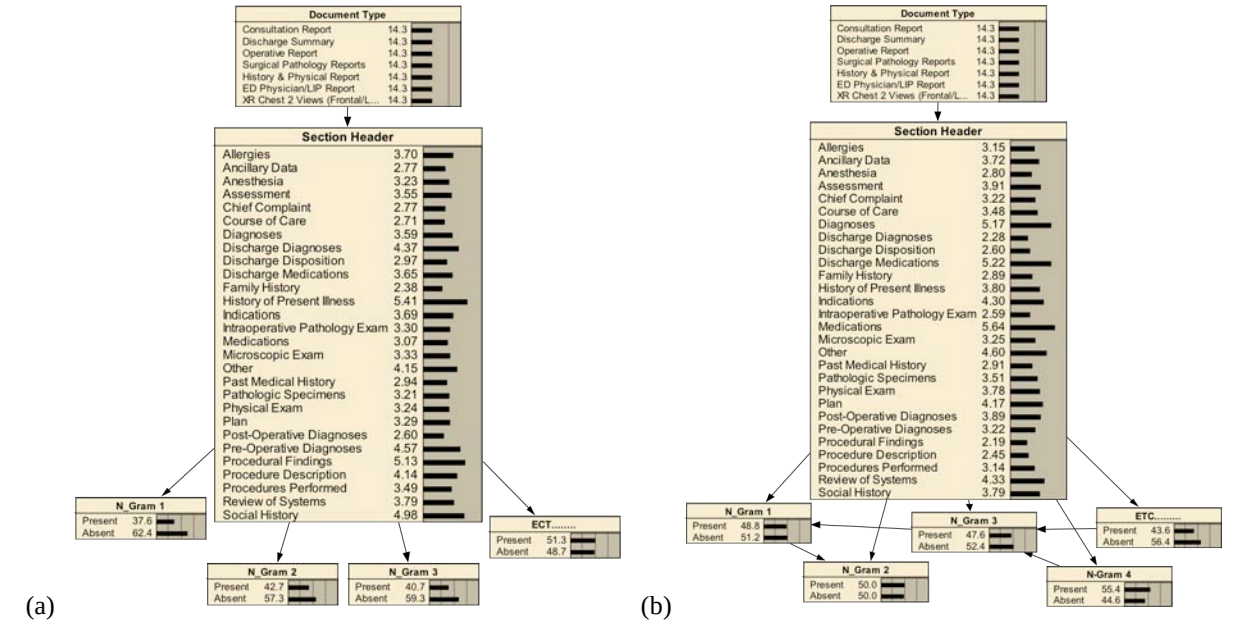


Figure 1: Bayesian networks were used to recognize section topics. In three cases (a), the core of the network employed a naïve Bayesian model. N-grams generated from 1) the section headers, 2) the section content, and 3) a combination of header and content were used to detect the topics of the individual sections. In the three additional cases (b), the network used these same feature sources but applied a tree-augmented naïve (TAN) Bayesian model. In all cases, the network was extended by adding a node representing the document type being processed.

The second way in which these models differed was that three were produced using an extended version of the naïve Bayes paradigm and three were developed using an extended version of tree-augmented naïve (TAN) Bayesian networks[7]. In developing these networks we extended the standard Bayesian paradigm by adding a node

representing the document type. This allowed for the differing distributions of section topics in the various document types to be easily captured by the algorithm. Table 3 (below) indicates the different algorithms tested.

The use here of Bayesian network techniques reflects our continuing interest in this technology. We have previously used BN-based systems in a number of NLP experiments. These include systems to extract findings from chest x-ray reports[8,9], systems to extract interpretations from ventilation/perfusion lung scan reports[10], and tools for syndromic detection[11]. A description of a system (MPLUS) that uses this approach to NLP can be found in Christensen et al[12]. Our enduring enthusiasm for this technology reflects its ability to incorporate information from multiple linguistic sources and to provide graphical tools through which to develop and inspect the resulting semantic models.

Analysis

We analyzed the different algorithms described above by employing a 10-fold cross validation technique. In this approach, the sample was divided into 10 parts, and for each part, the algorithm was trained with 90% of the sample and tested with the remaining 10%. The results are aggregated at the end of the procedure.

In this testing we looked at three different measures of parsing accuracy. We calculated the F-statistic and the standard one-versus-all area under the ROC curve (AROC) (for each topic, the analysis compares it to the combination of all competitors). However, when choosing a topic with this method, we actually choose the most likely topic from a list of 28. Therefore, we also calculated the pair-wise AROC[13] designed for multiclass classification problems.

Results

Within the initial dataset, 27,645 sections were identified. This had been reduced from 40,441 sections by the collapse of sub-sections into the parent sections. Ninety-eight different topics were assigned during annotation. These ranged in frequency from two instances (Nutrition Status, and Post-Operative Course) to 2933 instances (History of Present Illness). We determined to focus our efforts on topics with a frequency greater than or equal to 1% in the total corpus. Those topics whose frequency was less than 1% were collected into a category called "Other". This reduced the number of distinct topics to 27 plus the broad category of “Other”. These are listed in table 2.

In this way we narrowed the number of distinct section topics for consideration to 27. As anticipated, a review of the documents showed that these semantic labels were associated with a broad range of section header text. The 27 topics corresponded to 584 different text strings used as headers for sections from these clinical documents. For example, the single topic, “Medications”, was assigned to 1355 different sections; 53 different textual representations were found among these section headers ranging from “ADMISSION MEDICATIONS” to “She has been taking the following medications:” Across the range of topics we noted high variability in the headers produced during routine medical documentation. Interestingly, although it was not our intent, 23 of the 27 topics lineup well semantically with the section types described in the Consolidated CDA DSTU.

The goal of the Clinical Section Labeler is to use the N-grams from the section headers and content to appropriately label each section with a topic representing a principal category for the medical facts represented within that section. To accomplish this, we developed a tool that 1) extracted N-grams from section content and/or section headers, 2) executed a feature selection algorithm to identify subsets of N-grams able to discriminate among the different section topics, and 3) trained the Bayesian models described above to identify the appropriate topic for a section from among the 28 choices. The feature selection algorithm applies the Chi-square statistic to identify and discard the subset of features whose contribution to topic assignment is anticipated to be minimal. Researchers can inspect the distribution of Chi-squares and choose a threshold that will exclude irrelevant features.

Inspection of the N-grams generated by processing the section headers indicated that all useful information for topic detection could be gleaned using the top 800 N-grams. However, for topic detection using text from the section content or the text from the combination of section headers and contents, the number was larger. Over 1 million N-grams were generated from these sources and Chi-square testing suggested that a large subset of these could

contribute to topic detection. We chose to use the strongest 4000 of the N-grams produced for the topic identifier based on the section content text and the strongest 3000 of the N-grams produced for the identifier designed to process the combined section header and content. (Initial efforts with the combination of header and content N-grams (approximately 5000 total N-grams) proved disappointing; reducing the combined number appeared to give better results.)

Table 2: Section topic selected for analysis.

Section Concept	Case Count
Allergies	1269
Ancillary Data	1189
Anesthesia	417
Assessment	1865
Chief Complaint	1286
Course of Care	407
Diagnoses	817
Discharge Diagnoses	538
Discharge Disposition	418
Discharge Medications	378
Family History	934
History of Present Illness	2933
Indications	341
Intraoperative Pathology Exam	502
Medications	1355
Microscopic Exam	448
Other	2319
Past Medical History	2020
Pathologic Specimens	564
Physical Exam	1455
Plan	703
Post-Operative Diagnoses	473
Pre-Operative Diagnoses	695
Procedural Findings	663
Procedure Description	538
Procedures Performed	723
Review of Systems	1180
Social History	1215

Six section topic identifiers were built and tested using the 10-fold cross validation procedure described above. They ranged in raw accuracy from 61.77% to 98.96%. Table 3 displays results from these 6 different models. Shown are their accuracy (percentage of section topics correctly classified), the one-versus-all area under the receiver operating characteristic curve (AROC), the pairwise AROC, and the F-measure.

Based on these results, it appears that, for this population of documents, processing the section header with an extended, tree-augmented naive Bayesian model is most likely to provide an appropriate topic for the section. However, the measures reported are averages (weighted for AROC and F-measure) across all topics. For a tool of this sort, one hopes for consistent accuracy across the range of topics included in the model.

Table 3: Preliminary results of analysis of 6 topic-identification models.

Topic Identification Algorithms	Accuracy	AROC	Pairwise AROC	F-Measure
Naïve BN/Header Only	95.83%	0.9986	0.9992	0.9590
Naïve BN/Content Only	61.77%	0.9317	0.9551	0.6245
Naïve BN/Header + Content	84.08%	0.9801	0.9869	0.8400
TAN BN/Header Only	98.96%	0.9996	0.9997	0.9869
TAN BN/Content Only	67.55%	0.9578	0.9714	0.6792
TAN BN/Header + Content	90.90%	0.9911	0.9950	0.9116

Therefore, to further characterize the accuracy of the TAN BN using header text only, we evaluated its accuracy across the 27 individual topics (plus "Other"). Table 4 shows the most ("Pathologic Specimens") and least ("Course of Care") accurate topics identified with the TAN BN-based model using features from the section headers alone. We had anticipated that "Other" would be the least accurately identify topic, but were mistaken as indicated in the table.

Table 4: Statistics for the most and least accurate topics identified using the TAN BN model and the section header text.

Topic	Recall	Precision	F measure (95% confidence Intervals)	AROC (95% confidence Intervals)
Pathologic Specimens (Most accurately detected topic)	0.9982	1.0	0.9991 (0.9966, 1.0)	1.0 (1.0, 1.0)
Course of Care (Least accurately detected topic)	0.9165	0.9739	0.9443 (0.9266-0.9607)	0.9999 (0.9998-0.9999)
Other	0.9621	0.9339	0.9477 (0.9411, 0.9539)	0.9977 (0.9968, 0.9986)

Discussion

The documents used in this evaluation represent a typical collection of the kinds of reports produced when clinicians operate in a flexible authoring environment. Transcription and dictation, speech recognition, and manual authoring with and without templates all played a part in creating the medical documentation represented here. We expect a high degree of variability in the *content* of many of the sections that appear in medical reports. Patient characteristics are highly variable and this will be reflected in descriptions of their conditions.

However, we had originally hoped to take advantage of consistent section headers to help us find those locations in the document where specific types of information can consistently be located. Unfortunately, section headers also showed wide variability. We therefore chose to treat them like other medical concepts, which must be derived from strings of text in medical documents.

In electronic record systems where clinicians compose their text within a standard report template, section headers can be restricted to those provided by the template system. In these cases, where clinicians used standardized headers for sections, the challenges of automated section topic recognition are reduced. But in EHRs with flexible authoring systems, different wordings and formats occur frequently for section headers. Yet these variable representations can still be mapped to a common set of underlying medical concepts. This is the goal of the Clinical Section Labeler: to assign topics to sections that guarantee the semantic character of similarly labeled sections to be

consistent from report to report. Information about medications is to be grouped in a “Medications” section, and information about physical exam or past medical history tends to be found in “Physical Exam” or “Past Medical History” sections.

The Clinical Section Labeler succeeds in this to a degree. However, a review of the documents in this collection suggests that this modeling effort failed to accommodate a valuable group of document components. Many of the documents sections contained subsections, each with its own sub-heading. The initial semi-automated annotation tagged these sub-headings as headings, but then appropriately generated a pointer back to the relevant parent section heading. During the initial annotation, commonly encountered subsections (such as those representing the components of the physical exam) were assigned their own topic labels. The goal was to be able to independently identify subsections such as "Cardiovascular Exam" or "Eye Exam". However, subheadings and subsections were used erratically in many of the documents. For example, in some of the documents discharge diagnoses were used as subheadings and the subsections were descriptions of the evaluation and course of the individual diseases. As a result we decided to focus on labeling top-level sections. Future efforts will need to accommodate more complex document models where subheadings may be either identifiers for the subcomponents of a typical section or may represent individual clinical conditions used as alternative organizing foci within the documents.

Additional observations include the following:

- N-grams have limitations in large and complex document collections. We have anticipated this and plan to approach section labeling with other feature generation techniques in the future. None-the-less, we continue to find N-gram-based feature generation useful as an initial, brute force technique for configuring NLP systems.
- The use of Bayesian networks looks promising. In this experiment, we were able to develop both naïve Bayesian and TAN Bayesian models from our annotated data with relative ease. Extending the model with a network node reflecting document type was also simple. In the future, additional opportunities to extend the model are available. These include incorporation of new ways to combine information from the section headers and content using a different BN structure or, perhaps, adding an HMM component to the model to take advantage of the typically consistent sequences of section topics seen in medical documents.
- Our current section topic annotations will require revision. The current collection of topics represents an initial categorization driven in part by inspection of the documents extracted from our enterprise data warehouse. The goal was to investigate the existing “wild-type” section authoring process. In future work, we will refine this system. The focus will be on an organization of these concepts in the way that best supports extraction of key clinical information for specific care delivery and research activities. The section topics suggested as a part of the Consolidated CDA will help guide this revision. We envision capturing this organization in an ontology that can assist in future natural language processing efforts.
- Any approach that standardizes the use of section headers will ease the section labeling problem. Our next generation of EHR tools is expected to allow us to standardize many of our templates for collecting this data. None-the-less, research that wishes to exploit the several decades worth of collected reports in our EDW will continue to face the challenge of variable document structure and section heading expression.
- The reference standard for this project may have introduced an element of bias into the analysis. Not all of the 3483 documents were read through by the annotator. Instead, texturally identical section headers were grouped and examples of section content were reviewed to make sure the topic assigned was consistent with section semantics. It is apparent that some section content in fact belonged to different semantic categories than the header would indicate. Indeed, this is frequently seen when, for instance, elements of the “Social History” or “Family History” are included in the “History of Present Illness” due to their apparent relevance to the patients presenting complaint. To the extent that this occurs, it may explain the reduced accuracy of section identifiers that include section content. Another contributor to this reduction in accuracy is the huge number of n-grams generated for the section content. The Bayesian algorithms could

accommodate only a few thousand features, whereas accurate assignment of these sections to their semantic categories would have required tens of thousands or perhaps hundreds of thousands of N-gram-based features.

Conclusion

In this report, we have focused on tools to identify the semantic character of sections commonly found in medical documents. We describe this as assigning "topics" to these sections. The technology tested appears promising for this task and can be leveraged for other recognition tasks in natural language processing as well. We will continue to refine it and to study other approaches appropriate to semantic labeling tasks.

This work focuses on achieving accuracy in assigning topics to previously identified document components. In the future, we will embed this technology into a system designed to completely automate the identification of report components. This system will use both the text of medical documents and local document formatting characteristics to locate section headers and content. Subsequently, the tools described here will assign topics to these sections. Further processing of these labeled documents can then take advantage of an automatically generated document map to determine where relevant clinical information might be found. The ability to focus targeted natural language processing in those sections where relevant information is likely to be found should assist us in developing natural language processing systems that are both efficient and accurate.

The information extracted from medical documents has substantial value. It can contribute to research into the character and course of human illness and, in the future, will inform decision support systems capable of participate in clinical decision making at the bedside. We hope and expect that tools designed to identify sections in clinical text will help to realize the benefits of natural language processing systems.

This research was made possible by funding from the Strategic Health IT Advanced Research Projects (SHARP) Program (90TR002) administered by the Office of the National Coordinator for Health Information Technology.

References

1. Nadkarni PM1, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. J Am Med Inform Assoc. 2011 Sep-Oct;18(5):544-51.
2. HL7 Implementation Guide for CDA® Release 2: IHE Health Story Consolidation, DSTU Release 1.1. Draft Standard for Trial Use. Health Level 7: July 2012.
3. http://loinc.org.
4. J. Denny, A. Spickard, K. Johnson, N. Peterson, J. Peterson, and R. Miller. Evaluation of a method to identify and categorize section headers in clinical documents. Journal of the American Medical Informatics Association, 16(6):806, 2009.
5. Ying Li, Sharon Lipsky Gorman, and Noémie Elhadad. Section Classification in Clinical Notes Using a Supervised Hidden Markov Model. 2010. ACM International Health Informatics Symposium (IHI), pp. 744-750. Washington, DC.
6. Pearl J. Probabilistic reasoning in intelligent systems. San Francisco: Morgan-Kaufmann, 1988.
7. Friedman N, Geiger D, Goldszmidt M. Bayesian Network Classifiers. Machine Learning 1997;29:131–63.
8. Chapman WW, Fiszman M, Chapman BE, Haug PJ. A comparison of classification algorithms to automatically identify chest X-ray reports that support pneumonia. J Biomed Inform 2001 Feb;34(1):4-14.
9. Fiszman M, Chapman WW, Aronsky D, Evans RS, Haug PJ. Automatic detection of acute bacterial pneumonia from chest X-ray reports. J Am Med Inform Assoc. 2000 Nov-Dec;7(6):593-604.
10. Fiszman M, Haug PJ, Frederick PR. Automatic Extraction of PIOPED Interpretations from Ventilation/Perfusion (V/Q) Lung Scan Reports. Proceedings of the 1998 AMIA Annual Fall Symposium, pp. 860-864.

11. Chapman WW, Christensen LM, Wagner MM, Haug PJ, Ivanov O, Dowling JN, Olszewski RT. Classifying free-text triage chief complaints into syndromic categories with natural language processing.. Artif Intell Med. 2005 Jan;33(1):31-40.

12. Christensen L, Haug PJ, Fiszman M. MPLUS: a probabilistic medical language understanding system. Proceedings of the 40th Annual Meeting of the ACL (Association for Computational Linguistics) 2002.

13. Hand DJ, Till RJ. A simple generalization of the area under the ROC curve for multiple class classification problems. 2001 Machine Learning: 45, 171-186.

Security Concerns in Android mHealth Apps

Dongjing He, Muhammad Naveed, Carl A. Gunter, Klara Nahrstedt
Dept. of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL

Abstract

Mobile Health (mHealth) applications lie outside of regulatory protection such as HIPAA, which requires a baseline of privacy and security protections appropriate to sensitive medical data. However, mHealth apps, particularly those in the app stores for iOS and Android, are increasingly handling sensitive data for both professionals and patients. This paper presents a series of three studies of the mHealth apps in Google Play that show that mHealth apps make widespread use of unsecured Internet communications and third party servers. Both of these practices would be considered problematic under HIPAA, suggesting that increased use of mHealth apps could lead to less secure treatment of health data unless mHealth vendors make improvements in the way they communicate and store data.

1. Introduction

The mHealth trend is evident: as of March 2013, Research2Guidance reported that there were about 97,000 mHealth apps across 62 app stores[1]. According to a report from MarketsandMarkets, the global mHealth market is predicted to grow from $6.21 billion in revenue in 2013 to $23.49 billion by 2018 at a compound annual growth rate (CAGR) of 30.5 percent over the five-year-period from 2013 to 2018. The mobile fitness and wellness market is expected to grow at a CAGR of 36.7 percent from 2013 to 2018[2]. This rising mHealth market threatens changes in the way significant amounts of health data will be managed, with a paradigm shift from mainframe systems located in the facilities of healthcare providers to apps on mobiles and storage in shared cloud services. This trend is paralleled by a new openness in which devices that were once only available in hospitals become widely available to individuals while flexible mHealth applications tempt clinicians away from the hospital-based systems they used in the past. This popular market will disruptively challenge traditional approaches by being cheap and accessible.

Security and privacy of health data could be significantly affected by this trend. Freed from the bonds of HIPAA, mHealth apps are free to handle data using lower assurances than those typically applied to HIPAA entities. However, the data they handle is often as sensitive as the data handled by HIPAA entities. Typical Google Play apps such as Self-help Anxiety Management, iCardio, Epocrates CME, and Clinical Advisor provide assistance with mental health concerns, activity monitoring, and information services that reveal user interests in particular symptoms or diseases. It is important to develop guidelines for the security and privacy of mHealth apps that suit a dynamic market while assuring that the growth of mHealth does not lead to a cavalier vendor attitude toward personal data. New security and privacy risks particular to mobile computing and communications technology abound in mHealth apps[3, 4]. The aspects of mHealth make it different from other health information systems: First, mHealth apps allow a much larger amount of data being collected from the patient, as mobile devices can collect data over extended periods of time. Second, a much broader range of health-related data is being collected, as many mHealth apps collect patient activities and lifestyle, not only physiological data, but also include physical activity, location tracking, eating habits and diet details, social interactions and so on. Third, the nature of communications technology and mobile computing exposes many new attack surfaces to the outside world.

The goal of this paper is to carry out a three-stage study of the security and privacy status of free mHealth apps offered on Google Play. In the first study, the top 160 free mHealth apps in Google Play are classified and examined to formulate a list of attack surfaces that need attention in this area. These are shown in Table 1. Then a random sample of 27 apps is selected from the top 1080 apps and analyzed with respect to these seven attack surfaces. Significant issues are found in three attack surfaces: *Internet, Logging,* and *Third Party Services*. Since our concern about *Logging* will be addressed to a significant degree by deployment of a new version of Android, we focus our attention on the other two: *Internet* and *Third Party Services*. A random sample of additional 22 apps is taken involving Internet communications. Examination confirms that many of these 22 apps display significant risks to security and privacy on these two attack surfaces. Our primary conclusions are that the mHealth apps in Google Play commonly send sensitive data in clear text and store it on third party servers whose confidentiality rules may not be as strong as they need to be for the type of data being stored.

Table 1. Description of attack surfaces.

Attack Surface	*Description*
Internet	Sensitive information is sent over the Internet with insecure protocols, e.g. HTTP, misconfigured HTTPS, etc.
Third Party	Sensitive information is stored in third party servers
Bluetooth	Sensitive information collected by Bluetooth-enabled health devices can be sniffed or injected
Logging	Sensitive information is put into system logs where it is not secured
SD Card Storage	Sensitive information is stored as unencrypted files on SD card, publicly accessible by any other app
Exported Components	Android app components, intended to be private, are set as exported, making them accessible by other apps
Side Channel	Sensitive information can be inferred by a malicious app with side channels, e.g. network package size, sequence, timing, etc.

The remainder of this paper is organized as follows. We first discuss background and related work, then describe our methods for the three experiments. The next three sections describe the three studies respectively. We end with discussion.

2. Background and Related Work

In recent years, we have seen an increased adoption of mobile health applications by patients and physicians as well as the general public[1, 2]. Mobile computing and communications technology bring about new security and privacy concerns[3, 4]. The main objective of our study is to systematically investigate the security and privacy risks in mHealth apps on the Android platform. To the best of our knowledge, our study is the first study for classifying Android mHealth apps and summarizing their security and privacy risks.

2.1. Related Work

Recently, researchers have been actively involved in mHealth research. Mosa et al.[5] review articles discussing the design, development and evaluation of mHealth apps and discuss the differences between apps for healthcare professionals, for medical and nursing students, and for patients. Martínez-Pérez et al.[6] review on commercial mHealth apps for the most prevalent health conditions in the Global Burden of Disease list provided by the World Health Organization. Kotz[4] develops a taxonomy of the privacy-related threats to mHealth. Through an extensive survey of the literature, Avancha, Baxi, and Kotz[3] develop a conceptual mHealth privacy framework and discuss the technologies that could support privacy-sensitive mHealth systems. Aarathi et al.[7] investigate patients' privacy concerns about sharing their health information collected from mHealth devices with their family, friends, third parties and the public. Our goal is to review commercial mHealth apps from Google Play in order to classify, analyze and demonstrate their security and privacy risks.

2.2. Android Operating System

Our work focuses on researching the security and privacy risks on Android platform. Android is an open-source platform supported by Google that has become the most common OS for mobile devices. A report by F-Secure[8] shows that Android attracts much more malware attacks than iOS, which is another popular mobile platform. There are many mHealth apps and solutions have been built for the Android platform[9, 10, 11, 12]. Android is based on Linux for mobile devices. It provides a rich application framework to allow developers to build apps written in Java. App components are the essential building blocks of an Android app. There are four different types of components: Activity, Service, Content Provider and Broadcast Receiver. Android uses *Intents* for inter-component communication. Intents are used to start an Activity, to start a Service, or to deliver a Broadcast message. An *Intent Filter* is an expression composed from action strings that specifies the types of Intents a component would like to receive. Android provides a *permission* mechanism to enforce restrictions of inter-component communication and access to system resources.

3. Methods

This paper investigates the security and privacy risks in Android mHealth apps. More specifically, we will investigate such threats in three studies:

Study 1: What are the potential attack surfaces?

We review 160 apps identified by selecting the top 80 free apps in Health & Fitness category and another top 80 free apps in Medical category from Google Play. In order to get a sense of the context of Android mHealth apps, we first divide the 160 apps into two groups with regard to their target users and classify them into eight categories according to their functionalities. To develop a list of attack surfaces that are most representative, we review research papers[13, 14, 15, 16, 17] and documents[8, 18], and we analyze the 160 mHealth apps to find evidence of threats. Based on this review, the following seven attack surfaces represent areas that need protection: *Internet, Third Party Services, Bluetooth, Logging, SD Card Storage, Exported Components,* and *Side Channels*.

Study 2: How widespread is the threat?

After identifying the potential attack surfaces, Study 2 takes a further step to learn how widespread these attack surfaces are. The top 1080 free apps are identified from the Medical and Health & Fitness categories on Google Play, 540 from each. By using a random number generator without replacement through random.org, 27 apps are selected for the dataset for Study 2. Of these apps, we analyze them one by one in detail with respect to the seven attack surfaces identified in Study 1. Three attack surfaces are identified as important ones: *Internet, Third Party Services* and *Logging*, because the majority of the 27 apps evidence issues with these attack surfaces. Figure 1 shows how we include and exclude apps for Study 2.

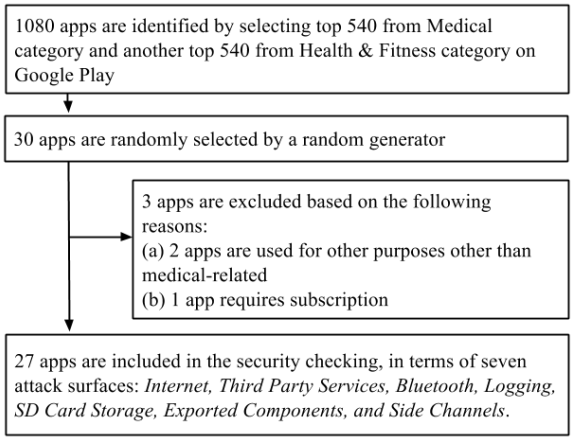


Figure 1. App selection flow graph for Study 2

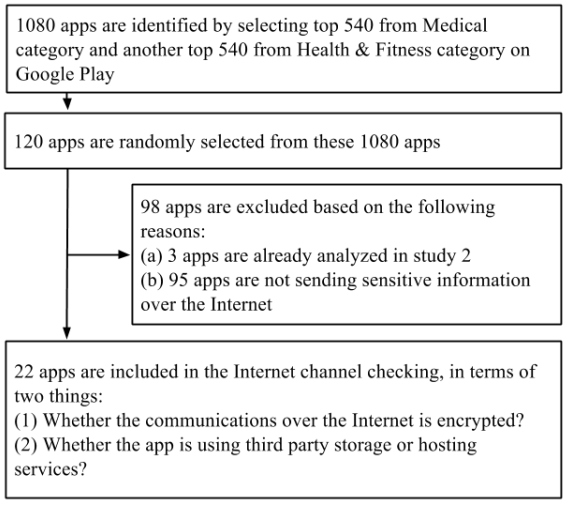


Figure 2. App selection flow graph for Study 3

Study 3: How serious is the threat?

Since security concerns in *Logging* will be addressed significantly by an Android version upgrade, we focus our attention on the other two attack surfaces, *Internet* and *Third Party Services*. The app selection process in Study 3 is similar to that of Study 2, but only apps that are sending sensitive information are selected. We randomly select 120 apps from the top 1080 free mHealth apps from Google Play. Then we use purposive sampling: the apps that have already been studied in Study 2 are excluded and the apps that are not sending sensitive information over the Internet are also excluded. In the end, 22 apps are included and analyzed in details to understand how serious the threat involving *Internet* communications is. Figure 2 shows how we include and exclude apps for Study 3.

4. Results

4.1. Study 1: What are the Potential Attack Surfaces?

To investigate the potential attack surfaces, we first want to understand the context of Android mHealth apps. By studying the 160 apps collected as described in the Section 3, we developed the classification system for Android mHealth apps shown in Table 2. We divide the top 160 free mHealth apps into two groups by their expected users. *Patient* apps are the ones mainly used by the individual whose health is being monitored. In most cases, the monitoring is done by the individual herself. *Healthcare Professional* apps are the ones mainly used by physicians, nurses, medical students, and other healthcare professionals to support their activities, which includes the monitoring

of patients. According to their functionalities, eight categories are used. Categories targeted at Patients include: *Lifestyle Management, Sensor-based Health Monitoring, Medical Contact, Medication and Disease Management,* and *Personal Health Record (PHR) Management*. Categories targeted at Healthcare Professionals include: *Medical References, Medical Training,* and *Clinical Communication*. A mHealth app may be useful for both Patients and Professionals (*e.g.* a pill identifier app can be used by patients to organize pills or by pharmacists to prevent errors in dispensing medications). Also, a mHealth app may belong to more than one category, since it may serve multiple functionalities (*e.g.* a fitness tracking app can monitor lifestyle data as well as manage PHR).

Table 2. Classification of popular free mHealth apps on Google Play.

Target users	*Category*	*Functionality Examples*	*Modules Used*	*Number of Apps (%)*
Patients	Lifestyle Management	Count calories; track eating habits, exercise, sleep, period, pregnancy, and etc.	Accelerometer, Gyroscope, GPS, Network	96 (60%)
	Medical Sensor-based Monitoring	Monitor health metrics such as: heart rate, blood pressure, blood glucose, insulin, cholesterol, and etc.	Externally connected health devices, Network	15 (9.38%)
	Medical Contact	Contact registered nurses, doctors or hospitals	Network, Phone call, Email	14 (8.75%)
	Medication and Disease Management	Manage prescription records; identify pills; shop medication online; look up symptoms; manage chronic diseases	Network	27 (16.88%)
	PHR Management*	Manage and/or synchronize PHR with health services	Network	75 (46.88%)
Healthcare Professionals	Medical References	Look up drug, disease and condition; anatomy tool; medical calculator; medical dictionary	Network	26 (16.25%)
	Medical Training	Aid medical students studying medical theories	Network	9 (5.63%)
	Clinical Communication	Emergency alert; photo sharing	GPS, Network	2 (1.25%)

* Here we define PHR management as patients syncing and managing user health information with an online health service provider.

Most of the applications in the categories are appropriate for our study but we exclude one app because it lacks a medical or healthcare purpose, and we exclude another app because its language is not English. Among the included 158 apps, we have 129 (81.65%) that are Patient-facing, 32 (20.25%) that are Professional-facing, and 3 (1.90%) drug identifier apps that are both. All the Patient-facing apps are from the Health & Fitness category and 41.03% of the apps from the Medical category are Professional-facing. In Table 2, the majority (60%) of the most popular Android mHealth apps are in the Life Management category. Nearly half (46.88%) of the apps manage and synchronize user health information to online service providers. The average rating score for the Patient-facing apps is 3.92, which is less than 4.18, the average score for the Professional-facing apps. However, the Patient-facing apps have almost 4 times more user installations, whose average is 502,263, than that of the Professional-facing apps, whose average is 139,125.

By studying previous literature[13, 14, 15, 16, 17] and online documents[8, 18], many different attack surfaces on Android apps have been identified. We study the 160 selected apps to have an understanding of what commercial mHealth apps are doing and whether risks exist in these attack surfaces. Real security issues are found within these Android mHealth apps, and the seven attack surfaces in Table 1 were identified as the most important ones. Here we use four specific examples in Android commercial mHealth apps to demonstrate the attack surfaces can lead to realistic and serious consequences.

Case 1 (Unencrypted Internet): Many mHealth apps send unencrypted information over the Internet. For example, both Doctor Online[19] (patients can talk to doctors online) and Recipes by Ingredients[20] (patients can search recipes according to their illness or ingredients suitable for their diseases), send unencrypted sensitive information,

including the user's email and password, in clear text over the Internet. Figure 3 shows the network traffic from Doctor Online captured by WireShark that contains user's name, email and password.

Figure 3. Network traffic from Doctor Online containing sensitive information.

Case 2 (Logging Sensitive Information): Many mHealth apps put sensitive user information into logs. For example, CVS/pharmacy[21], a popular app with millions of installations on Google Play, puts user login credentials and personal medical information in its log messages. Figure 4 shows the log messages with sensitive information from CVS/pharmacy. In the Example 1, CVS/pharmacy logs the prescription refill details from use inputs, including name, email address, store number, and Rx number. In the Example 2, CVS/pharmacy puts user login credentials in a debug log message. With this information a malicious party can view user profile and prescription history, which could support medical identity theft. A malicious party can even do pharmacy online shopping with users' stored credit card information.

Example 1: I/HttpDataClient(21039): https://native.usablenet.com/mt/ws/cvs.com/v2/refill?first=**tina**&last=**fey**&mail=**tina.health.droid%40gmail.com**&store=**24536**&orig=prod&rx1=**1524949**

Example 2: D/LOGIN (21039): https://native.usablenet.com/mt/ws/cvs.com/v2/login?username=**tina.health.droid%40gmail.com**&password=**password**&orig=prod

Figure 4. Log messages from CVS/pharmacy containing sensitive information.

Note that in both cases, the sensitive information is contained in the query strings of HTTPS URLs. Some developers may have a misconception that all HTTPS requests using GET or POST are sent over encrypted TCP connections so that sensitive information can be safely put into HTTPS URLs. However, even if sensitive information is not seen during transit, it remains visible in other places, such as mobile app logs, server logs, browser history and so on. Developers should avoid as far as possible including sensitive information in logs since it may be hard to know or control who is able to access the logs.

Case 3 (Exported components): Several apps in our study have component exposure threats. For instance, Noom Weight Loss Coach[22], an app with more than 10 million installations, exposes its Content Providers to external apps, which means any app can access the exposed Content Providers without declaring any permission. After searching for "content://" paths in the manifest and decompiled source code, we get a list of content URIs defined in the app. By using Drozer[23], an automatic security analysis tool, we attempted to access sensitive information with each content URI. Figure 5 shows the ability to read user workout history stored in the app's Content Provider with the content URI "content://com.wsl.noom.exerciseinfo".

dz> run app.provider.query content://com.wsl.noom.exerciseinfo

accessCode	key	timestamp	exerciseType	duration	caloriesBurnt	isManual	stepsCalories
WUX2G6EF	0	1393052360350	Walking	50892	0.484559	0	0
WUX2G6EF	1	1393009200000	Running	1800000	262	1	175

Figure 5. Access Noom Weight Loss Coach's user workout history by using Drozer.

Case 4 (Unencrypted SD card storage): Some sleep monitoring apps, such as SnoreClock[24] and Sleep Talk Recorder[25], record the sleep sounds of users and store them as unencrypted audio files on an external storage. For instance, the Sleep Talk Recorder explicitly stores sleep-recording unencrypted audio files on the SD card with name format YYYY-MM-DD-HH-MM-SS.wav. With read storage permission, a malicious app can read a user's sleep recordings; with internet permission, it can further send this information to remote servers. Another example is that Urgent Care[26] stores system logs in an unencrypted file on SD card, potentially leaking symptom lookup history.

4.2. Study 2: How Widespread is the Threat?

The complexity of the Android system has led to numerous potential attack surfaces that could be exploited by a malicious party to gain unauthorized access to sensitive data in mHealth apps and cause serious consequences. Analyzing these attack surfaces can help security specialists do security assessment and help mHealth users and developers understand and manage security risks. In Study 2, we analyze these attack surfaces with a new set of 27 random selected Android mHealth apps. The process of selecting these apps is described in Section 3.

Internet: Android mHealth apps access the Internet for various purposes, including to transfer information to a remote server and to retrieve ad to display to users. The information transferred over the Internet to a remote server

includes sensitive health information and ideally all such communication with the remote server should be encrypted. The 27 randomly sampled apps are analyzed to study why they require the Internet access (i.e. to transfer information or to display ads). Furthermore, we analyzed if the encrypted communication is used in network transmission.

Any app can get access to the Internet with Andoid's INTERNET permission. To study if the apps are using Internet for displaying ads or transferring information to the remote server, we study the description of the apps and check the functionality of the apps by installing and using them on a Samsung Galaxy SII phone. As a result, 85.2% (23/27) of the apps have the permission to access the Internet. 70.4% (19/27) use the INTERNET permission to display ads, while 29.6% (8/27) of them use it to communicate user information over the Internet.

To study whether the communication with remote servers is encrypted we installed and ran each of the apps while capturing network traffic using the "Shark for Root" Android app, and used WireShark to see if the traffic is encrypted. The result shows 7.4% (2/27) of the apps allow the users to use the blog or social network associated with the app via the Internet but only one of the apps used encrypted communication. We found that 25.9% (7/27) transmit medical information to the remote server, 57.1% (4/7) use encrypted communication, and 42.9% (3/7) use unencrypted communication to transfer the sensitive health related information.

We analyzed if the three apps sending unencrypted data over the Internet are actually sending sensitive information. The first app searches for nearby pharmacies, doctors, etc. The second app tracks exercise workouts, and the third (Doctor Online from Spain) facilitates finding and talking to doctors online. Doctor Online sends email, username and even password unencrypted over the Internet.

Third Party Services: Android apps use storage and hosting services such as Amazon instead of maintaining their own infrastructure. This is an economical as well as scalable solution for mobile apps. But storing sensitive health information on these third party services can have serious implications even for large and widely-trusted services like Amazon. We study if these seven apps communicating with remote servers are hosted on the cloud or on-premises servers owned by the app vendors. To this end, we analyze the IP addresses of these apps in the communications with their respective servers. IP addresses have a publicly available record of whom it belongs to and we use this to find out where the traffic is going. We found that 85.7% (6/7) apps are hosted on third party servers. Three of them are hosted on Amazon and rest on other hosting services. We were not able to tell if data on the remote third party servers is stored in encrypted fashion such that the hosting companies do not have access to this data. However, the four apps mentioned in the previous Internet section are using encryption for the communication only.

Bluetooth: Many mHealth sensing apps primarily use Bluetooth to collect data from health sensors to mobile devices. One app (3.7%) of the 27 apps in our dataset connects to a Bluetooth health device to collect personal health information. Supporting Bluetooth devices is more common among the 160 most popular Android mHealth apps, where 15/160 (9.5%) provide Bluetooth connectivity to collect health data. 12 of the 15 apps declare and use both BLUETOOTH and BLUETOOTH_ADMIN permissions, so that they can use Bluetooth to connect and collect data from external health sensors, while the remaining three of them collect health data via the Internet or by connecting with other apps. The apps collect various types of health information, including heart rate, respiration, pulse oximetry, electrocardiogram (ECG), blood pressure, body weight, body temperature, quality of sleep, exercise activities. Apparently, Bluetooth is a major communication technology for sensor-based health monitoring in Android mHealth apps. Naveed et al.[14] present a problem of external-device misbonding (DMB) for Bluetooth-enabled Android devices and health sensors. They show how a malicious app can stealthily collect user data from an Android device or spoof a device and inject fake data into the original device's app. One app of the 27 apps connects to external health sensors and uses default PIN code 0000, which makes it vulnerable to the DMB attack. To defend against the Bluetooth-based threats on mHealth apps, Naveed et al.[14] propose an OS-level protection, which generates secure bonding policies between a device and its official app and enforces these rules when establishing and terminating Bluetooth connections.

Logging: the Android logging system enables developers to collect and view debugging output for apps. The logging facility allows a system-wide logging, including both application information and system events. If an app is granted READ_LOGS permission, the app is allowed to read the low-level system log messages. With the READ_LOGS permission, a malicious app may be able to extract sensitive information from log messages. To find such logging vulnerabilities we used a tool called logcat from the Android Debug Bridge (ADB) shell to view system log messages.

In our dataset, 9 out of the 27 apps (33.3%) put sensitive information in log messages. Among the 9 aforementioned apps, two (22.2%) disclose GPS coordinates, three (33.3%) disclose Facebook friend information, and one (11.1%) divulges more sensitive data such as user sign up data, which includes name, location and profession of the user. Three (33.3%) apps leak disease and drug browsing history in the app logs. From the study on our dataset, it indicates a large number (33.3%) of the mHealth apps leak sensitive information in system logs that could support cause serious attacks such as medical identity theft.

SD Card Storage: Each Android app gets a dedicated part of file system where it can write its private data. However, if an app writes files to an external storage, such as an SD card, the files are not guaranteed to be protected. With READ_EXTERNAL_STORAGE or WRITE_EXTERNAL_STORAGE permissions, any app can read or write files from an external storage. Before API level 19, the READ_EXTERNAL_STORAGE permission is not enforced and all apps still have access to read from an external storage.

In our dataset, 66.7% (18/27) of the apps declare the WRITE_EXTERNAL_STORAGE permission, which means they write data to external storage that can be read by any app with the READ_EXTERNAL_STORAGE permission. We used Dex2jar[27] to decompile the application package (apk) files for these 27 apps to get their Java source code. We searched this code for the "ExternalStorage" and "ExternalFiles" keywords in the source code to construct all possible paths for files stored on SD card. Then, we executed all possible operations with the studied apps and exhaustively went through the resulting directories to check their file contents. This search did not reveal evidence that any of the apps store sensitive information in external storage files.

Exported Components: Android app developers can specify if a component (Activity, Service, Broadcast Receiver, or Content Provider) is public to external apps. A component can be declared as *exported*, or public, if its declaration sets the EXPORTED flag or includes at least one Intent Filter without permission protection. However, setting a private component improperly as *exported* enables a malicious app to send unwanted Intents to the component, which can cause security problems with broadcast injection, activity launch or service launch[17]. In addition, if the Content Provider is exported, a malicious app can read or write the exported Content Provider without declaring any particular permission. The Content Provider supports the basic "CRUD" (create, retrieve, update, and delete) functions and the data in a Content Provider is addressed via a "content URI". Knowing the "content URI" from an exported Content Provider, a malicious app can retrieve or modify the data according to the Content Provider's schema. An example we gave earlier in our discussion of Study 1 illustrates an unauthorized access to an exported Content Provider to read the app's sensitive information.

Side Channels: All the attack surfaces discussed above are using explicit channels in Android system, where a malicious party has chance to directly read sensitive data from the attack surfaces. Besides the explicit channels, side channels can be exploited by a malicious party to infer sensitive information from apps, even if they are well-designed and implemented by their developers. Zhou et al.[13] find a correlation between network payload size, which is publicly accessible in Android system, and the disease condition a user selects on WebMD mobile[28]. With this correlation even an app with no permissions, a "zero-permission" malicious app, can monitor WebMD's network payload in the background, and map their monitoring results to the disease condition that a user searches on WebMD. Side channel information leakage has also been discovered from motion sensors, such as accelerometer and gyroscope[15, 16]. Fine-grained motion sensor monitoring can be used to infer keystrokes, such as 4-digit PIN codes, on touch screen smartphones with only soft keyboards.

To circumvent the network-payload-based side channel attack, Zhou et al.[13] present a mitigation mechanism which enforces limitations to accessing Android public resources (*e.g.* network payload size) by modifying the Android kernel. MHealth developers can pad blank information to network packets to ensure they are fixed-length, or develop offline strategies for downloading sensitive data. For the motion-sensor-based side channel attack, Adam *et al*.[15] propose disabling untrusted access to motion sensors whenever a trusted input function (*e.g.* password entry) is being performed.

4.3. Study 3: How Serious is the Threat?

Three vulnerabilities in Study 2 are revealed to be common and serious: sending sensitive information unencrypted over the Internet, storing it on third party services, and including it in logs. Since logging can be addressed by an Android version upgrade, we focus our investigation on the other two threats, Internet traffic and third party services.

As only seven apps in Study 2 are actually sending sensitive information over the Internet, we carried out Study 3 to understand the prevalence of these threats with a larger number of apps using the Internet. Another 120 apps are

randomly sampled from the 540 top Health and Fitness apps and another 540 top medical apps (1080 in total) from Google Play. These apps are then manually analyzed to rule out those not sending any sensitive information over the Internet. After the filtering, 22 apps are found to be sending sensitive data over the Internet (some apps requiring subscription are filtered out as well). To analyze their Internet traffic, we installed these 22 apps and captured their traffic using the same methods described in Study 2. The results reveal that 63.6% (14/22) of these apps are sending unencrypted data over the Internet and 81.8% (18/22) are using third party storage and hosting services such as Amazon's cloud services. One of our randomly selected apps (Fitbit) uses encryption over the Internet, but is also using third party storage and hosting services. The four apps that are using their own servers to store and host their apps are big companies such as Aetna, United Healthcare, Caring Bridge and US Dept. of Health and Human Services. We were not able to obtain ground truth about whether apps encrypt data when they store it with third parties, but one may conjecture that apps that do not encrypt data over the Internet probably also do not encrypt it on third party storage. Even though the data might be hosted in an isolated environment (e.g. on an isolated VM in the cloud), storing unencrypted data on third party storage makes the data vulnerable to insider attack, where the service provider is malicious.

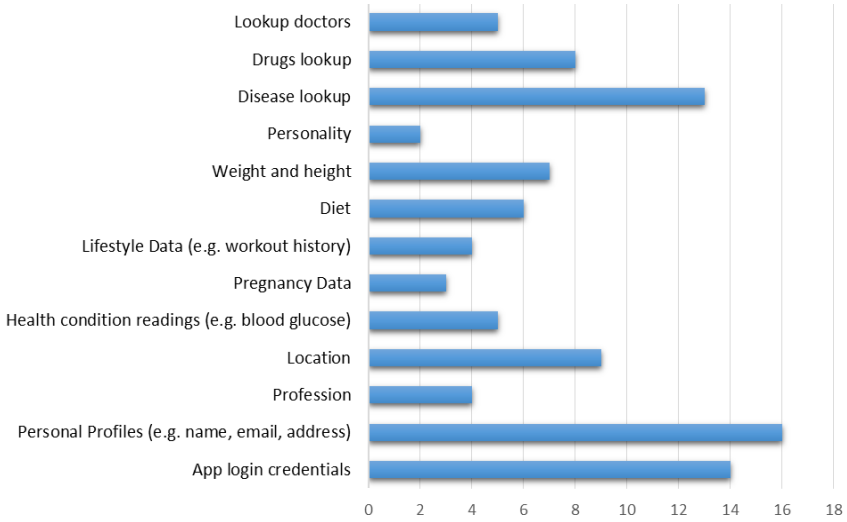


Figure 6. Sensitive information distribution in the 22 apps dataset for Study 3.

When used as intended, a variety of sensitive user data are collected, stored, and transmitted through these Android mHealth apps. Figure 6 shows the distribution of sensitive information in the 22 apps (x axis means the number of appearances of sensitive information in the 22 apps). Based on our study, the information includes at least personal profiles, health sensor data, lifestyle data, medical information browsing history, and third-party app data (*e.g.* Facebook account information). Depending on the type, sensitivity, and volume of mHealth data breaches, disclosure or tampering with these sensitive data may lead to serious consequences, such as profiling, medical identity theft, and healthcare decision-making errors. According to the information collected from World Privacy Forum[29], thefts have used stolen medical information for a resourceful collection of nefarious purposes. For example, a Colorado man whose Social Security number, name and address had been stolen received a bill for $44,000 he presumably owed to a hospital because his identity had been used by a thief to get medical services in his name. In another case, another identity thief in Missouri used the personal data of multiple victims to establish false driving licenses and was able to use them to obtain prescriptions in the victims' names at a regional health center.

5. Discussion

5.1. Summary of Findings

Our three-stage study raises many concerns and shows some serious problems with Android mHealth apps. The major issue is unencrypted communication over the Internet and use of thirty party hosting and storage services. Our study shows that a significant number of the apps in the top health related apps from the Google Play market have these issues. These issues need attention and are not easily fixable because they require extra effort and security expertise from developers and computational capabilities from platforms. Third party cloud and hosting services

provide a very economical solution for hosting app services and storing data and many app vendors may feel it is not economical to maintain their own servers or even encrypt stored data on the third-party services.

5.2. Compliance Recommendations

The increased use of mHealth results in greater risks to health-related information on mobile devices. Developers and healthcare service providers would be wise to make efforts to ensure that mHealth apps facilitate security compliance even if they are not legally required to do so (at the current time). Based on our study on the risks from mHealth apps, here are some important compliance recommendations: encryption is essential to secure personal data stored on mobile devices; when accessing web-based services, TLS/SSL should be deployed throughout the Internet transmission session; even though the network transmission session is protected and encrypted, using third party services to store users' sensitive data must be closely reviewed and users should be informed when it is happening; developer guidelines or training can be helpful in avoiding many of the common mistakes that are rooted from development with poor security practices; risk assessment provided by authorities can further minimize the security risks that may harm users. Experience with Happtique, which was forced to suspend app certifications after some of the apps it certified were found to have some of the problems above, provides evidence of strong incentives for better security and privacy practices[30]. A report from Symantec also raises questions about security risks in self-tracking devices and apps[31]. A good possible direction is for mHealth app developers to create a set of security and privacy guidelines that offer a baseline for protections.

5.3. Limitations of the Study

Android version upgrade. Android is constantly making behavior changes in order to circumvent newly found threats. For example, to mitigate the logging information leakage problem, since Jelly Bean (Android 4.1), an app can only collect and view log messages originating from itself. However, on a rooted device (i.e., a device allows any app to run with administration permissions on Android)[32], a malicious app can, by executing a '*pm grant'* command, grant itself a READ_LOGS permission. This means it is still dangerous for an app to keep sensitive information in system logs. According to the Android platform distribution[33] collected in March, 2014, almost 40% of the overall Android devices are under the version of Jelly Bean. Due to a large number of Android device users and mHealth apps, it is lucrative for malicious parties to investigate ways to harvest sensitive personal healthcare information from mHealth apps.

User agreements. We observed that many apps may ask users to share their private health information by providing privacy policy agreed by users themselves. In our study, most of the apps do make privacy policies available to users either via an URL link in the app or shown when the app is launched for the first time. How health data is managed and transmitted is generally out of control or visibility of the users, but the apps should at least encrypt all data in transit and at rest. We believe that understanding the privacy policies of mHealth apps is an interesting future research topic. Users should know what they are agreeing to in order to use the app and how their data can be used.

6. Conclusion

A study of Android mHealth apps reveals common shortcomings in security and privacy when using communications and storage. Steps should be made to encourage mHealth app vendors to assure encrypted network links for communications and the use of third party storage only when adequate security and privacy guarantees are obtained.

Acknowledgements

We acknowledge the grant HHS-90TR0003/01 (SHARPS, sharps.org) from the Office of the National Coordinator for Health Information Technology at the Department of Health and Human Services (HHS) and NSF 13-30491 (THaW, thaw.org). The views in this paper represent opinions of the authors only.

References

1. Aitken M, Gauntlett C. Patient apps for improved healthcare from novelty to mainstream. Parsippany (NJ): IMS Institute for Healthcare Informatics; 2013 Oct.
2. MarketsandMarkets. Mobile health apps & solutions market by connected devices (cardiac monitoring, diabetes management devices), health apps (exercise, weight loss, women's health, sleep and meditation), medical apps (medical reference) – global trends & forecast to 2018. 2013 Sep. Report No.: HIT 2104
3. Avancha S, Baxi A, Kotz D. Privacy in mobile technology for personal healthcare. ACM Computing Surveys (CSUR). 2012; 45 (1): 3.

4. Kotz, D. A threat taxonomy for mHealth privacy. 2011 Third International Conference on Communication Systems and Networks (COMSNETS); 2011 Jan 4-8; Bangalore.
5. Mosa A, Yoo I, Sheets L. A systematic review of healthcare applications for smartphones. BMC Medical Informatics and Decision Making; 2012 Jul 10; 12(7). BioMed Central.
6. Martínez-Pérez B, de la Torre-Díez I, López-Coronado M. Mobile health applications for the most prevalent conditions by the world health organization: review and analysis. J Med Internet Res 2013 Jun 14; 15(6):e120.
7. Prasad A, Sorber J, Stablein T, Anthony D, Kotz D. Understanding Sharing Preferences and Behavior for mHealth Devices. Proceedings of the 2012 ACM Workshop on Privacy in the Electronic Society (WPES). October 15; Raleigh, NC. New York, NY: ACM; 2012. p. 117-128.
8. F-Secure Labs. Mobile threat report. Helsinki, Finland; 2013 Jul-Sep. Report No. 2013 Q3. http://www.f-secure.com/static/doc/labs_global/Research/Mobile_Threat_Report_Q3_2013.pdf
9. Rajput Z, Mbugua S, Amadi D, Chepng'eno V, Saleem J, Anokwa et al. Evaluation of an Android-based mHealth system for population surveillance in developing countries. J Am Med Inform Assoc; Feb 24; 2012.
10. Altini M, Penders J, Roebbers H. An Android-based body area network gateway for mobile health applications. Proceedings in Wireless Health 2010; 2010 Oct -710-13; San Diego, CA. New York, NY: ACM; 2010. p. 188-189.
11. Wei R, Yang Z. Design and implementation of doctor-patient interaction system based on android. 2012 International Symposium on Information Technology in Medicine and Education (ITME); 2012 Aug 3-5; Hokodate, Hokkaido; 2: 580-583.
12. Gregoski M, Vertegel A, Treiber F. Photoplethysmograph (PPG) derived heart rate (HR) acquisition using an Android smart phone. Proceedings of the 2nd Conference on Wireless Health; 2011 Oct 10-13; San Diego, CA. New York, NY: ACM; 2011: 23.
13. Zhou X, Demetriou S, He D et al. Identity, location, disease and more: inferring your secrets from Android public resources. 20th ACM Conference on Computer and Communications Security (CCS); 2013 Nov 4-8; Berlin, Germany. New York, NY: ACM, 2013.
14. Naveed M, Zhou X, Demetriou S, Wang XF, Gunter CA. Inside job: understanding and mitigating the threats of external device mis-bonding on Android. Proceedings of the 21st Annual Network and Distributed System Security Symposium (NDSS); 2014 Feb 23-26; San Diego, CA. Reston, VA: The Internet Society, 2014.
15. Aviv A, Sapp B, Blaze M, Smith J. Practicality of accelerometer side channels on smartphones. Proceedings of the 28th Annual Computer Security Applications Conference (ACSAC). December 9-13; Orlando, FL. New York, NY: ACM; 2012. p. 41-50.
16. Cai L, Chen H. On the practicality of motion based keystroke inference attack. Proceedings of the 5th International Conference on Trust and Trustworthy Computing (TRUST). June 13-15; Vienna, Austria. Berlin, Heidelberg: Springer-Verlag; 2012. p. 273-290.
17. Chin E, Felt AF, Greenwood K, Wagner D. Analyzing inter-application communication in Android. Proceedings of the 9th international conference on Mobile systems, applications. June 28-July 1. New York, NY: ACM; 2011. p. 239-252.
18. Seven ways to hang yourself with Google Android. http://www.cs.berkeley.edu/~emc/slides/SevenWaysToHangYourselfWithGoogleAndroid.pdf
19. Doctor Online. https://play.google.com/store/apps/details?id=com.airpersons.airpersonsmobilehealth
20. Recipes by Ingredient. https://play.google.com/store/apps/details?id=com.abMobile.recipebyingredient
21. CVS/pharmacy. https://play.google.com/store/apps/details?id=com.cvs.launchers.cvs
22. Noom Weight Loss Coach. https://play.google.com/store/apps/details?id=com.wsl.noom
23. Drozer. https://www.mwrinfosecurity.com/products/drozer/
24. SnoreClock. https://play.google.com/store/apps/details?id=de.ralphsapps.snorecontrol
25. Sleep Talk Recorder. https://play.google.com/store/apps/details?id=com.madinsweden.sleeptalk
26. Urgent Care. https://play.google.com/store/apps/details?id=com.greatcall.urgentcare
27. Dex2jar. https://code.google.com/p/dex2jar/
28. WebMD mobile. http://www.webmd.com/mobile
29. Dixon P. Medical Identity Theft: The Information Crime that Can Kill You. World Privacy Forum; 2006 May 3.
30. Happtique suspends mobile health app certification program. http://mobihealthnews.com/28165/happtique-suspends-mobile-health-app-certification-program/
31. Symantec Corporation. How safe is your quantified-self? 2014 July.
32. Rooting – is it for me? Some Q&A. http://www.androidcentral.com/rooting-it-me-some-qa
33. Android historical version distribution. https://developer.android.com/about/dashboards/index.html

Fostering Multilinguality in the UMLS: A Computational Approach to Terminology Expansion for Multiple Languages

Johannes Hellrich and Udo Hahn

Jena University Language & Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena, Jena, Germany

Abstract

We here report on efforts to computationally support the maintenance and extension of multilingual biomedical terminology ressources. Our main idea is to treat term acquisition as a classification problem guided by term alignment in parallel multilingual corpora, using termhood information coming from of a named entity recognition system as a novel feature. We report on experiments for Spanish, French, German and Dutch parts of a multilingual UMLS-derived biomedical terminology. These efforts yielded 19k, 18k, 23k and 12k new terms and synonyms, respectively, from which about half relate to concepts without a previously available term label for these non-English languages. Based on expert assessment of a novel German terminology sample, 80% of the newly acquired terms were judged as reasonable additions to the terminology.

Introduction

The life sciences are one of the most active areas for the development of terminological resources covering a large variety of special thematic fields such as human anatomy, cell structure, diseases, chemical substances, gene products, etc. These individual activities have already been bundled in several portals which combine up to some hundreds of these specialized terminologies and ontologies, such as the Unified Medical Language System (UMLS),[1] the Open Biological and Biomedical Ontologies (OBO),[2] and the NCBO BIOPORTAL.[3]

This stunning variety of terminological resources has grown over time primarily for the English language. Despite ambitious efforts by non-English language communities to provide translated "mirror" terminologies in their native languages the term translation gaps encountered here remain huge—even for prominent and usually well-resourced European languages, such as French, Spanish or German. The non-English counterparts of the UMLS for these languages lack between 65 to 94% of the coverage of the English UMLS (see Table 1), and there is an even much lower long tail for under-resourced languages, e.g. Greek or Hungarian. Missing terminological coverage not only decouples the English-speaking medical language community from the non-English ones but also hampers all sorts of meaningful international data exchange or data compilation (e.g. gathering reasonably sized cohorts for rare diseases).[4] Further multilingually rooted healthcare problems arise in the age of ever-increasing cross-border mobility of people due to world-wide tourism and for countries with a substantial proportion of immigrants lacking sufficient proficiency of the country's language which is not their primary language.[5]

Table 1. Number of UMLS terms and synonyms per selected language in a slightly pruned UMLS version (taken from the CLEF-ER challenge data),[6] term coverage relative to the English UMLS version, number of additional entries acquired through our terminology extraction system and lexical extension rate (increase relative to the number of previously available entries for the selected language). Data are based on classifier 7 (see Table 5).

Language	UMLS terms & synonyms	Coverage rate (w.r.t. English)	Additional entries	Lexical extension rate (w.r.t. selected language)
English	1,822k	100.0%	–	–
Spanish	644k	35.3%	18k	2.8%
French	137k	7.5%	19k	13.9%
German	120k	6.6%	23k	19.2%
Dutch	116k	6.4%	12k	10.3%

We here report on automatic efforts to partially close this gap by treating term acquisition as a classification problem, capitalizing on lexical and phrasal alignment techniques from the field of statistical machine translation (SMT) and named entity recognition (NER) techniques applied to selected parallel corpora. SMT uses statistical models to align and thus map expressions from a source language into expressions of a target language. Nowadays most SMT translation models are phrase-based, i.e. they use the probability of two text segments being translational equivalents of each other and the likelihood of a text segment existing in the target language to model translations.[7] SMT model building requires an ample amount of training material, usually *parallel* corpora. Such corpora are manually constructed, direct translations from a source into a target language text. In the meantime, open source SMT systems have been demonstrated to automatically generate translations of considerable quality for languages such as German, Spanish or French, even for specialized domains such as biomedicine.[8]

Since the availability of parallel corpora is so crucial for the operation and performance of SMT alignment systems, we assembled several corpora that are thematically rooted in the biomedical domain and took care as much as possible of such issues as linguistic variability and variance in text genres. For the purposes of this paper, we focus on two types of parallel corpora: MEDLINE, on the one hand, supplies English translations of titles from a wide range of non-English scientific journals (especially numerous for Spanish, French and German). On the other hand, the EMEA corpus provides patient information leaflets in 22 languages, among them about one million sentences each for Spanish, French, German and Dutch.[9] All these documents have carefully been crafted by the authors of journal articles or educated support staff such as editors or translators so that translation quality is generally high. Using parallel corpora for biomedical terminology extraction has already been proposed by Déjean et al.[10] and Deléger et al.,[11,12] with German and French as the respective target languages from English sources. Compared with today's standards for biomedical NLP system engineering this early work, however, suffers from limited scalability and generalizability and lacking transparency of the evaluation frameworks (for a more detailed discussion, see below).

Since parallel corpora are mostly rare (for some domains they are even unavailable) and limited in scope, researchers have been looking for reasonable alternatives such as *comparable* corpora. Just as parallel corpora they consist of texts from the same domain, but unlike parallel corpora no direct mapping between content-related sentences can be assumed for the different languages. Comparable corpora are easier to get and much larger in size and thematic diversity,[13] yet provide less accurate results due to the increased level of inherent noise. Therefore, they are sometimes used in combination with parallel corpora which leads to better results than employing each on its own.[10] Comparable corpora are typically accompanied by context-based projection methods, i.e. collecting context vectors for words in both languages and partially translating these with a seed terminology—words with similar vectors are thus likely to be translations of each other. A brief overview of the use of comparable corpora in SMT and the influence of different parameters is provided by Laroche & Langlais.[14] Skadia et al.[15] show how a subset can be extracted from a comparable corpus which can be treated as a parallel corpus.

Aker et al.[16] treat term extraction as a classification task as well. Yet they use single-word translation probabilities generated from parallel corpora to extract terms from comparable corpora rather than phrase probabilities generated and terms directly extracted from parallel corpora as in our approach.

Non-English terminology can also be harvested from documents in a mono-lingual fashion, without an *a priori* reference anchor in English; hence, no parallel corpora are needed. These approaches are primarily useful, if no initial multilingual terminology exists. After extraction, these terminologies need to be aligned, which can be done using purely statistical or linguistically motivated similarity criteria.[17,18]

Terminology extraction in the biomedical field is a tricky task because terms not only appear as single words (e.g. *'appendicitis'*) but much more as multi-word expressions (e.g. *'Alzheimer's disease'* or *'acquired immunodeficiency syndrome'*). Approaches towards finding these multi-word expressions can be classified as either pattern-based, using e.g. manually created part-of-speech (POS) patterns that signal termhood, or statistically motivated, e.g. utilizing phrase alignment techniques. Pattern-based approaches such as the ones described by Déjean et al.[10] or Bouamor et al.[19] are fairly common. The downside of these approaches is their need for term-indicative POS patterns that are often hand-crafted and may become increasingly cumbersome to read, write and maintain. Alternative approaches rely on statistical data, like the translation probabilities of the single words of a term (treated as a bag of words)[20] or phrases. Those phrases can be linguistically motivated, i.e. use POS information,[21] or be purely statistically informed, e.g. derived from the model produced by a phrase-based SMT system.

Also in terms of evaluation, a large diversity of approaches can be observed. Some groups report only precision data based on the number of correct translations produced by their system;[11] others issue F-scores based on the system's ability to reproduce a (sample) terminology.[10] In these studies, numbers for new and correct translations range

between 62% and 81%.[11,12] Unfortunately, these results are not only grounded in the systems' capabilities, but also strongly influenced by both the particular terminology and parallel corpora being used—thus, comparisons between systems are generally hampered by the lack of a common methodology and standards for evaluation.

An earlier state of our own work was decribed by Hellrich & Hahn.[22] This contribution supersedes that former version in numerous technical and methodological respects. In particular, it contains the following pieces of new information: We here evaluate the performance of the JCoRe named entity tagger (Table 3), provide an in-depth empirical analysis of the term re-invention performance of our approach under different experimental conditions (e.g. various choices of cut-offs and feature combinations, including phrasal translation probabilities, named entity termhood, as well as string length and string similarity criteria) by considering the UMLS as a gold standard (Table 5), and evaluate the term mining performance under two different conditions in a human assessment task. We also include a quantitative assessment of the relative frequency of UMLS terms of different lengths which is crucial for the length limitation of string similarity computations (Figure 1).

Methods

Methodologically speaking, our approach primarily combines reusable off-the-shelf components, namely the LingPipe[a] gazetteer for lexical processing, Giza++[23] and Moses[24] for generating phrasal alignments, the JCoRe[25] biomedical named entity recognizer (to indicate the degree of termhood of phrases) and Weka[26] as the machine learning repository from which we took a Maximum-Entropy classifier to combine NER and SMT information. We distinguish three steps in our approach: corpus set-up, term candidate generation and term candidate classification. Both the corpora and the UMLS version we used were taken from the CLEF-ER 2013 challenge material.[b,6]

1. Corpus Set-Up

We started by merging the MEDLINE and EMEA parallel corpora, resulting in one single corpus per language pair. These corpora contained 860k, 713k, 388k, and 195k parallel title phrases (from MEDLINE) and sentences (from EMEA) for the German, French, Spanish and Dutch language, respectively (see Table 2 for details), with English always acting as the reference language. Subsequently, a flat annotation for the common "Biomedical Entity" type was performed. For this task, we used a LingPipe-based gazetteer informed by the UMLS-derived terminology for all languages involved; this was used as gold data for training and evaluation. The gazetteer approach relies on only exact matching and cannot disambiguate terms (about 2.5% of them correspond to more than one concept), yet was chosen for its capability to work for any language.

Table 2. Overall corpus size, number of MEDLINE titles and EMEA sentences for each language pair. These numbers reflect the final version of the CLEF-ER data and are more recent than the original ones from the challenge website.

Language pair	Corpus size	MEDLINE titles	EMEA sentences
German-English	860k	719k	141k
French-English	713k	572k	141k
Spanish-English	388k	248k	141k
Dutch-English	195k	54k	141k

10% of each of the automatically annotated corpora were set aside and used to train language-specific JCoRe NER systems in order to find biomedical entities, in general, without any deeper distinction of more fine-grained semantic groups, such as Disease or Anatomy (common in the UMLS framework). Thus, the probabilities provided by the systems merely characterize domain-specific, in our case biomedical, termhood of a word sequence under scrutiny. Performance data for these systems can be found in Table 3. The remaining 90% of the corpora were used to generate phrasal alignments with Giza++ and Moses, the basis for term extraction. The gazetteer-based annotations were exploited for the training and evaluation of the term extraction system. The corpora were split to prevent NER re-classification effects from inflating term extraction performance.

[a] http://alias-i.com/lingpipe/

[b] https://sites.google.com/site/mantraeu/clef-er-challenge

Table 3. Performance of the language-specific JCoRe NER systems (10-fold cross-validation for each language).

Language	F_1 Score	Precision	Recall
German	0.78	0.88	0.70
French	0.77	0.83	0.72
Spanish	0.82	0.87	0.79
Dutch	0.58	0.92	0.43

2. Term Candidate Generation

The so-called phrase table generated by Moses contains alignment information in terms of phrase pairs, e.g. *'an indication of tubal cancer' :: 'als Hinweis auf ein Tubenkarzinom'* and translation probabilities for each such pair—conditional probabilities for a phrase in the target language *t* (either German, French, Spanish or Dutch) and a phrase in the source language *s* (English) being translations of each other. More specifically, we take into account the direct phrase translation probability $\phi(t|s)$, the inverse phrase translation probability $\phi(s|t)$, the direct lexical weighting lex(t|s), and the inverse lexical weighting lex(s|t). Term candidates for enriching the UMLS were produced by selecting those phrase pairs that translate a known English biomedical term into one of the target languages, as long as the term candidates were different from all synonyms in the target language with respect to all predecessors and sucessors of the concept in question.

3. Term Candidate Classification

A naïve baseline system enriches the terminology with all term candidates forming these pairs, yet most of them will presumably be wrong. Since the phrases in this system are statistically rather than linguistically motivated, a simple mapping from text segments onto other text segments is given. This may also lead to scruffy translations like *'cancer' :: 'Krebs zu'* ['cancer to'], which not only contain the relevant biomedical terms (*'cancer'* and *'Krebs'*) but also (sometimes multiple) text tokens merely adding noise (here, *'zu'* ['to']).

A first refinement step keeps only those term candidates that are the most probable ones according to the phrase table, as those will be both relatively frequent and specific. This can be achieved by introducing a cut-off which selects the *n* most likely target term candidates based on the direct phrase translation probability, $\phi(t|s)$, for each English biomedical source term. In our experiments, a cut-off of '5' led to recall statistics identical with the baseline while massively reducing the number of odd translations. Hence, this cut-off was taken for all subsequent filtering steps.

Additional filtering conditions became instrumental in a binary classifier for term candidates. We trained a Maximum-Entropy classifier from Weka, using the gazetteer-based annotations generated for the corpus set-up as positive training examples (all other potential term candidates were taken as negative examples, so that the system is biased towards proposing new terms) and tested combinations of the following features:

- *Translation probabilities* from the phrase table;
- Judgments on each term candidate's termhood based on its classification as a *biomedical named entity*. We generated this piece of evidence by running the NER system on all sentences containing the candidate. We then summed the probabilities for each candidate being a named entity and divided the sum by the total number of the sentences containing that candidate (using '0', if no match was found);
- A *scaled ratio belancing the length* of the term candidates against their putative translation equivalents (the rationale was to exclude overly length-biased and thus unlikely translation pairs);
- Five *string similarity criteria* based on cognate features (following Aker et al.[16]). Since our work is based on word sequences (see Figure 1 for an overview of the number of words per term in the UMLS) rather than single words, we tested several reordering variants that are likely to appear in the UMLS as well. Besides comparing the unmodified term candidate and its translational equivalent we also tested all permutations of the term candidate's constituent words (for terms up to length '5' only, due to computational constraints). This was done both for independent reordering, selecting the permutation yielding the highest value for each feature and with a single uniform reordering for all features, selecting the permutation leading to the maximal harmonic mean for all five cognate features. Collapsing these five features into their harmonic

mean as a single feature was pre-tested, yet led to slightly degraded performance. Overall, this feature turned out to be Janus-sided during the evaluation in the following sense: it inceased the performance for reconstructing UMLS fragments (i.e. known terms), whereas it decreased the recognition of acceptable new terms.

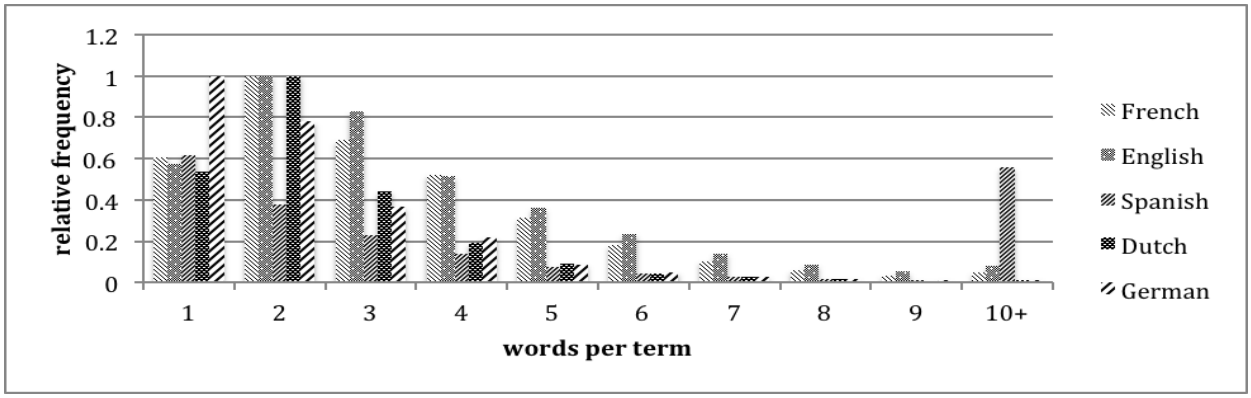


Figure 1. Relative frequency of UMLS terms with *n* words per language; normalized by dividing through the absolute frequency of the most frequent number of words per term for each language.

Results

In general, three types of translation equivalents can be found for the non-English terminologies using our approach: already *known terms*, *additional synonyms for concepts* already labeled in a terminology, as well as *entirely new terms for concepts* with no prior label in a terminology. In order to evaluate the validity of the extracted terms we conducted both an automatic evaluation procedure (for already known terms and additional synonyms) and elicited human expert assessments (for entirely new terms).

1. Automatic Evaluation: Matching Terminologies for Term Re-Invention

In the first setting, we actually measured the system's ability to *re-invent UMLS fragments*, using 10-fold cross-validation for the classifier-based systems. This evaluation was carried out concept-wise, i.e. a term being a synonym for multiple concepts, or a translation thereof, was examined multiple times depending on its ambiguity rate. A term was counted as being correct, if it was already contained in the set of synonyms in the considered language for the UMLS concept. False negatives (for recall calculation) were counted by using those concepts as an upper bound for which the gazetteer system had annotated terms in two aligned sentences (see Table 4 for an overview); we call such terms *traceable*. An obvious drawback of this approach is that additional synonyms and entirely new terms are erroneously counted as being incorrect, since they are not yet contained in the terminology. This may also hamper training, since valid additions are treated as negative examples during training. Any alternative avoiding these shortcomings would have required manual annotation.

Table 4. Number of traceable terms and synonyms, i.e. known terms and synonyms in parallel sentences, per language, used as an upper bound estimate to calculate recall.

Language	Traceable
Spanish	19,797
French	17,492
German	15,428
Dutch	2,017

The automatic analysis allows us to assess the influence of different approaches and feature combinations on term extraction for all four languages (results are listed in Table 5). A system without any candidate classification, i.e. merely enriching the terminology with all translations contained in the phrase table, is listed as a *naïve baseline* and trivially dominates recall. We also list a *cut-off based system* discarding all but the most probable *n* translations of an English term (*n*=1,3,5), an approach that achieves medium F_1 scores. These cut-off selection systems have recall values comparable to the baseline, indicating a strong capability of the aligment step to identify relevant phrases. Finally, we list the *classifier-based systems* considering different feature combinations. The versions with all features described in Section 3, i.e. incorporating translation probabilities, NER termhood, length scaling, and similarity measures, performed consistently best by F_1 scores. Especially for German, uniformly reordering words during similarity calculation improved recall and outperformed the F_1 score for other feature combinations. Classifier systems without translation probabilities as a feature (classifier 1) performed under par, whereas systems using only this information (classifier 2) performs close to the overall best systems, indicating again the high quality of the phrases generated by the SMT system.

Performance is comparable over all languages, with Dutch suffering from tremendous precision problems while achieving above-average recall—this could be caused by the low number of traceable Dutch terms (i.e. examples for classifier training; see Table 4) or the smaller corpus size (see Table 2). The former guess is more likely, since Spanish achieves the overall best F_1 scores and precision values despite a medium corpus size, yet has the highest number of traceable terms. The effect of the NER feature on the performance of classifiers 3 and 4 seems to be strongly language-dependent, yet there is no clear connection to the F_1 scores reported in Table 3, with Spanish and Dutch profiting equally little despite being the best and worst, respectively, in terms of NER performance.

Table 5. Evaluation of the re-invention performance by considering the UMLS as a gold standard, with F_1 score (F), Precision (P) and Recall (R) for all four languages. We list a naïve baseline (without candidate filtering), cut-off selection systems picking up the top *n* candidates (by direct phrase translation probability) for each term and classifier-based systems with different feature combinations.

Method	German			French			Spanish			Dutch		
	F	P	R	F	P	R	F	P	R	F	P	R
Naïve Baseline	0.14	0.07	**0.98**	0.13	0.07	**0.98**	0.19	0.11	**0.97**	0.11	0.06	**0.95**
Cut-off: top 1	0.40	0.25	0.96	0.37	0.23	0.97	0.48	0.32	0.95	0.39	0.25	0.91
Cut-off: top 3	0.22	0.12	**0.98**	0.22	0.13	**0.98**	0.28	0.16	**0.97**	0.19	0.10	0.94
Cut-off: top 5	0.19	0.18	**0.98**	0.18	0.10	**0.98**	0.25	0.14	**0.97**	0.15	0.08	**0.95**
Classifier 1 (NER only)	0.18	0.36	0.13	0.31	0.50	0.22	0.28	0.48	0.21	0.14	0.29	0.10
Classifier 2 (phrase translation probabilities only)	0.62	0.55	0.72	0.54	0.55	0.54	0.72	0.63	0.84	0.55	0.40	0.84
Classifier 3 (Classifier 2 + NER termhood)	0.65	0.61	0.70	0.59	0.57	0.61	0.72	0.62	0.86	0.54	0.40	0.84
Classifier 4 (Classifier 3 + length ratio)	0.67	0.60	0.75	0.62	0.60	0.63	0.74	0.65	0.87	**0.56**	**0.41**	0.88
Classifier 5 (Classifier 4 + similarity)	0.68	0.63	0.74	**0.73**	**0.64**	0.85	**0.80**	**0.70**	0.94	**0.56**	**0.41**	0.89
Classifier 6 (Classifier 4 + re-ordered similarity)	0.68	0.63	0.74	0.72	**0.64**	0.83	**0.80**	**0.70**	0.93	**0.56**	**0.41**	0.89
Classifier 7 (Classifier 4 + uniformly reordered similarity)	**0.72**	**0.64**	0.81	0.72	**0.64**	0.83	**0.80**	**0.70**	0.93	**0.56**	**0.41**	0.90

2. Manual Evaluation: Expert Judgment for Term Mining

We performed a manual evaluation for classifier 4 and 7 as terminology acquisition systems using phrase translation probabilities, NER termhood, string length scaling and, in case of classifier 7, string similarity based on uniformly reordered words as features. Both systems found a comparable number of entirely new terms and additional synonyms for each of the four languages. Classifier 7 (with string similarity) came up with 23k, 19k, 17k and 12k new terms for German, French, Spanish and Dutch, respectively, whereas classifier 4 (without string similarity) produced 23k, 18k, 19k and 12k new terms for these languages. Table 1 reveals the over-all lexical extension rate, Table 6 contains a detailed breakdown of the output of classifier 4. Results are similar for German and French, yet the much smaller size of the Dutch corpus leads to a higher number of entirely new terms while the inverse seems to be true for Spanish, probably due to the comparably larger *a priori* coverage of the Spanish UMLS.

In order to evaluate the terms and synonyms not yet contained in the UMLS we elicited judgments from two experts, both native speakers of German (with decent knowledge of the English biomedical terminology) and graduate students in the biomedical domain. They judged a random sample of 100 German terms each for classifier 4 and 7 according to the following three categories:

- Correct translation and perfectly suited as a terminology entry (including abbreviations and loan words);
- Correct translation, yet not directly suited as a terminology entry, e.g. due to vagueness or deviation from inflection norms (only nominative singular and plural terms are permitted);
- Incorrect translation.

Based on these samples 63% of the terms extracted by classifer 7 and 80% of those extracted by classifer 4 can be assumed to be both correct translations and suited as terminology entries. Thus, string similarity measures seem to be particularly apt for the reconstruction of the UMLS, yet they are not really helpful for the mining of novel terms.

A big chunk of the terms (49 items) produced by classifier 4 are spelled identically as an already known English term for the same concept, e.g. the name of the tuna species '*Thunnus thynnus*'. This is caused by a large number of entries marked as "English" in the UMLS, despite being used worldwide to denote the same concept. Furthermore, 8 terms are linguistically correct translations of an English term, although not yet suited as terminology entries, e.g. the inflected form '*linken Nebenniere*' (accusative) instead of '*linke Nebenniere*' ('*left adrenal gland*', nominative). Only a small portion of the errors of our system (6) raise serious concerns as mistranslations, e.g. caused by partial matches during phrasal alignment, such as with '*Flussmessung*' ('flow measurement') as a synonym for '*cardiac flow*'. The rate of serious errors for classifier 4 was comparable to that of classifier 7 (5 items); partial matches like *'intermittens'* as a synonym for *'claudication'/'limping'* due to their combined appearance in *'claudicatio intermittens'* ['temporary limping'] are again a main reason for errors. The 17-word gap between both classifiers was primarily caused by the judges' disagreement on the terminological validity of terms in 12 cases. Six of these were due to inconsistent use of annotation guidelines regarding inflected forms, e.g. the plural form '*Osteopenien*' ['bone losses'] was rejected by one judge, two were disagreements regarding the termhood of names, e.g. the protein nickname *'Harakiri'*, and four were caused by disagreement regarding abbreviations like *'CA2'*, a brain areal.

Table 6. Number of terms extracted by classifier 7 (see Table 5) with all features for all four languages. We distinguish three types of extracted items: "already known terms and synonyms" are already contained as a label for this concept and language (*term re-invention*), "additional synonyms" supplement terms/synonyms known for this concept and language by new ones (*term enrichment*), "entirely new terms" comprise really novel terms for a concept in that language, i.e. these concepts were previously unlabeled (*term mining*).

Language	Already known terms and synonyms	Additional synonyms	Entirely new terms
German	11k	11k	12k
French	11k	9k	10k
Spanish	18k	15k	3k
Dutch	4k	4k	8k

Discussion

The terminology acquisition system we developed for German, French, Spanish and Dutch parts of a UMLS-derived terminology yielded 23k, 19k, 18k and 12k additional synonyms (for terms already listed in the terminology) and entirely new terms (for concepts lacking any official label in this language), respectively. The extension rate for the corresponding languages relative to the already existing resources are 19.2%, 13.9%, 2.8%, and 10.3%, respectively.

About 80% of these new terms can be assumed to be correct and reasonable for this domain based on human judgments on a sample of 100 newly acquired German term candidates. We also showed that similarity features as proposed by Aker et al.[16] might be useful for reconstructing portions of the UMLS, yet they may be disadvantageous for the extraction of new biomedical terms.

For our experiments, we benefited a lot from the availability of parallel corpora. However, in order to further boost the rates for term mining we might have to enlarge our set of text corpora substantially. Using texts from other sources, like webpages[13] or Wikipedia[27], including the use of comparable corpora,[15,17] should be straightforward alternatives to improve our system and decouple it from its reliance on strictly parallel corpora. We also plan to incorporate alternative classifiers (such as CRFs or SVMs) with an enriched feature set, in particular, including "phrasal" types of features such as term collocations[28] and additional indicators of termhood such as term extraction metrics.[29]

Acknowledgements. This work was funded by the European Commission's 7th Framework Programme for small or medium-scale focused research actions (STREP) from the Information Content Technologies Call FP7-ICT-2011-4.1, Challenge 4: Technologies for Digital Content and Languages, Grant No. 296410.

References

1. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research. 2004;32(Database issue):D267–D270.
2. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters, W, et al. The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology. 2007;25(11):1251–1255.
3. Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, Musen, MA. BIOPORTAL: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Research. 2011;39(Web Server Issue):W541–545.
4. Balk EM, Chung M, Chen ML, Chang LKW, Trikalinos TA. Data extraction from machine-translated versus original language randomized trial reports: a comparative study. Systematic Reviews. 2013;2:97.
5. Turner AM, Mandel H, Capurro D. Local Health Department translation processes: potential of machine translation technologies to help meet needs. In: AMIA 2013 – Proceedings of the Annual Symposium of the American Medical Informatics Association. Washington, D.C., November 16-20, 2013; 2013. pp. 1378–1385.
6. Rebholz-Schuhmann D, Clematide S, Rinaldi F, Kafkas Ş, van Mulligen EM, Bui, C, et al. Entity recognition in parallel multi-lingual biomedical corpora: the CLEF-ER laboratory overview. In: Forner P, Müller H. Paredes R, Rosso P, Stein B, editors. Information Access Evaluation. Multilinguality, Multimodality, and Visualization – Proceedings of the 4th International Conference of the CLEF Initiative, CLEF 2013. Valencia, Spain, September 23-26, 2013. Lecture Notes in Computer Science, 8138. Berlin: Springer; 2013. pp. 353–367.
7. Koehn P, Och FJ, Marcu D. Statistical phrase-based translation. In: HLT-NAACL 2003 – Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. Vol. 1. Edmonton, Canada, May 27 - June 1, 2003; 2003. pp. 48–54.
8. Wu C, Xia F, Deléger L, Solti I. Statistical machine translation for biomedical text: are we there yet? In: AMIA 2011 – Proceedings of the Annual Symposium of the American Medical Informatics Association. Improving Health: Informatics and IT Changing the World. Washington, D.C., USA, October 22-26, 2011; 2011. pp. 1290–1299.
9. Tiedemann J. News from OPUS: a collection of multilingual parallel corpora with tools and interfaces. In: Nicolov N, Bontcheva K, Angelova G, Mitkov R, editors. Recent Advances in Natural Language Processing. Vol. V. Amsterdam, Philadelphia: John Benjamins; 2009. pp. 237–248.
10. Déjean H, Gaussier E, Renders JM, Sadat F. Automatic processing of multilingual medical terminology: applications to thesaurus enrichment and cross-language information retrieval. Artificial Intelligence in Medicine. 2005;33(2):111–124.

11. Deléger L, Merkel M, Zweigenbaum P. Enriching medical terminologies: an approach based on aligned corpora. In: Hasman A, Haux R, van der Lei J, De Clercq E, Roger France FH, editors. MIE 2006 – Proceedings of the 20th International Congress of the European Federation for Medical Informatics. Ubiquity: Technologies for Better Health in Aging Societies. Maastricht, The Netherlands, August 27-30, 2006. Vol. 124 of Studies in Health Technology and Informatics. Amsterdam: IOS Press; 2006. pp. 747–752.
12. Deléger L, Merkel M, Zweigenbaum P. Translating medical terminologies through word alignment in parallel text corpora. Journal of Biomedical Informatics. 2009;42(4):692–701.
13. Resnik P, Smith NA. The Web as a parallel corpus. Computational Linguistics. 2003;29(3):349–380.
14. Laroche A, Langlais P. Revisiting context-based projection methods for term-translation spotting in comparable corpora. In: COLING 2010 – Proceedings of the 23rd International Conference on Computational Linguistics. Beijing, China, 23-27 August 2010; 2010. pp. 617–625.
15. Skadja I, Aker A, Mastropavlos N, Su F, Tufiş D, Verlic M, et al. Collecting and using comparable corpora for statistical machine translation. In: LREC 2012 – Proceedings of the 8th International Conference on Language Resources and Evaluation. Istanbul, Turkey, May 21-27, 2012; 2012. pp. 438–445.
16. Aker A, Paramita M, Gaizauskas R. Extracting bilingual terminologies from comparable corpora. In: ACL 2013 – Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Volume 1: Long Papers. Sofia, Bulgaria, August 4-9 2013; 2013. pp. 402–411.
17. Ştefănescu D. Mining for term translations in comparable corpora. In: BUCC 5 – Proceedings of the 5th Workshop on Building and Using Comparable Corpora: Language Resources for Machine Translation in Less-Resourced Languages and Domains @ LREC 2012, the 8th International Conference on Language Resources and Evaluation. Istanbul, Turkey, 26 May 2012; 2012. pp. 98–103.
18. Weller M, Gojun A, Heid U, Daille B, Harastani R. Simple methods for dealing with term variation and term alignment. In: TIA 2011 – Proceedings of the 9th International Conference on Terminology and Artificial Intelligence. Paris, France, 8-10 November 2011; 2011. pp. 87–93.
19. Bouamor D, Semmar N, Zweigenbaum P. Identifying bilingual multi-word expressions for statistical machine translation. In: LREC 2012 – Proceedings of the 8th International Conference on Language Resources and Evaluation. Istanbul, Turkey, May 21-27, 2012; 2012. pp. 674–679.
20. Vintar, Š. Bilingual term recognition revisited; the bag-of-equivalents term alignment approach and its evaluation. Terminology. 2010;16(2):141–158.
21. Lefever E, Macken L, Hoste V. Language-independent bilingual terminology extraction from a multilingual parallel corpus. In: EACL 2009 – Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Athens, Greece, March 30 - April 3, 2009; 2009. pp. 496–504.
22. Hellrich J, Hahn U. Exploiting parallel corpora to scale up multilingual biomedical terminologies. In: Andersen S, Hasman A, Lovis C, Pape-Haugaard L, Saka O, Séroussi B, editors. MIE 2014 – Proceedings of the 25th Medical Informatics in Europe Conference. e-Health: for Continuity of Care. Istanbul, Turkey, August 31 - September 3, 2014. Vol. 205 of Studies in Health Technology and Informatics. Amsterdam: IOS Press; 2014, pp. 575-578.
23. Och FJ, Ney H. A systematic comparison of various statistical alignment models. Computational Linguistics. 2003;29(1):19–51.
24. Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, et al. MOSES: open source toolkit for statistical machine translation. In: ACL 2007 – Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL) on Interactive Poster and Demonstration Sessions. Prague, Czech Republic, 25-27 June 2007; 2007. pp. 177–180.
25. Hahn U, Buyko E, Landefeld R, Mühlhausen M, Poprat M, Tomanek K, Wermter J. An overview of JCORE, the JULIE Lab UIMA Component Repository. In: Proceedings of the LREC '08 Workshop "Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP". Marrakech, Morocco, 31 May 2008; 2008. pp. 1–7.
26. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. ACM SIGKDD Explorations. 2009;11(1):10–18.
27. Bouamor D, Popescu A, Semmar N, Zweigenbaum P. Building specialized bilingual lexicons using large scale background knowledge. In: EMNLP 2013 – Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. A meeting of SIGDAT, a Special Interest Group of the ACL. Seattle, WA, USA, 18-21 October 2013; 2013. pp. 479–489.
28. Wermter J, Hahn U. Paradigmatic modifiability statistics for the extraction of of complex multi-word terms. In: HLT/EMNLP 2005 – Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing. Vancouver, B.C., Canada, 6-8 October 2005; 2005. pp. 843–850.

29. Frantzi KT, Ananiadou S, Mima H. Automatic recognition of multi-word terms: the C-value/NC-value method. International Journal on Digital Libraries. 2000;3(2):115-130.

Problem Management Module: An Innovative System to Improve Problem List Workflow

Chad M. Hodge, MS[1, 2], Kathryn G. Kuttler, PhD[2], Watson A. Bowes III, MD, MS[1, 2], Scott P. Narus, PhD[1, 2]

[1]Department of Biomedical Informatics, University of Utah, Salt Lake City, UT
[2]Intermountain Healthcare, Salt Lake City, UT

Abstract

Electronic problem lists are essential to modern health record systems, with a primary goal to serve as the repository of a patient's current health issues. Additionally, coded problems can be used to drive downstream activities such as decision support, evidence-based medicine, billing, and cohort generation for research. Meaningful Use also requires use of a coded problem list. Over the course of three years, Intermountain Healthcare developed a problem management module (PMM) that provided innovative functionality to improve clinical workflow and boost problem list adoption, e.g. smart search, user customizable views, problem evolution, and problem timelines. In 23 months of clinical use, clinicians entered over 70,000 health issues, the percentage of free-text items dropped to 1.2%, completeness of problem list items increased by 14%, and more collaborative habits were initiated.

Background

A problem list is a fundamental component of an electronic health record (EHR), serving as a thumbnail view of ongoing diagnoses, findings, symptoms, and other health concerns that pertain to an individual patient [1]. When clinicians maintain an accurate problem list, patients generally receive better care [2], have less frequent omissions of care [3], and benefit from more frequent adherence to evidence-based care. Clinicians also benefit from using the problem list by receiving clinical decision support (CDS) alerts [4] that guide them to better outcomes, viewing detailed information about a disease via infobuttons [5], and more easily referring to a list of ongoing health issues, without searching through past notes. Furthermore, medical researchers benefit from an accurate problem list as they seek/mine for specific diseases in order to generate study cohorts and conduct comparative effectiveness research based on diagnoses and other patient health issues [6–9]. Effective use of a problem list is necessary for certification through Joint Commission [10], ONC HIT certification [11], and American Recovery and Reinvestment Act/Meaningful Use (ARRA/MU) [12], which can ultimately lead to incentive pay and/or avoidance of penalties.

Despite their many benefits, problem lists have been found to be sub-optimal, being inaccurate, incomplete, and out of date [13]. Studies have shown that even major health issues such as coronary artery disease and hypertension have a 49-51% probability of inclusion on a patient's problem list [2,14]. There is disagreement about what actually belongs on a problem list [15], and who owns the list, consequently leaving clinicians often unwilling to update a problem to its proper state [2] even if they know it to be inaccurate. Adding a duplicate problem to the list rather than updating an existing item adds clutter, which further reduces the value of the problem list. This behavior can be compounded by the fact that many organizations maintain multiple problem lists per patient - often segregated by discipline, care site, or acute/chronic state of the problem.

At Intermountain Healthcare (IH), an electronic problem list has been available for almost 20 years [16]. However, low utilization of the problem list, high prevalence of free-text problems, the need to replace an aging problem list system in the NICUs, and Meaningful Use (MU) requirements prompted the design of a new problem management module (PMM) that expanded the notion of the problem list to include a wider scope of health issues and workflow associated with them. When this project began in January 2011, only 47% of patients had any problems on their list. In fact, as a whole, IH averaged 8 problems per patient, with a median of just 1 problem per patient, 72.5% of which were coded. In the entire IH inpatient population, only 34% had an active, coded problem as required by MU.

Objective

The objective of this paper is to describe the PMM developed at IH and the functionality designed to overcome the following limitations of the prior system: underutilization, duplicate entry, out-of-date problems, free-text problems, missing attributes, inability to find the proper problem in a large list, cumbersome addition of new problems, an inability to tell the clinical story, and problem maintenance tension between care givers, resulting in incorrect

problem status. We present new and unique functional improvements incorporated into the PMM, and present quantitative results of pilot implementations at IH care locations.

Materials and Methods

Setting

The PMM was built at IH, in Salt Lake City, Utah, a not-for-profit integrated health care delivery network. IH operates 22 hospitals across Utah and southern Idaho and employs over 33,000 employees, 1,000 of whom are multispecialty clinicians working in 185 ambulatory clinics [17].

In order to meet the specific needs of the clinicians, and IH in general, a steering committee was formed, comprised of clinicians from different disciplines and care settings. This committee met every week, along with representatives from informatics, terminology, data modeling, knowledge management, clinical analysis, user interface design, and software engineering. Weekly meetings focused on forthcoming work, demonstration of the current state of the PMM, and resolution of issues that arose during the development process. Additionally, several structured user-acceptance testing sessions occurred with clinicians from many specialties. This iterative, user-centric, participatory approach culminated in the PMM tool that was initially released to a select, primarily inpatient audience at IH, including newborn and adult Intensive Care Units (ICU), acute care floors, and Tele-Health. The goal of the participatory design was to develop a common model that could be used across diverse care settings.

For quantitative results, we queried usage statistics for the PMM and legacy problem list application from our enterprise data warehouse and terminology logs for the period from the original pilot implementation on August 24, 2012, to the most recent data available on July 21, 2014. In some cases we compared data from before PMM implementation to post-implementation. Simple statistics were generated by the authors from these data. IRB approval was obtained prior to study initiation.

System Description

Master List

When opening the PMM, a longitudinal list of problems is presented, containing all problems from birth to death, whether active, inactive, acute, chronic, or historical (Figure 1). Viewing historical problems alongside active problems made it easier to ascertain a picture of a patient's overall health status. However, for the master list to be useful and not overwhelming, focus on data presentation and mechanisms for sorting, filtering, and grouping was needed.

One such mechanism aggregated multiple instances of the same problem, reducing overall length of the master list. Once aggregated, only the most recent instance of the problem was visible, with a superscript denoting the actual number of instances. Clicking the superscript opened a list of all prior occurrences with their detail.

Another mechanism for managing data was custom sorting. Each list column defined a custom sort order that tailored sorting of problems to an individual clinician's needs. When applied to body system, a clinician may choose to see cardiovascular-related problems first, respiratory-related problems second, and fluids/electrolytes-related problems last. Once set, data are automatically sorted according to the clinician's own preference.

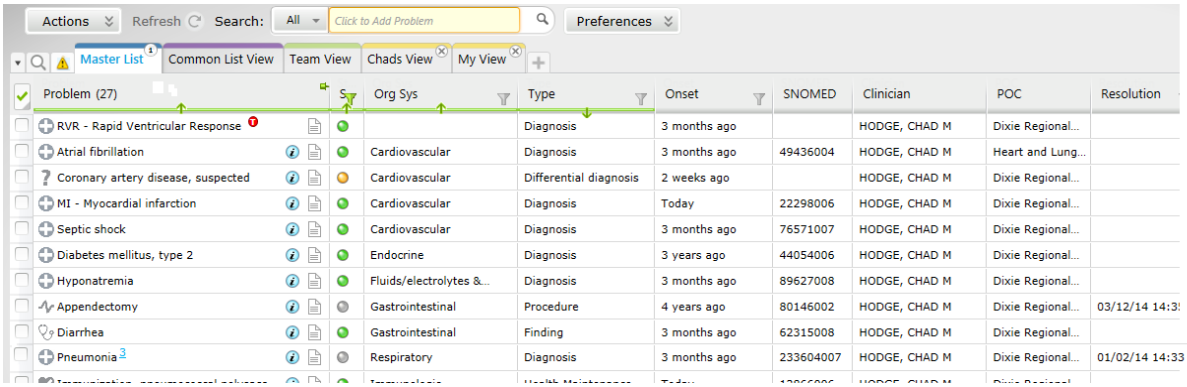


Figure 1 - Master problem list view within the PMM.

Smart Search

Scrolling through a long list of problems on the master list was a tedious task. IH created the Smart Search field to be the integral means by which clinicians searched for existing problems within the problem list, and as the sole means of adding new problems. Smart Search allowed clinicians to find a problem of interest by typing in a portion of the problem description. The description was used to find matches in three data sources: the patient's master list, the terminology database, and the clinician's common list. The results from searches against these three sources were then merged, ordered, and displayed as a single result set (Figure 2).

Searching the master list first allowed identification of the patient's existing problems. Presenting existing problems helps to keep the problem list uncluttered and up to date. An active instance of a problem cannot be added to the list until the current instance is resolved.

The second source searched is the terminology database. The search term is matched against a constrained domain of problem concepts that have been explicitly included based on identified needs and historical usage patterns. The matches are also translated into possible synonyms, so closely related terms and acronyms may be quickly found. For example, if a clinician enters 'heart', the search will return exact matches with the word 'heart', but also synonyms or acronyms, such as CHF, myocardial infarction, cardiomyopathy, etc. Search results are pre-coordinated so that body system, problem type, and status are automatically selected for the clinician [18]. The pre-coordinated attributes for each problem are available in an XML document residing in our knowledge repository [19].

The third source searched is the clinician's common list, consisting of a clinician's most commonly used problems. Matches against the common list are used to prioritize search results, so the clinician's preferred terms (typically specialty or setting specific) are shown at the top of the result set. This allows faster identification of higher-interest problems that a patient may have. The search results are limited to 40 items to avoid overwhelming clinicians. Clicking on one of the search results will add the pre-coordinated problem to the patient's problem list. Problems are mapped to SNOMED-CT terms where possible.

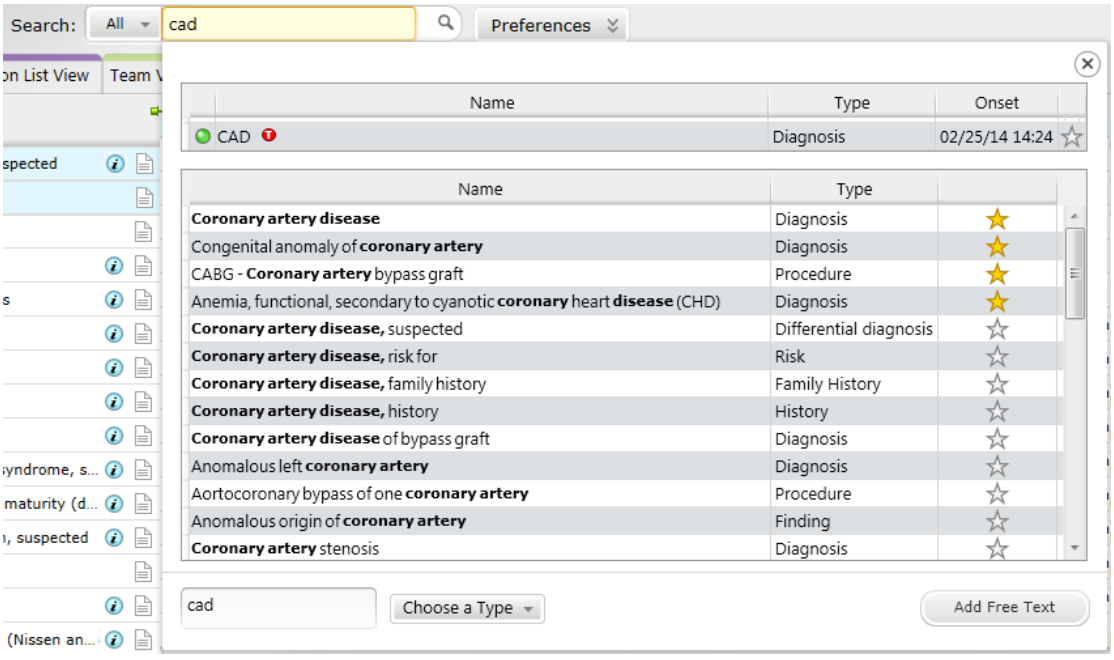


Figure 2 - Smart Search results. Existing problems are shown first, followed by matches from terminology, both results are ordered by common list preference

Feedback Mechanism

IH works with clinicians prior to implementation of vocabularies to define, pre-coordinate, and load only the concepts and representations that are clinically relevant, as evidenced by usage and clinician review. Post-implementation, clinicians require new content, and request synonyms for existing content that allow for efficient

searching. When terminology is incomplete, clinicians revert to adding patient data as free-text rather than as coded concepts. Current processes for handling terminology requests are ad-hoc and entail months of effort. This can leave clinicians disengaged from content governance, and patient data remain uncoded and thus unavailable for decision support. To address these issues, IH developed a new terminology feedback process to engage clinicians. Novel and efficient mechanisms were developed to identify gaps and allow clinicians to interactively review and approve recommended content in the context of patient care.

When clinicians added a free-text problem, a structured request was automatically routed to the terminology work queue. Clinical modeling engineers (CME) then inspected the free-text term, and searched for and evaluated matches for the term in the existing dictionary, to determine if the term was misspelled, an acronym, a synonym, or missing. CMEs then added the new content or created mappings between the synonym/acronym and root concept, ensuring future searches return proper results. CMEs then created a list of candidate substitutions for the clinician's free-text problem, and sent that list back to the application for review by the clinician.

The next time a clinician opened the same patient's problem list, the application presented the clinician with the substitution candidates for the free-text problem (Figure 3). When reviewing, the clinician viewed the original free-text alongside the list of candidates. The clinician could select a single choice to serve as the replacement for the free-text problem, or could choose to reject all provided choices with an accompanying reason. If rejected, the process was performed one more time, resulting in a new set of candidates. If the clinician chose to accept one of the newly provided terms, the free-text problem was automatically replaced by the coded concept.

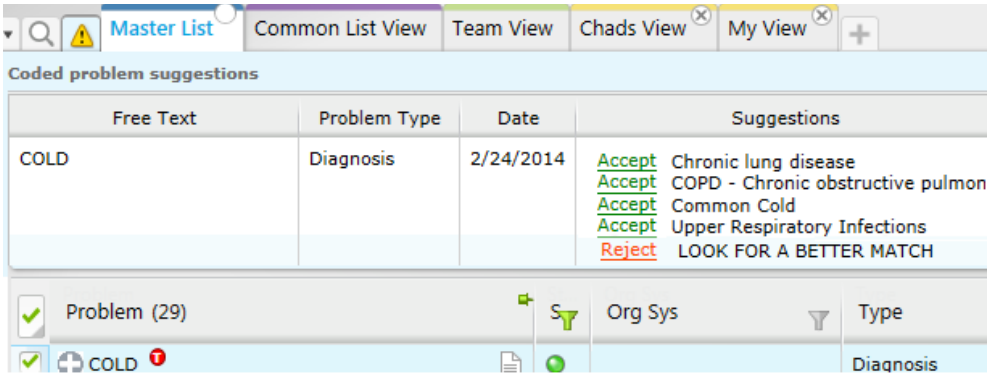


Figure 3 - Feedback Mechanism presenting a clinician with suitable substitutes for an entered free-text problem.

Problem Evolution

Problems tend to follow a predictable progression, or evolution. When treating a patient, clinicians begin creating a mental list of differential diagnoses, ruling out or confirming each one as data become available. Once confirmed, a differential diagnosis is evolved to a diagnosis, meriting treatment and monitoring. When a diagnosis is resolved, if a clinician decides that a reminder of the condition should be available to other caregivers, the diagnosis is evolved to a historical problem.

To support the evolution of problems, IH worked with a team of clinicians to enumerate 10 distinct problem types (Table 1). Through experience, IH clinicians felt that the problem list was the logical place to record and communicate these various types of health issues managed by their interdisciplinary teams. (*Health issue* is the more correct term for items on the PMM list, but we will interchangeably refer to them as *problems* throughout the paper for brevity.) A mapping between the types was created, defining the evolution pathway. For each coded concept in the problems terminology domain, a knowledge document was created that described how the specific problem may progress along the evolution pathway. Each of the 10 problem types was given a set of unique statuses that, when used, automatically triggered the progression of a problem along its evolution pathway.

The evolution document describes possible evolution pathways and the coded concepts that represent an evolved problem. When a clinician changes a problem's status to one defined in the document, evolution is triggered. For example, when a differential diagnosis is 'Confirmed' the problem automatically evolves to a diagnosis concept. The new diagnosis concept is added to the problem list and linked to the existing differential diagnosis. Only the most recent evolution for the problem is shown in the list, though all linked previous evolutionary states may be seen in the change history (Figure 4).

Once the diagnosis is resolved, a clinician may choose to evolve the problem to the terminal state of *historical* problem. Selecting the status of 'Resolved, Create History' will trigger a traversal of the evolution document one more time, locating the equivalent concept for the historical problem. That new concept is added to the list and linked to the end of the problem chain.

Having so many different statuses for each type of problem was confusing to the clinicians. Thus, a second knowledge document was created that mapped each problem status to one of four overarching meanings: active, proposed, inactive, or error (Table 2). This equivalency mapping was used to present a color-coded status indicator, agnostic of the actual status wording. For instance, a 'Recommended' health maintenance issue and a 'Scheduled' procedure both mapped to a yellow indicator, meaning both were in a *proposed* state. This allowed clinicians to focus on the meaning of the status, not the wording of the status.

Table 1 - Problem Types & Statuses.

Type	Description	Possible statuses
Finding	Observation about the patient that is less than a diagnosis occurring in relation to a physical exam. E.g., Abdominal pain.	Active, Inactive, Proposed, Resolved, Error, *Resolved, Create History <Evolve>*
Procedure	An action taken on a patient's body, typically diagnostic or treatment related. E.g., Lumbar puncture.	Active, Cancelled, Candidate, Resolved, Error, Unconfirmed, Ordered, Scheduled, *Resolved, Create History <Evolve>*
Differential Diagnosis	A possible diagnosis that has yet to be confirmed or ruled out. E.g. Sepsis, suspected.	Working Up, Ruled out, Error, *Confirmed <Evolve>*
Diagnosis	A confirmed health issue, describing the nature of a disease, injury, or congenital defect that requires monitoring or treatment. E.g. Carotid artery stenosis, or Trisomy 8 mosaic syndrome.	Active, Inactive, Proposed, Resolved, Error, *Working Up <Devolve>, Resolved, Create History <Evolve>*
Health Maintenance	Care steps taken to maintain good health, focused on early detection and prevention. E.g., Well child check-ups or vaccinations.	Due, Error, Recommended, Up-to-date, Resolved, *Create History <Evolve>*
Risk	An event with high probability of occurrence. E.g., Falls Risk.	Active, No Longer of Concern, Proposed, Resolved, Error
Event	A specific major occurrence. E.g., Automobile accident.	Active, No Longer of Concern, Proposed, Resolved, Error, *Resolved, Create History <Evolve>*
Device	An implantable device designed for a specific function. E.g., Pacemaker, or Port-a-cath.	Unconfirmed, Present, Removed, Error, Removed, *Create History <Evolve>*
History Of	Any medical and surgical item that has resolved, but continues to be pertinent to patient care. E.g., Breast cancer, or organ transplant.	Validated, Unknown, Proposed, Resolved, Error
Social History	Lifestyle choices. E.g., smoking, alcohol, or illicit drug use.	Active, Unknown, Proposed, Resolved, Error, *Resolved, Create History <Evolve>*
Family History	Major diseases which were/are present in the patient, but which are present in family members, potentially increasing the likelihood the patient will acquire the disease. E.g., cystic fibrosis.	Validated, Unknown, Proposed, Error

Table 2 - Status Equivalency

Status Equivalent	Description
Active (Green)	The problem is currently present in the patient, or is actively monitored and treated.
Proposed (Yellow)	The problem is potentially of concern, but more work is needed to confirm.
Resolved / Inactive (Grey)	The problem is resolved or no longer a concern.
Error (Red)	The problem was added in error, perhaps on the wrong patient.

History:

Description	Problem Type	Status	Org Sys	Onset Date	Resolutic
Sepsis, history	History	Validated	Infectious Disease	02/17/14 15:07	03/12/14
Sepsis	Diagnosis	Resolved, create history	Infectious Disease	02/17/14 15:07	03/12/14
Sepsis	Diagnosis	Active	Infectious Disease	02/15/14 15:07	
Sepsis	Diagnosis	Active	Infectious Disease	02/15/14 15:07	
Sepsis, suspected	Differential diagnosis	Confirmed	Infectious Disease	02/13/14 15:07	
Sepsis, suspected	Differential diagnosis	Working up	Infectious Disease	02/13/14 15:07	

Figure 4 - Change history of Sepsis showing evolution.

New Problem Indicator

When a clinician sees a patient and updates the problem list, there is a sense of ownership over what is in the problem list. However, when the same patient is seen later, the list may be significantly different because other clinicians have seen the patient and added and updated problems. When this occurred, the sense of ownership over the list was lost because it was too difficult to easily determine what specifically had changed, and a clinician may no longer be willing to maintain the list, even if they know it is inaccurate. To combat this, an indicator was created that alerts clinicians to changes in a patient's problem list.

Each time a clinician views a patient's problem list, a time-stamp is logged for that clinician-patient combination. On subsequent views, the time-stamp is retrieved and compared against the last edit time of the patient's problems. A running count is generated that tracks how many problems were added or updated since that date. This number appears directly on the master list tab (Figure 5). The clinician can click on the number to get more details about what has changed since last review. When clicked, a popup appears listing newly added or updated problems and relevant details. Selecting a problem from the list navigates the clinician directly to the problem in the master list.

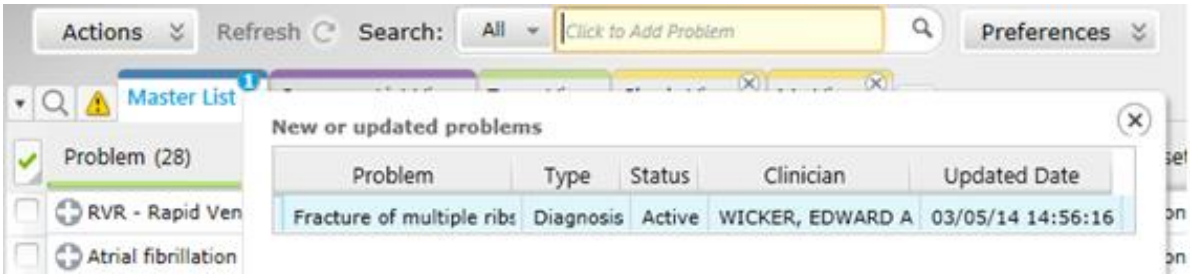


Figure 5 - New problem indicator. The superscript "1" on Master List tab indicates one new problem on the list.

Assessment and Plan

The prior version of the IH problem list application required clinicians to open a separate module to create the assessment and plan (A&P) for a problem. To find previous A&Ps written for a specific problem, one would have to review all recent clinical notes for any mention of the problem, which was a tedious task because previous A&Ps were sometimes copied forward without adding new information. IH clinicians expressed a desire to add and review A&P notes while reviewing a specific problem, using a single application.

To address this workflow issue, the PMM allowed an A&P to be created and stored for an individual problem. All A&Ps pertaining to a problem are listed in chronological order and are readily accessible from a note icon next to the problem. This makes it easier to determine progress towards treatment goals.

When an overarching clinical note needs to be created, such as a progress or discharge note, the latest A&P of each active problem was retrieved and automatically inserted into the note. This generated note contained much of the necessary detail of each problem, requiring only small changes before finalizing it.

No Known Problems

To meet MU requirements, functionality added to the PMM allowed attestation that a patient has "no known problems". If there are no MU-defined problems on a patient's list, the PMM displayed a link to a "no known problem" form (Figure 6). By using the form, the clinician affirms that the patient had no known problems. This attestation counted as a coded problem in the MU numerator. When attested to, the date and clinician name is shown. Once an actual problem was added to the problem list, the "no known problem" assertion was automatically

removed since it is mutually exclusive to coded problems on the problem list. If all problems were resolved, attestation would begin anew.

Patients may visit a clinician to address any of the health issue types shown in Table 1, allowing entry of a broader range of health issues than just diagnoses and medical history. For example, if the visit is for purely preventative reasons, a health maintenance problem can be added to the problem list. However, the numerator criteria set forth in MU guidelines do not recognize all 10 health issue types that PMM supports. In the case that none of the items on the patient's problem list meets the MU criteria for problems, the "no known problems" logic is initiated.

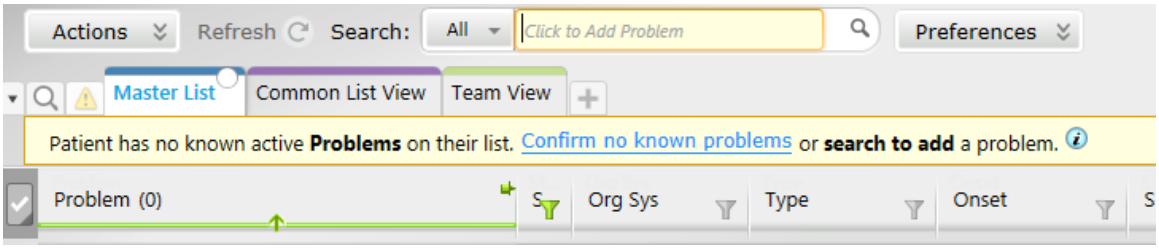


Figure 6 - No Known Problems indicator.

Timeline View

To improve the comprehension of complex problem lists, and to see patterns in a patient's health history, a timeline view of the problem list was created, similar to work reported by Plaisant et al. [20].The timeline view graphs health issues currently in view, in relation to time. The x-axis is time, starting at the patient's date of birth, minus 9 months, on the far left, until present day on the far right. The y-axis lists each distinct health issue visible on the patient's list, placing the earliest onset health issue on top, and the most recent onset health issue at the bottom. When multiple occurrences of the same health issue are present, they are mapped on the same line, and each occurrence is placed in the appropriate period on the x-axis. Each health issue is drawn as a rectangle with the onset of the health issue as the left edge, and the resolution date as the right edge of the rectangle, showing the lifespan of the issue. A patient's encounters/visits are overlaid on the graph to correlate them in time with health issues. Thin transparent bars represent a short visit to an outpatient provider, and a wider bar denotes a longer inpatient encounter (Figure 7).

The timeline view provides immediate detail about how many problems a patient has had, how long these problems lasted, what health issues occurred concurrently, if a caregiver was seen during that time, and when the most recent instance occurred. Long-standing problems will show prominently, prompting clinicians to inquire further. Additional patterns, such as problem overlaps, may emerge when viewing data on a lifetime scale that are otherwise not obvious.

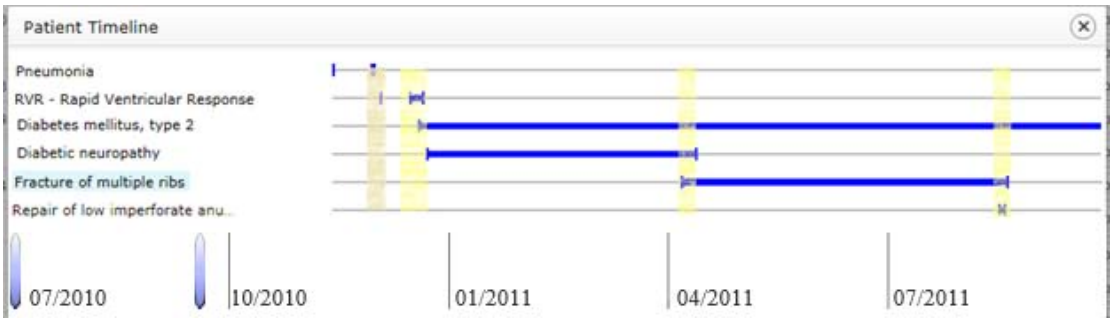


Figure 7 - Plotting problems in chronologic order for a holistic representation of the patient's problem history.

Views

During clinician testing of the PMM, it was observed that some clinicians occasionally became frustrated that there were problems on the list that did not specifically pertain to the patient's current visit. These frustrated clinicians chose to resolve all the problems on the list and then add problems specific to the current visit, rather than deal with the existing large problem list. This action is detrimental to the accuracy of the patient's problem list, and may

destroy years' worth of problem history. This phenomenon was termed *Tension*, because a clinician was acting as a force to deform the problem list, pushing it out of harmony with the patient's true health state.

To resolve the underlying cause of this behavior, *Views* were introduced (Figure 8). A View is simply a way for a clinician to separate problems from the large master list, narrowing to a focused problem list. A View references selected problems from the master list, allowing the clinician to choose only those problems that are of interest, imbuing the clinician with a sense of ownership over those few problems.

This concept reduced clinician's frustration with large lists; however, they became curious about what other clinicians' Views looked like for the same patient. To satisfy that curiosity, a search function was added that allowed discovery of other Views, and allowed them to be opened in read-only mode. The viewing of other clinician views was useful when dealing with referrals to a specialist.

Now that IH clinicians could create a list of problems for their own needs, they wanted to show the interrelatedness of those problems, to better express the clinical story. The PMM allowed clinicians to drag and drop one problem onto another, building a subsumption hierarchy in whatever manner best depicts the clinicians thought processes. When the PCP searched for the cardiologist's View, he or she would not only see the problems the specialist was concerned with, but also the relations between them. For example, the hierarchy might inform the PCP that the patient has underlying heart issues caused by poorly managed diabetes, prompting a discussion with the patient.

When presented with Views for individual clinicians, it was realized that an entire care team could also benefit from the Views concept. The Team View allows groups of clinicians in a specific care setting, such as an ICU or outpatient clinic, to define a set of problems with relationships. The Team View allows the entire care team to collaborate on what problems should be monitored and treated, giving all shifts and specialties a better understanding of the team's priorities. Just as personal views could be searched for and inspected in read-only mode, so too could Team Views. Using another team's view is useful when patients transfer from one team to another. The prior team's thoughts and work could be referred to before receiving the patient in a new unit so that preparations could be made, thereby improving handoffs.

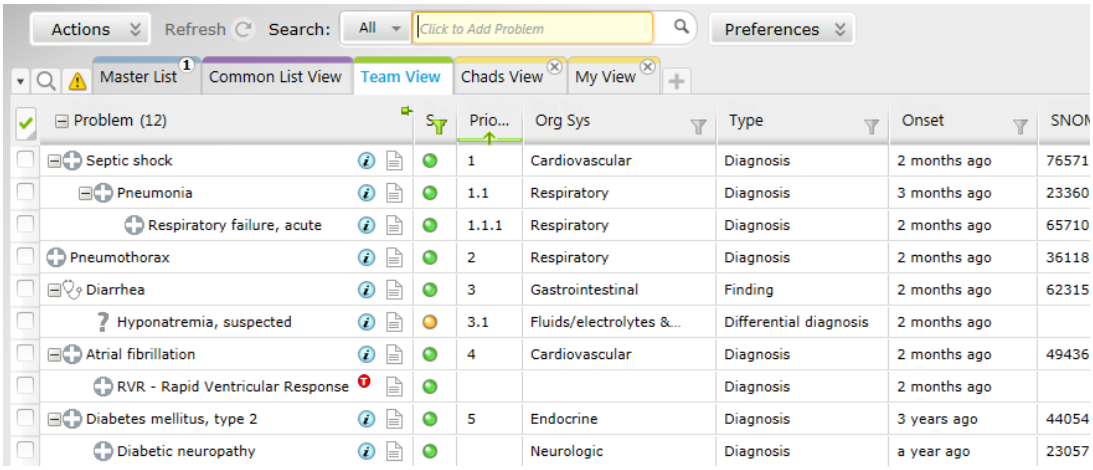


Figure 8 - A team view that depicts subsumed and prioritized problems for a care team.

Results

During the time period examined, 580 clinical users, including physicians, nurse practitioners, and physician assistants from 9 locations, have used the PMM. These users entered information on 5,557 patients, adding 71,838 total health issues. On average, 110 clinicians are entering or updating 3,900 health issues per month, with a peak of almost 200 clinicians entering over 4,700 issues during one month. Clinicians used the Assessment and Plan tool to attach over 997,000 A/P notes to problems on 5,079 patients.

Free-text problems entered using PMM were 1.2% of the total problems entered during the study period. During the same period, 11% of problems entered using the legacy problem list application were free-text. We also found that roughly half of the problem terms selected by users in the Smart Search tool contained pre-coordinated information.

The number of missing data elements has fallen from an average of 27.63 per 100 problems using the old problem list application, to 13.73 using the PMM, for fields such as body system, problem type, and status.

Not only are data more complete, problem evolution has added a new dimension of insight to the problem list. Using PMM, clinicians evolved 3,098 health issues, most commonly by changing a status to "Resolved, create history." But clinicians also evolved a differential diagnosis to a confirmed diagnosis 1314 times: The most frequently evolved differential diagnoses were *Jaundice due to preterm infant, suspected* (308), *Respiratory distress syndrome, suspected* (208), *Jaundice in a term infant, suspected* (175), and *Sepsis, suspected* (142).

Clinicians utilized views, subsumption hierarchies, and prioritization to organize problem lists in clinically-meaningful ways. Care-teams selected problems off the master list and built Team Views 9,352 times; individual clinicians created clinician-specific Views 116 times. Arbitrary segregation of problems into a View is one way of organizing problem data; an additional method is to subsume problems into a hierarchy of relatedness. The maximum depth any view subsumed a problem is 5 levels deep, with an average depth of 1.5 per view. When prioritizing problem data in a View, the largest number of problems prioritized was 18, with an average of 3. The most commonly prioritized problems were apnea, nasal cannula oxygen administration, hyperbilirubinemia, and septic shock.

Discussion

The large number of problems added per patient (average 12.9/patient) using the PMM can be explained in three ways. First, problem evolution results in many new problems being created automatically for a single issue as the original problem's status is changed to "resolved" and its newly evolved instance is created, which is not readily apparent to the clinical user. Second, most of the users of the PMM were in ICU settings, which average a much higher number of problems per patient than other care settings. Third, organizational incentives for clinicians to meet MU goals coincided with use of the PMM, possibly inflating results. Future studies will further compare use of the PMM with the legacy problem list application, which is still in use in certain locations (primarily outpatient clinics), to determine if the new features are the cause of a perceived improvement in problem management at IH. Workflow and clinician engagement improved as a result of the new problem terminology search and maintenance functionality. The large percentage of pre-coordinated terms selected in the Smart Search tool likely helped clinicians find appropriate terms and reduced free-text entry. The usefulness of the Feedback Mechanism was also evident and will be incorporated in other parts of our EHR.

Not only has the PMM created much more data than a typical problem list application, it has created entirely new types of data for study, such as how clinicians aggregate data in a view and evolve problems. It is evident from our usage statistics that clinicians in our currently ICU-centric pilot sites prefer to create team views rather than personal views. This is largely due to the fact that these clinicians use a team-oriented model for care in which information and treatment plans are shared. Future work will explore how clinicians and teams use Views, and the manner in which they impart order to the Views by subsuming problems into a hierarchy. Additional studies will investigate if Views lead to a more accurate problem list, and if they reduce tension between caregivers of the same patient.

The goal of the PMM project was to increase clinician adoption of the problem list by addressing issues associated with the use of problem lists in general. Each of the enhanced features described was designed to address one or more of these issues. Current usage statistics indicate that these interventions have been successful. However, a more formal study will be necessary in order to determine the exact impact of each of these enhancements on problem list usage and patient care, and their generalizability to more clinicians and settings. In the 23 months of use, feedback from clinicians has been extremely positive, affirming that the PMM provides more value, is easier to use, and offers innovative features that entice use of the system.

Acknowledgements

The envisioning and development of this application involved dozens of people over three years, spanning both technical and clinical domains. The authors would like to acknowledge the invaluable input of Jean Millar, Lisa Gleed-Thornton, Ed Wicker, Chris Wood, Larry D. Eggert, James Orme, Ning Zhuo, Annie Bigler, Harsha Puthalapattu, and Srinivasulu Reppale Venkat.

References

1. Weed LL. Medical Records That Guide and Teach. N Engl J Med [Internet]. 1968;278(11):593–600. Available from: http://www.nejm.org/doi/full/10.1056/NEJM196803142781105

2. Hartung DM, Hunt J, Siemienczuk J, Miller H, Touchette DR. Clinical implications of an accurate problem list on heart failure treatment. J Gen Intern Med [Internet]. 2005 Feb;20(2):143–7. PMID 15836547
3. Wright A, Maloney FL, Feblowitz JC. Clinician attitudes toward and use of electronic problem lists: a thematic analysis. BMC Med Inform Decis Mak; 2011 May 25;11(1):36. PMID 21612639
4. Wright A, Goldberg H, Hongsermeier T, Middleton B. A description and functional taxonomy of rule-based decision support content at a large integrated delivery network. J Am Med Inform Assoc. 2007 Jul-Aug;14(4):489–96. PMID 17460131
5. Cimino JJ, Elhanan G, Zeng Q. Supporting infobuttons with terminological knowledge. Proc AMIA Annu Fall Symp. 1997;528–32. PMID 9357682
6. Lin J-H, Haug PJ. Exploiting missing clinical data in Bayesian network modeling for predicting medical problems. J Biomed Inform. 2008 Feb;41(1):1–14. PMID 17625974
7. Abend A, Housman D, Johnson B. Integrating Clinical Data into the i2b2 Repository. AMIA - Summit on Translat Bioinformatics. 2009 Mar 1;2009:1–5. PMID 21347159
8. Bayley KB, Belnap T, Savitz L, Masica AL, Shan N, Fleming NS. Challenges in using electronic health record data for CER: Experience of 4 learning organizations and solutions applied. Med Care. 2013 Aug;51 (8 Suppl 3):S80-6. PMID 23774512.
9. Narus SP, Srivastava R, Gouripeddi R, Livne OE, Mo P, et al. Federating clinical data from six pediatric hospitals: Process and initial results from the PHIS+ consortium. AMIA Annual Symp Proc 2011;2011:994-1003. PMID 22195159
10. Joint Commission Resources. Comprehensive Accreditation Manual for Hospitals: The Official Handbook. Oakbrook Terrace, IL; 2008.
11. Department of Health and Human Services. Federal Register:45 CFR Part 170;75(144). Health Information Technology: Initial set of standards, implementation specifications, and certification criteria for electronic health record technology. 2010. [cited 2014 Jul 14]. Available from: http://www.gpo.gov/fdsys/pkg/FR-2010-07-28/pdf/2010-17210.pdf
12. Centers for Medicare & Medicaid Services. Eligible Hospital and Critical Access Hospital Meaningful Use Core Measures [Internet]. 2013 [cited 2014 Jan 22]. p. 22–4. Available from: http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/downloads/3_Maintain_Problem_List.pdf
13. Luna D, Franco M, Plaza C, Otero C, Wassermann S, Gambarte ML, et al. Accuracy of an electronic problem list from primary care providers and specialists. Stud Health Technol Inform. 2013;192:417–21. PMID 23920588
14. Szeto HC, Coleman RK, Gholami P, Hoffman BB, Goldstein MK. Accuracy of computerized outpatient diagnoses in a Veterans Affairs general medicine clinic. Am J Manag Care. 2002 Jan;8(1):37–43. PMID 11814171
15. Holmes C, Brown M, Hilaire DS, Wright A. Healthcare provider attitudes towards the problem list in an electronic health record: a mixed-methods qualitative study. BMC Med Inform Decis Mak [Internet]. BMC Medical Informatics and Decision Making; 2012 Nov 11;12(1):127. PMID 23140312
16. Clayton PD, Naus SP, Bowes WA, Madsen TS, Wilcox AB, Orsmond G, et al. Physician use of electronic medical records: issues and successes with direct data entry and physician productivity. AMIA Annu Symp Proc. 2005;2005:141–5. PMID 16779018
17. Fast Facts - Intermountain Healthcare - Salt Lake City, Utah [Internet]. [cited 2014 Mar 9]. Available from: http://intermountainhealthcare.org/about/overview/Pages/facts.aspx
18. Rosenbloom ST, Miller RA, Johnson KB, Elkin PL, Brown SH. Interface Terminologies: Facilitating Direct Entry of Clinical Data into Electronic Health Record Systems. J Am Med Inform Assoc. 2006 May-Jun; 13(3):277–88. PMID 16501181
19. Hulse NC, Galland J, Borsato EP. Evolution in Clinical Knowledge Management Strategy at Intermountain Healthcare. AMIA Annu Symp Proc 2012; 2012: 390–9 Epub 2012 Nov 3. PMID: 23304309.
20. Plaisant C, Mushlin R, Snyder A, Li J, Heller D, Shneiderman B. LifeLines: Using visualization to enhance navigation and analysis of patient records. Proc AMIA Annu Fall Symp; 1998:76–80. PMID 9929185

Reasoning Based Quality Assurance of Medical Ontologies: A Case Study

Matthew Horridge PhD[1], Bijan Parsia PhD[2], Natalya F. Noy[1] PhD, Mark A. Musen MD, PhD[1]
[1]Stanford University,CA, USA; [2]The University of Manchester, UK

Abstract

The World Health Organisation is using OWL as a key technology to develop ICD-11 – the next version of the well-known International Classification of Diseases. Besides providing better opportunities for data integration and linkages to other well-known ontologies such as SNOMED-CT, one of the main promises of using OWL is that it will enable various forms of automated error checking. In this paper we investigate how automated OWL reasoning, along with a Justification Finding Service can be used as a Quality Assurance technique for the development of large and complex ontologies such as ICD-11. Using the International Classification of Traditional Medicine (ICTM) – Chapter 24 of ICD-11 – as a case study, and an expert panel of knowledge engineers, we reveal the kinds of problems that can occur, how they can be detected, and how they can be fixed. Specifically, we found that a logically inconsistent version of the ICTM ontology could be repaired using justifications (minimal entailing subsets of an ontology). Although over 600 justifications for the inconsistency were initially computed, we found that there were three main manageable patterns or categories of justifications involving TBox and ABox axioms. These categories represented meaningful domain errors to an expert panel of ICTM project knowledge engineers, who were able to use them to successfully determine the axioms that needed to be revised in order to fix the problem. All members of the expert panel agreed that the approach was useful for debugging and ensuring the quality of ICTM.

Introduction

The World Health Organisation International Classification of Diseases (ICD)[1] is a large and widely used resource. The current version, ICD-10, which was published in 1990, is essentially a large hierarchical terminology that contains codes, titles, and textual descriptions of diseases. It is used in areas such as epidemiology in order to compile health statistics, for remuneration in various jurisdictions, to monitor health-related spending, and to guide health policy decisions. In 2007, the World Health Organization (WHO) initiated work on ICD-11. The development of ICD-11 is a large social/collaborative effort involving over 250 subject matter experts and is being carried out using the *iCAT* tool. iCAT, which stands for "ICD Collaborative Authoring Tool", is a highly customized version of WebProtégé[2] – part of the well known Protégé suite of tools – and allows users to simultaneously review, edit and extend ICD-11[†].

In contrast with all previous versions of ICD, it was decided by the WHO that ICD-11 would be represented in a well-structured, computable form, defined by a *Content Model*. The main motivation for this was the hope that it would make it possible for subject matter experts to view, edit and curate content in a variety of software tools. It was also hoped that it help to resolve ambiguities in terms of what information needed providing and would also allow some form of automated error checking and quality assurance to be implemented. Initially, the Content Model was a three layer UML model, which specified requirements such as "a term should have a title, a list of synonyms, a code etc.". However, the decision to re-implement the Content Model as an ontology was made, and the UML model was re-implemented using the widely used *Web Ontology Language* (OWL)[3]. The OWL *Content Model Ontology* essentially provides a "schema" for *data*, which corresponds to descriptions of diseases, to be entered. The entered data then forms the basis for ICD-11.

The main benefits of using ontologies, and in particular OWL, for the Content Model were stated to be that:

- Other ontologies and applications would be able to reuse ICD or portions of ICD more easily than if it was represented using some other language or semi-structured data format such as XML Schema. Thus, the chances of interoperability with other software systems is maximized;
- Expressive OWL constructs could be used to formalize the Content Model and could be used to generate the (form based) user interface for entering content;

[†] While iCAT is meant for ICD-11 content editors, end users of ICD-11 (and members of the public) may view nightly compilations at http://www.who.int/classifications/icd11

- OWL would be able to provide a path to future use of automated consistency checking and class hierarchy computation, which could, (a) be used for Quality Assurance purposes; (b) reduce hierarchy construction mistakes (placing something in the wrong place); and, (c) be useful for the purposes of migrating ICD from a pre-coordinated terminology to a post-coordinated terminology.

While the first two points have been proven to some extent[4], automated reasoning has not yet been used, either at Content Model Ontology design time, or at runtime when data is being entered into the system by subject matter experts. In this paper, we therefore explore how standard reasoning, in the form of consistency checking, and non-standard reasoning, in the form of justification finding, can be used to debug and explain inconsistencies that arise during the development of a Chapter of ICD. We look at the potential utility of such services and reflect on the current state of tool support. More specifically, we investigate the following: Are inconsistencies that arise during the development of real world complex medical ontologies indicators of meaningful errors in the ontologies? Is it possible to deal with a large number of errors? Are there regularities in the reasons for inconsistencies and if so can these regularities be exploited to triage the errors? Can off-the-shelf explanation tools provide pointers so that these errors can easily be corrected?

We focus our investigation on the development of Chapter-24 of ICD, which deals with the specification and description of Traditional (Chinese) Medicine and medical practices. This chapter was developed as part of the *ICTM* project, where a customized version of iCAT (iCAT-TM – herein simply referred to as iCAT) and a Content Model Ontology that extends the ICD Content Model Ontology were used to create it. The reasons for focusing on this Chapter are two-fold: (1) The chapter provides a smaller subset of ICD that is more convenient to experiment with compared to the complete ICD classification; (2) The development of the chapter uses the same underlying workflows, tooling and technologies that are being used to create other chapters of ICD-11. Thus, any results of experiments on this chapter should be generalizable to ICD as whole. Furthermore, while we use ICTM as a case study, the approach should be adaptable to any system that uses a logic based ontology as a schema for data entry.

Preliminaries – Ontologies, OWL, Reasoning and Justifications

Ontologies: Ontologies are machine processable artifacts that capture the important concepts, and the relationships that hold between them, in some domain of interest. Ontologies are used in many diverse domains, but they are heavily used and are important in the areas of medical-informatics. For example, SNOMED-CT[5] is a huge multilingual healthcare ontology. It provides a vocabulary for clinical findings, symptoms, diagnoses, procedures, and body organs amongst other medical domain knowledge.

OWL: The latest standard in ontology languages is The Web Ontology Language, OWL[3] (or to be more precise, OWL 2). The main units of currency in an OWL ontology are *axioms*. An axiom is basically a *statement* of truth. For example, Heart SubClassOf Organ is a subclass axiom, which states that every instance of Heart is also an instance of Organ. Axioms tie together *entities*, which are names for basic concepts and properties in the domain. An entity may be a *class*, *property*, or an *individual* (an individual may also be called an instance). In this paper we use the convention that, class names start with uppercase letters, and property and individual names start with lowercase letters. Furthermore, we write all entity names in a sans-serif font, and property names in particular in an *italic sans-serif font.* We can broadly categorize axioms into *TBox* axioms and *ABox* axioms. TBox axioms represent the "schema" part of an ontology – they are used to specify descriptions of classes. ABox axioms represent the "data" part of the ontology – they are used to specify descriptions of individuals.

Entailment: OWL is a *logic-based ontology language*, which means that the semantics of the language is precisely specified by a mapping to an underlying logic. This gives OWL axioms a well-defined, unambiguous meaning. It also allows the specification of a consequence, or rather *entailment*, relation. In OWL, it is possible to deterministically compute whether or not an axiom follows from, or is implied by, what has been stated in an ontology. For example, if an ontology states that, Heart SubClassOf Organ, Organ SubClassOf AnatomicalStructure. Then it is possible to determine that Heart SubClassOf AnatomicalStructure also follows from the ontology. If an axiom follows from an ontology we say that the axiom is *entailed* by that ontology. We also refer to an entailed axiom as an *entailment*.

Reasoning: OWL was designed with favorable computation properties in mind and it is possible to automate the computation of entailments using off-the-shelf tools called *reasoners*. Reasoners are software systems that can be used both at ontology design time and at application runtime. At design time, a reasoner can be used to check that certain desirable entailments follow from what has been stated, and, more typically, that certain undesirable entailments (logic bugs) do not follow from what has been stated. One of the most serious logic bugs that can occur

for an ontology is for it to be inconsistent[‡]. When an OWL ontology is inconsistent, it entails *anything*. In other words, any conclusion may be drawn from it and under classical semantics an inconsistent ontology is essentially meaningless.

Justifications (explanations): In this work we use a reasoning service that can generate explanations for inconsistent ontologies. These explanations are known as *justifications*. A justification is a minimal subset of an ontology that is sufficient for an entailment to hold. In the case of an inconsistent ontology, the axioms in a justification explain what has been stated that makes the ontology inconsistent. For example, the justification, J = {PatientC2 InstanceOf Man, PatientC2 InstanceOf Woman, DisjointWith(Man, Woman)} explains an inconsistency caused by the individual PatientC2 being an instance of the classes Man and Woman which also happen to be disjoint (disjoint classes cannot share any instances). There may be multiple justifications that explain why an ontology is inconsistent, reflecting the fact that there can be multiple reasons for a single inconsistency. Furthermore, these justifications may overlap with each other, that is, they may share axioms. Finally, justifications have emerged as the dominant form of explanation in the OWL world. This is because they are conceptually simple, there are off-the-shelf tools for computing them (which work with any OWL reasoner), and searching for the cause of logic bugs without some kind of pinpoint device like justifications is like looking for a needle in a haystack – that is, it is a wretched and error prone process.

Preliminaries – iCAT(-TM) and WebProtégé

The content for ICTM, which forms Chapter-24 of ICD-11, is created, edited and viewed with the *iCAT-TM* tool (ICTM Collaborative and Authoring Tool, herein simply referred to as iCAT). iCAT is a customization of WebProtégé, a Web-browser-based collaborative ontology development environment that can be used by geographically distributed users to simultaneously edit OWL ontologies.

Figure 1. A screenshot of an iCAT-TM form. Selecting the Disorder/Syndrome radio button at the top of the form causes property assertions to be added to the underlying ontology stating that the underlying individual, representing an ICD code has a type relationship to Disorder/Syndrome. The end user (an ICTM subject matter expert) does not need to know what an individual is, or what a property is, or what a property assertion axiom is. The grid at the bottom of the form shields the user from editing multiple complex expressions and axioms that define the given term.

A key feature of iCAT, and one that has been heavily used in the context of the ICTM project, is form based entity editing. Figure 1 shows an example of an iCAT form. Forms can consist of checkboxes, radio buttons, text fields, tables and the like. They provide a simple graphical way of editing OWL axioms under the hood. The advantage of this is that an end user of the system does not necessarily need to have any prior knowledge of OWL or ontologies to be able to contribute to the authoring of an ontology such as ICD or ICTM. In essence forms allow an ontology to be edited and extended "by stealth".

The forms that iCAT displays are declaratively specified using a custom markup language. This markup language specifies how form elements are rendered on the screen. It specifies the kind of widgets that should be used to display content, and also specifies how to translate specific input values of widgets into domain specific vocabulary and axioms in the underlying ontology. A key feature of the system is that it uses a loose, manually specified,

[‡] Note that if an ontology is inconsistent then we say that the whole ontology is inconsistent. That is, an ontology is either consistent or inconsistent. While there may be multiple subsets of an ontology that cause it to be inconsistent (that is, multiple reasons for the inconsistency), there is still only a single inconsistency.

coupling of forms to the underlying ontology. This provides a certain level of flexibility in translating the form inputs into ontology constructs.[§] Additionally, it makes it possible to provide tighter constraints on form entry values without having to over-constrain or pollute the ontology with baroque modeling choices that do not necessary sit well with languages like OWL. One drawback of the loose coupling is that the person who configures the forms has to take some care in ensuring that the forms do not allow data to be entered that could be *inconsistent* with axioms specified in the underlying ontology. For example, a form that allows multiple distinct values to be entered for a functional property (single valued property) in the ontology would result in an inconsistency.

Preliminaries – The ICTM Content Model Ontology

The *ICTM Content Model Ontology*, developed by the HiM-TAG, and herein simply referred to as the Content Model Ontology, is an expressive OWL ontology. The main purpose of the Content Model Ontology is to provide a formalization of a "schema" for the data that is entered into the system as ABox axioms via the iCAT forms. It is used by the internals to actually store data and, to some extent, it is used to design and guide the construction of the forms that are displayed to end users who author content.

While the Content Model Ontology was initially fairly simple, it has evolved considerably since the project began. New versions of the Content Model Ontology have been developed as modelers have used the iCAT tool and the requirements of what has to be captured have become clear. A Content Model Ontology update cycle typically starts by subject matter experts indicating that they cannot specify the necessary details for some type of entity. For example, they might say that they cannot specify the location for a disease. Next, a discussion ensues with the ICTM working group and engineers about how the required information would be best presented (on a form) in the system. Based on this, the Content Model Ontology is then extended and updated to capture what can be represented. Finally, the forms are altered to enable new data to be captured.

Using Reasoning and Justification Finding for Quality Assurance

Having introduced iCAT and discussed the relationship between the Content Model Ontology and the Forms Configuration used in iCAT, we now present an exploration of using off-the-shelf reasoning and explanation generation technology that we believe is useful for debugging and carrying out Quality Assurance checks on the combined Content Model Ontology, entered data and Forms Configurations in iCAT.

The investigation is split into two experiments. The first experiment (**EXPERIMENT-A**) involved computing, examining and categorizing justifications for an inconsistent version of the ICTM Content Model Ontology plus data. The goals of this experiment were to answer the questions: Are there a large number of justifications for the inconsistency? Do the justifications have any regularities or patterns that can be exploited in order to triage the problem? What is the makeup of the justifications – are axioms in them purely TBox axioms, or a mixture of TBox and ABox axioms? (Justifications that only contain TBox axioms are highly likely to indicate problems with the Content Model Ontology alone, while justifications that contain a mix of TBox and ABox axioms could indicate a problem with the Content Model Ontology, Data, or both). The second experiment (**EXPERIMENT-B**) involved semi-structured interviews of people on an expert panel that was made up from the team of researchers/engineers that work on the design and evolution of the Content Model Ontology and forms in iCAT. The purpose of the interview was to determine the nature of the problems that cause the inconsistencies (whether the problems lie in the Content Model Ontology, the Data or the Forms); to determine how they would go about fixing the inconsistencies; and, to ascertain whether they believe that the techniques used are generally useful and could fit into some future tool chain. The goals of this experiment were to answer the questions: Do the justifications for inconsistencies indicate meaningful errors to the Content Model Ontology engineers and Forms Configuration engineers? What is the nature of the errors? Can the justifications be used to easily correct or suggest a strategy for correcting the errors? What do the Content Model Ontology engineers and Forms Configuration engineers think about reasoning and justifications as a way of pinpointing errors?

EXPERIMENT-A - Materials: A revision of the ICTM Content Model Ontology plus entered content (data) from May 2013 was used in the experiments that follow. The revision (Content Model Ontology with data) was found to be

[§] It is worth noting that this is rather different to Protégé 3 (desktop) forms generation mechanism, where the forms are automatically directly-generated from the ontology. In Protégé 3, there is a tight coupling between the "schema level" of the underlying ontology and the forms used for data entry. In WebProtégé, the loose coupling it makes it possible to provide highly customized layouts and simple widgets for editing multiple complex constructs, which is either not easy or not possible with the Protégé 3.

inconsistent during examination with the HermiT reasoner (version 1.3.8)[**]. It should however be noted that the Content Model Ontology alone (separate from the entered content/data) was consistent. All experiments were performed on a 2.2 GHz Intel Core i7 MacBook Pro, running Java 1.6 with 3GB of RAM allocated to the virtual machine. For computing justifications we used the justification finding library that is distributed with the Explanation Workbench plugin in Protégé 4.3. The library essentially implements a black-box justification finding algorithm[6].

EXPERIMENT-A - Methods: We computed justifications for the inconsistent ICTM Content Model Ontology plus Data. Given that inconsistent ontologies can typically have large numbers of justifications, we imposed a 30-minute timeout on the justification finding run. We sorted the justifications that had been computed, by eye, into categories according to similarity patterns – justifications that appear in the same category are structurally similar to each other, meaning that they are either isomorphic to each other[7] or that they only differ by a subset of axioms that are used to entail the same intermediate entailment/step that is necessary in some other part of the justification. For example, an individual can be entailed to be an instance of a particular class by a direct assertion, or it can be entailed to be an instance of the class by being an instance of a subclass of the class.

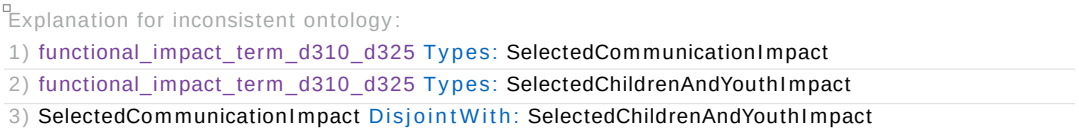


Figure 2(a) Justification-1. functional_impact_term_d310-d325 is an instance of SelectedCommunicationImpact (Axiom-1) and also an instance of SelectedChildrenAndYouthImpact (Axiom-2). However, these two classes are in fact disjoint with each other (Axiom 3), which makes the ontology inconsistent. Intuitively, an individual cannot be an instance of two disjoint classes.

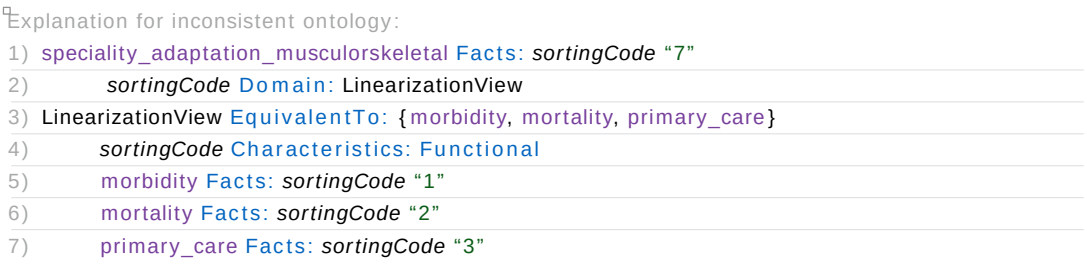


Figure 2(b) Justification-2. Recall that OWL does not have the Unique Name Assumption. That is, two individuals that have *different names* could denote the *same* object. In this justification, speciality_adaptation_muskuloskeletel is an instance of LinearizationView because it has a *sortingCode* (Axiom-1), and anything that has a *sortingCode* must be an instance of LinearizationView (Axiom-2). Furthermore, anything that is an instance of LinearizationView must be equal to one of morbidity, mortality, or primary_care (Axiom-3). Therefore, speciality_adaptation_muskuloskeletel must be equal to one of morbidity, mortality, or primary_care. However, speciality_adaptation_muskuloskeletel has a *sortingCode* value of "7". This means that it cannot be the same as any of these three individuals as any given individual can have at most one value for *sortingCode* (Axiom-4), and they each have a different value for *sortingCode* (Axioms 5-7). Intuitively, there are three possible values for LinearizationView, but other assertions contradict this and require a fourth distinct value (speciality_adaptation_muskuloskeletel). This makes the ontology inconsistent.

EXPERIMENT-A - Results: 621 justifications for the inconsistent ontology were computed within the 30-minute timeout (300 were computed within the first 5 minutes, with most being computed soon after, followed by a trickle of justifications in the remaining time). Browsing the 621 justifications (in the Protégé 4.3 Explanation Workbench) revealed that there were three main categories of justifications. The first category contains the single justification shown in Figure 2(a), the second category contains 531 justifications that are similar to the justification shown in Figure 2(b), and the final category contains 89 justifications that are similar to the justification shown in Figure 2(c).

EXPERIMENT-A – Analysis: A large number of justifications, 621, for the inconsistency were computed. Despite the large number, the justifications could be boiled down to *three distinct patterns*, indicating *three main kinds of*

[**] HermiT is an OWL reasoner that is bundled with Protégé. It is also available at http://hermit-reasoner.com. Note that we also used FaCT++ (https://code.google.com/p/factplusplus/) to double check that the inconsistency was a real inconsistency and not just due to a reasoner bug.

errors. It is worth noting at this point that some tool support for automating, or semi-automating, this process would have been extremely useful. However, after browsing the set of justifications for several minutes it became obvious that many of them followed common patterns. The main reasons for the large number of justifications in relation to the kinds of errors that can be observed are that (a) the Content Model is rather rich, meaning that there are many ways of entailing the same thing; and (b) the data (or entered content) is quite regular, meaning that the same patterns crop up over and over again, leading to many similar but distinct justifications. From Figure 2, it is noticeable that Justification-2 and Justification-3 are non-trivial and the ontology is inconsistent due to the interplay of many different kinds of axioms. The justifications *contain TBox axioms*, that are present in the Content Model Ontology, *and also ABox axioms*, that have been created via the iCAT tool (as data). Note that, although Justification-1 initially appears trivial, in the sense that it is straightforward to understand, it is highly unlikely that this could have been spotted by eyeballing the ontology. Finally, although it is possible to guess a repair for the inconsistent ontology from these justifications, in reality, some domain knowledge is required to understand where the problem really lies and to understand which axioms should be altered or removed.

Explanation for inconsistent ontology:

1) *factorBasedDiagnosis* Characteristics: Functional
2) influenza_A_H1N1 Facts: *exteriorOrInterior* exterior
3) *exteriorOrInterior* SubPropertyOf: *factorBasedDiagnosis*
4) influenza_A_H1N1 Facts: *yinOrYang* yang
5) diabetes_mellitus Facts: *yinOrYang* yang
6) *yinOrYang* SubPropertyOf: *factorBasedDiagnosis*
7) diabetes_mellitus Facts: *coldOrHeat* cold
8) *coldOrHeat* SubPropertyOf: *factorBasedDiagnosis*
9) *coldOrHeat* Range: ColdOrHeat
10) exterior Types: ExteriorOrInterior
11) ColdOrHeat DisjointWith: ExteriorOrInterior

Figure 2(c) Justification-3. influenza_A_H1N1 has a *factorBasedDiagnosis* value of exterior (Axioms 2 and 3), and a *factorBasedDiagnosis* value of yang (Axioms 4 and 6). Since any given individual can have at most one value for *factorBasedDiagnosis* (Axiom-1[††]). This means that exterior and yang is entailed to be the same individual. diabetes_mellitus has a *factorBasedDiagnosis* value of yang (Axioms 5 and 6) and a *factorBasedDiagnosis* value of cold (Axioms 7 and 8). Again, since any given individual must have at most one value for *factorBasedDiagnosis* (Axiom-1). This means that yang and cold must be the same individual. Since exterior is the same as yang and yang is the same as cold this means that exterior is the same as cold. Since cold must be an instance of ColdOrHeat (Axioms 7 and 9), exterior must also be an instance of ColdOrHeat. Since exterior must also be an instance of ExteriorOrInterior (Axiom-10), and ColdOrHeat is disjoint with ExteriorOrInterior (Axiom-11) this makes the ontology inconsistent.

Given the justifications computed in **EXPERIMENT-A**, three representative justifications, one from each category, were shown to the three members of the expert panel as part of a semi-structured interview in **EXPERIMENT-B**:

EXPERIMENT-B – Materials: A justification from each category was selected. The actual justifications are those shown in Figures 2(a) - (c). The set of guiding questions for the semi-structured interview is shown in Figure 3.

EXPERIMENT-B – Methods: A semi-structured interview based on the set of guiding questions presented in Figure 3 was conducted (by Horridge) with each participant on the expert panel. Each question in Figure 3 was asked with follow up questions where necessary to ensure the correctness in understanding of answers.

EXPERIMENT-B – Results: Three interviews lasted approximately 50 minutes, 40 minutes and 75 minutes. All participants described the ontology development process described earlier in this paper. None of the participants had used a reasoner during the evolution of the content ontology or for checking dumps of the Content Model Ontology plus data. *When asked to identify what would be a problem for the ontology, two out of the three suggested that logical errors, including inconsistency, would be problematic.* All of them suggested that incorrect classification of diseases etcetera would be problematic and that the long term goal is to eventually use reasoning to

[††] If a property is declared "Functional" this means that, for a given individual, it has at most one value. Multiple declared values will actually be entailed to be the same object. For example, in Figure 2(c), diabetes_mellitus essentially has two values, yang and cold, specified (via sub-properties) for the property factorBasedDiagnosis. Since factorBasedDiagnosis is functional, this means that yang and cold are entailed to be the same individual (same value). Note that OWL does not have the Unique Name Assumption hence, different names can refer to the same object.

assist the user in placing new terms into the class hierarchy. None of the participants had prior knowledge that the union of the Content Model Ontology plus the entered data was inconsistent. When presented with each justification, all of the participants were keen to understand what the problem was and all of them identified similar reasons for the problem. *All of the justifications indicated meaningful domain specific errors to the participants.*

For Justification-1 (Figure 2(a)), the participants identified either a possible problem with the modelling choices in the Content Model Ontology – in that the disjoint classes axiom should not be in the ontology – or some of them thought this might be due to a bug in the Form Configuration in iCAT. For Justification-2 (Figure 2(b)), all of the participants immediately identified the problem to be an issue with the Content Model Ontology. The EquivalentClasses axiom (Axiom-6) was introduced early on in the Content Model Ontology and was not updated as the Content Model Ontology evolved to represent more kinds of "linearization types" (instances of LinearizationView). Two participants said that the axiom should be updated and one said that it should either be updated or removed. Finally, for Justification-3 (Figure 2(c)), all participants spotted a problem with the Content Model Ontology, namely the functional property axiom (Axiom-1), which specifies that *factorBasedDiagnosis* is functional (single valued), should be deleted from the ontology. *All of the participants thought that the justifications were useful and the justifications identified significant errors in the Content Model Ontology*. For each justification all of the participants were able to suggest an initial possible repair to the Content Model Ontology or Forms.

General questions:
What was the methodology used for the development of the content model?
What was the measure of success for the content model ontology?
Did you use a reasoner (either during ontology construction or at runtime)?
What would be a problem for the content model ontology?
Would inconsistency or incoherence be problems for the content model ontology?
The ontology is inconsistent. Were you aware of this?

For each justification:
What do you think of it? What is your initial reaction?
Given the justification, can you explain what the nature of the problem is?
What kind of problem does it indicate?
How would you alter the ontology or data to fix the problem?
Would you change your modelling patterns as a result of this?

Finally:
Was this helpful as a QA exercise?
Would some kind of tool to assist help here?

Figure 3 – The guiding questions for the semi-structured interview.

EXPERIMENT-B – Analysis: Although one of the main goals of translating the Content Model to OWL and collecting the entered data as an OWL ontology was to have a more formal representation that could be automatically checked, this had not been carried out to date and was scheduled as part of future work in the project. None of the participants thought the Content Model Ontology would actually be inconsistent, and indeed, when taken in isolation, the Content Ontology Model is consistent. While some of the participants were initially surprised at this, when presented with the justifications they were able to quickly grasp what the problem was. In each case they were able to either immediately identify a way of fixing the problem themselves or they were able to identify a clear debugging/investigation strategy to be explored.

Two main categories of possible problems were identified: (1) *Problems with the content model* – the content model was deemed to be incorrect or buggy either because (a) it contained modelling slips, which had not been caught. In some cases the intended semantics were not the encoded semantics (for example the misuse of functional properties); or (b) because parts of it (in this case single axioms) were outdated with respect to an iteration in the evolution of the ontology. As stated previously, the Content Model Ontology evolves quite frequently and not all parts were updated to account for evolutions elsewhere in the ontology. (2) *Problems with the form configurations*. Ultimately, the forms accepted data that is inconsistent with respect to the Content Model Ontology (regardless or whether or not the Content Model Ontology is correct). In some cases for example, Justification-1, it was not immediately clear without further investigation whether the inconsistency was caused by a problem in the Content Model Ontology or by a problem in the Form Configuration in iCAT that allowed multiple choices for property values to be specified when only one choice should have been possible.

Finally, it was noticeable that all participants thought that having access to justifications was very useful (and interesting in itself – one of the participants would have liked to have gone through more justifications). They all

thought that this approach would prove quite valuable as a Quality Assurance procedure. All of the participants thought that some kind of tooling for use by Content Model Engineers and Form Configuration Engineers (as opposed to the subject matter experts who fill out details in iCAT) would be particularly useful.

Using Justifications to Repair the Content Model Ontology

For an inconsistency that can be traced back to problems in the Content Model Ontology, justifications for the inconsistency can be used to derive a repair plan for the ontology. A classic justification based repair requires all justifications for the problem in question: Given all justifications the simplest repair involves choosing one axiom from each justification and removing that axiom from the source ontology. Note that, this does not mean that for *n* justifications *n* axioms will need to be removed from the source ontology since justifications may overlap (share axioms) with each other. However, as mentioned previously, in **EXPERIMENT-A** not all justifications for the Content Model being inconsistent were computed. In this case it is necessary to perform an incremental repair – i.e. make changes to the ontology based on some axioms in *some* set justifications, next if the ontology is still inconsistent compute further justifications (which will not contain the justifications that were used for the repair), and repeat.

Given the descriptions of the problems provided by the expert panel we carried out a trial repair as follows. First, it was clear that the axiom LinerariztaionView EquivalentTo {morbidity, mortality, primary_care} was incorrect. We therefore removed this axiom from the Content Model Ontology (as suggested by one member of the expert panel) and recomputed the set of justifications. In this second round, 234 justifications were computed within a 30-minute timeout (with the majority being computed within 5 minutes). Most of these justifications were of the same form as the justification shown in Figure 2(c) and 233 of them contained the functional property axiom that makes *factorBasedDiagnosis* functional (single valued), which we chose to remove. We then recomputed the justifications again. This time, only 1 justification was computed – the justification shown in Figure 2(a), and we chose to remove the DisjointClasses axiom, which fixed the ontology so that it became consistent.

Discussion

The importance of detecting errors in ICD-11 and ICTM: The errors found in this case study were caused by errors with the ICTM Content Model Ontology, which the interview participants thought were genuine and significant. While ICTM and ICD have been under construction in iCAT for a period of time, it was a recent evolution of the Content Model Ontology where the inconsistency surfaced. Although the particular errors did not cause a specific problem with data entry via the iCAT forms at the time, it is arguable that maintaining a consistent Content Model Ontology is vital for the purposes of future proofing development and for preventing errors from creeping into Forms Configurations. Moreover, the success of the reusability of ICD-11 by other (OWL aware) software systems would be significantly impacted if the combination of the Content Model Ontology and data were inconsistent. Since one of the primary goals for ICD-11 is to achieve greater interoperability between software systems that use ICD-11, we consider a consistent Content Model Ontology and data to be of fundamental importance.

The importance of OWL reasoning and justification finding services for debugging ICTM: All of this work relies heavily on OWL reasoning and some kind of justification finding service. This was picked up by one of the interview participants who pointed out that it is highly unlikely, to impossible, that the errors would have been detected without using these technologies. Although they had not thought about using this kind of technology (particularly justification finding) for debugging and Quality Assuring ICD before, all of the participants thought that it was very useful for being able to identify and resolve these kinds of errors and saw genuine value in the technology going forward.

Using Justifications for Repair: The previous section describes a possible repair based on the computed justifications (i.e. which axioms to remove or edit in order to achieve consistency). While the decision of which axioms to alter or remove required the input from the expert panel – that is, people who are familiar with the domain, the use of justifications ultimately guided us in pinpointing the causes of the inconsistent Content Model and data, and also provided some guidance as to what a repair might be. Despite the very large number of justifications that were initially computed, it was not necessary to analyze them one by one in order to formulate a repair plan. Furthermore, it was not necessary to compute *all* justifications for the repair to be successful – an incremental repair was entirely satisfactory. It is worth noting that for this simple repair, we chose to modify the Content Model Ontology rather than the Forms Configuration – happily, this only required a few axioms to be touched. However, any ABox (instance assertions) axioms that appear in a justification could indicate problems with the Forms Configuration and provide a starting point of how to alter the Forms Configuration so that further data entry does not make things inconsistent.

Checking Form Configurations in advance: In our case study of ICTM, the Content Model Ontology by itself was consistent. It was only the combination of the Content Model Ontology and form data that was inconsistent (although the problems mainly originated from the Content Model Ontology). We only detected this inconsistency after "live data" had been entered into iCAT. Ideally, since forms should enforce consistency, it would be nice to have an *automated* testing tool that could detect any potential form configuration issues before forms "go live". While this could be tested by entering random data into the forms on some kind of staging platform and then inspecting the result, a smarter and more automatable strategy would be to generate an OWL ABox (instance assertion) representation of the forms and form data. These ABox statements could be generated automatically by examining the form structure, which could be obtained by parsing the form markup language. Since it is possible to map automatically generated ABox statements back to form widgets, any justifications for inconsistencies which contain these ABox statements could be directly used to highlight the parts of the form that cause the problems.

Taking the existing tools forward: The tools that we used in this work are essentially off-the-shelf OWL reasoning tools, which have been available in editors such as Protégé-4 for several years. While we foresee a situation where this tool chain will be used by expert administrators to debug the Content Model Ontology, it would be nice to have a single testing tool that integrates the tool chain into a single testing utility. Such a utility could provide a summary of the hundreds of justifications that get computed by using some kind of lemmatization strategy[8], or by using different notions of justification isomorphism[9]. Finally, a more advanced tool would have the ability to check forms prior to deployment and identify parts of forms that would cause a disparity (manifested as an inconsistency) between the entered data and the content model ontology. Such a tool could even suggest a variety of repair plans that the user might want to consider.

Generalizability and applicability to other medical ontologies: While we used ICD-11 Chapter-24 (traditional medicine) as a case study, it is likely that the techniques presented here should generalize to the complete development of ICD-11. In principle, the techniques used in this paper could also be generalized to the form-based editing of other medical ontologies where there is an underlying template (Content Model) ontology. While the content model is specific to a particular chapter of ICD-11, WebProtégé (on which iCAT is based) is a domain independent ontology editing environment. In terms of scaling the approach to much larger ontologies, such as the whole of ICD-11 (or other ontologies such as SNOMED-CT or the NCI thesaurus) there are three things to bear in mind: (1) In practice, the justification algorithm runtime performance does not directly depend on the size of the ontology[6] – It actually depends upon the number of justifications for a given entailment. While, in the theoretical worst case, the number of justifications is exponential in the size of the ontology, in practice, this is not the case and for most ontologies it is practical to compute all justifications for a given entailment; (2) Even if the *total* number of justifications is large, it is not necessary to compute *all* of them in order to carry out an interactive repair (simply seeing a handful of justifications was very informative in this case); and, (3) Even if the number of justifications computed is large, as was the case here, regularities, and the justificatory structure can be used to identify key patterns and problems and can be used to initiate a repair.

Related work: This project is not the first to use automated reasoning to assist in the construction of large medical ontologies. The ability to use reasoning as a kind of "complier" for medical ontologies was harnessed many years ago in the GALEN project. However, rather than being used as a debugging aid, the primary use of reasoning in GALEN was to compute an implied hierarchy and drive the end application Pen&Pad[10]. In work by Suntisrivaraporn[11] justification finding was used to explain subsumptions in the SNOMED-CT class hierarchy. The explanation tool in Suntisrivaraporn's work would theoretically be invoked by end users browsing SNOMED-CT who want to understand the reason for a particular subsumption (in a consistent version of SNOMED-CT). In contrast, we make use of the consistency checking during development to *directly flag errors*, indicated by inconsistencies, in the ontology plus data. We then compute the source of these errors using justification finding and use the computed justifications to point to whether the errors occur in the Content Model Ontology or the entered data. In recent years ontology quality assurance has become a hot topic, particularly in the area of bio-medical ontologies. Indeed, the Journal of Biomedical Informatics has devoted a special issue[12] to the problem. Rogers[13] presents a good overview of the quality assurance of medical ontologies.

Conclusion

Representing the ICTM (and ICD) Content Models in OWL was stated to have several benefits, one of which was that it would make it possible to use automated reasoning and error checkers to perform consistency checking and spot errors in the classification. In this paper we have successfully used automated OWL reasoning to detect inconsistencies in the union of the Content Model Ontology and the data entered into the ICTM authoring tool

iCAT-TM. We used an off-the-shelf reasoner to check consistency and used an off-the-shelf justification finding library to generate explanations for these inconsistencies. The justifications helped pinpoint and fix the causes of problems, which were identified as genuine problems by Content Model and Forms Configuration engineers. Whether the root of the problems lies in errors in the Content Model Ontology, or the Forms Configuration, the computed justifications make it possible to identify either a part of the Content Model Ontology that needed fixing, or narrow down the possible causes to a problem with interplay between the Content Model Ontology and the Forms Configuration. Without the use of automated reasoning, the inconsistency of Content Model Ontology plus data would not have been detected. Without the use of an explanation generation library to generate justifications, it would have been impossible to track down the reasons for the inconsistency.

All in all the use of OWL, automated reasoning, and non-standard reasoning services such as justification finding helped detect and debug errors in ICTM (and ultimately ICD-11). It seems promising that these tools could form part of a tool-chain that could be used to check Content Model Ontology and Form revisions before they are deployed in the runtime system. It also seems possible that such a tool chain could be used to regularly check for conflicts between the Content Model Ontology and data entered into the iCAT tool both as the Content Model Ontology evolve and the collected data grows.

References

1. World Health Organization, Brämer GR. International Statistical Classification of Diseases and Related Health Problems. 10th Revision. *Occup Health (Auckl)*. 1992;41:1–201.

2. Tudorache T, Nyulas C, Noy NF, Musen MA. WebProtégé: A Collaborative Ontology Editor and Knowledge Acquisition Tool for the Web. *Semant Web*. 2013;4:89–99. doi:10.3233/SW-2012-0057.

3. Motik B, Patel-Schneider PF, Parsia B. *OWL 2 Web Ontology Language Structural Specification and Functional Style Syntax*.; 2009. Available at: http://www.w3.org/TR/owl2-syntax/.

4. Tudorache T, Nyulas C, Noy NF, Musen MA. Using Semantic Web in ICD-11: Three Years Down the Road. In: *International Semantic Web Conference (2)*.Vol 8219. Lecture Notes in Computer Science. Springer; 2013:195–211.

5. Spackman KA, Campbell KE. SNOMED RT: A Reference Terminology for Health Care. In: Masys DR, ed. *Journal of the American Medical Informatics Association*. Bethesda, Maryland, USA: Hanley and Belfus Inc.; 1997:640–644.

6. Horridge M. Justification Based Explanation in Ontologies. 2011.

7. Horridge M, Bail S, Parsia B, Sattler U. The Cognitive Complexity of OWL Justifications. In: *International Semantic Web Conference*. Lecture Notes In Computer Science. Springer; 2011.

8. Horridge M, Parsia B, Sattler U. Justification Oriented Proofs in OWL. *Semant Web – ISWC 2010, Lect Notes Comput Sci*. 2010;6496:354–369.

9. Bail S, Parsia B, Sattler U. Diversity of Reason: Equivalence Relations over Description Logic Explanations. In: Kazakov Y, Lembo D, Wolter F, eds. *Description Logics*.Vol 846. CEUR Workshop Proceedings. CEUR-WS.org; 2012.

10. Rector AL, Horan B, Fitter M, et al. User Centered Development of a General Practice Medical Workstation: The PEN & PAD Experience. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '92. New York, NY, USA: ACM; 1992:447–453.

11. Baader F, Suntisrivaraporn B. Debugging SNOMED CT Using Axiom Pinpointing in the Description Logic EL+. In: *Proceedings of the 3rd Knowledge Representation in Medicine Conference (KR-MED'08): Representing and Sharing Knowledge Using SNOMED*.; 2008.

12. Geller J, Perl Y, Halper M, Cornet R. Guest Editorial: Special Issue on Auditing of Terminologies. *J Biomed Informatics*. 2009;42(3):407–411. doi:10.1016/j.jbi.2009.04.006.

13. Rogers JE. Quality assurance of medical ontologies. *Methods Inf Med*. 2006;45:267–274. doi:06030267.

Coordination of Care for Complex Pediatric Patients: Perspectives from Providers and Parents

Jan Horsky, PhD[1,4], Stephen J. Morgan, MD, FAAP[1,3,4], Harley Z. Ramelson, MD, MPH[1,2,4]

[1]Brigham & Women's Hospital, Boston, MA; [2]Partners HealthCare, Boston, MA; [3]Massachusetts General Hospital, Boston, MA; [4]Harvard Medical School, Boston, MA

Abstract

Coordinators help patients requiring complex chronic care manage frequent ambulatory visits and services received at home or from community-based agencies. EHRs directly support only a few of the required tasks as they do not allow access to all parties involved in care. Our goal was to examine how technology was used to coordinate efforts and to describe common barriers and facilitators. Insights may inform the design of tools that would effectively support identified goals. We conducted five hours of interviews with sixteen parents and six clinicians and characterized emergent themes from transcripts. Situational awareness, care and visit planning, document aggregation, abstraction and interpretation were tasks essential to coordination yet generally poorly supported by EHRs. Providers communicated primarily by email, telephone and by exchanging paper and scanned documents. A preliminary model of coordination that could be used in the planning and testing stages of a User Centered Design process is described.

Introduction

Inadequate coordination has been identified as a significant impediment to effective care,[1, 2] contributing to increases in costs and negatively affecting quality.[3-5] As specialist care is increasing and accounts for more than half of all ambulatory visits, teams consisting of several providers located at different institutions are common.[6] Patients with multiple chronic conditions may visit up to 16 physicians in a single year.[7] Each provider often needs to review results, findings, relevant patient history and other information gathered by the others in order to make care decisions. Exchange of information across technological and institutional boundaries, however, has been a time-consuming, complicated process fraught with care delays, duplication of effort and a source of frustration.

Most currently available electronic records systems (EHR) do not support coordination and collaboration among caregivers adequately.[8] A survey of clinicians using vendor EHRs found that the technology provided only limited support for many common coordination tasks.[9] For example, an accurate, updated list of care team members is not yet widely available.[10] The low prevalence of this resource in current systems is reflected in the modest requirement by the Office of National Coordinator (ONC) to maintain such lists for at least 10% of patients in order to receive Meaningful Use Stage 2 certification.[11] The referral process for specialty care is often described as a frustrating series of disjointed phone calls, faxes and mail that may result in delay, lapses and duplication of effort.[12]

Care for patients with chronic conditions, especially congenital and childhood-onset diseases, is highly complex, involves non-medical professionals and requires significant involvement of patients and their family members. Coordination, characterized as a flow of information between providers to ensure that they all act toward a common goal, encompasses health services such as ambulatory, hospital, post-acute and home care, and social support services from state, community and private organizations.[13-16] Vendors have so far not adequately addressed the limitations of EHRs in supporting coordination beyond institutional boundaries.[17] Professionals providing patient care and services therefore need to rely on other technologies to communicate[18] and often turn to web-based applications to improve the process.

We conducted a qualitative study to gather and analyze insights from clinicians, coordinators and parents of children with multiple chronic conditions to formulate a preliminary model of care coordination intended to inform the design of electronic support tools. It models information needs and flow between participants, describes common technical and organizational barriers and outlines the complexities of information exchange that advanced software tools will need to support. Studies of similar scale and purpose are usually performed in the initial phase of a User-Centered Design (UCD) process[19] as necessary precursors to functional and design specifications that are sensitive to appropriate fit to task and clinical context. UCD is established in software development as a robust, reliable process across many industries and has recently been required for HIT design under the 2014 Meaningful Use Act.[20, 21]

Methods

We interviewed twenty-two individuals at two institutions affiliated with a large urban healthcare delivery network in five sessions lasting an hour each. Two primary care physicians and one nurse practitioner were interviewed individually and a primary care physician, a social worker and a coordinator as a group. Sixteen parents of children with multiple chronic problems formed a focus group during their scheduled monthly meeting. The interviewer asked open-ended questions intended to elicit descriptive narratives about their routine tasks and activities in coordinating care for their children. Participants were encouraged to describe facilitators and barriers to effective care management, their most frequent modes of communication and to give examples of the types of documents they gather, receive, store or exchange with other parties, including patients and their families. The interview structure was gradually adapted over time to focus on salient themes emerging from preceding sessions and to follow up on specific topics. Interviewees were selected as a sample of convenience but attention was given to including providers from both hospital and community settings in order cover a broader set of personal perspectives.

All sessions were recorded, transcribed and analyzed using NVivo 10 software.[22] Codes were developed and refined in three iterations by one investigator in which statements representing discrete meaningful concepts were either aggregated into broader categories based on similarity or differentiated into more descriptive units (axial and selective coding).[23] The final structure was reviewed and refined in two consecutive sessions in collaboration with a second investigator.

We analyzed all transcripts by using a constant comparative method[24] that involves identifying ideas in discrete statements (units of observation) and aggregating them into sets. This method produces categories derived from the participants' expertise and language and those the researcher identifies as significant to the focus of inquiry (e.g., coordination, communication). The goal is to develop theoretical insights into processes that characterize the flow of information and construction of goals in a socio-technical system (STS).[25] The sociotechnical framework was used to interpret and analyze the findings as components of goal-oriented activity that encompasses humans, technology and artifacts.[26]

Interview analysis

We identified and coded statements about roles, communication and information technology and characterized actors, technological agents and artifacts. For example, we analyzed accounts of compiling, abstracting, and interpreting information from visit notes and other sources routinely done by physicians and administrative staff in order to update patients on the course of treatment and to plan future tests and visits. Analysis of the parent focus group followed themes that were unique to their experience and those that provided their perspective on processes and issues identified in interviews with providers.

Results

Descriptive and analytical results from the interviews are described below. We identified main participants in communication and information exchange (parents, families, care and assistance providers), technology (EHR, email, mail, phone, fax) and goals and activities characterizing coordination (aggregation, abstraction, interpretation, visit planning, situational awareness, care planning, medications and problems). We also describe how these findings would inform the design of software tools for coordination.

Patients and families

Families with children who had three or more chronic problems or disabilities requiring frequent treatment and continuous care were assisted by coordinators whose task was to manage most communications with the care team, interpret and clarify medical findings provided by specialists and advise parents on the progress of care and next steps. For example, one coordinator described the treatment complexity of most patients as "having four to five subspecialists who are seen regularly every few weeks to every few months; almost all having physical, occupational or speech therapy." Parents were regarded as the main source of information about previous care and as agents who maintained contacts to community-based assistance providers and mediated the forwarding and transferring documents between offices and agencies without direct formal affiliation. However, primary care providers and coordinators actively maintained communications and activities for those not able to carry out the tasks themselves.

Role	Description
Primary Care Physician	Located in large care centers or in patient's home community
Specialist Physician	Affiliated with hospital network or in independent clinics and offices
Nurse	Clinic, community, school, home care workers, visiting nurses
Therapists	Physical, occupational, speech, and specialized forms of therapy
Administrator	Supports PCPs and other clinicians in care management.
Coordinator	Services and care coordination at clinics, state and private agencies
Social Worker	Services and management at hospital centers and state agencies

Table 1 Roles of collaborating professionals

Parents described similar levels of care complexity as did clinicians and coordinators although individual experiences varied. Several parents reported receiving care regularly from about ten providers, "six of them out of the system," and worked extensively to maintain communication and timely updates. Their frequency of visits was typically "between six times a month to about five times every two months, with physical therapy sessions occurring several times per week." Collectively, they expressed the feeling of "being intermediaries, especially for episodic care like ED visits," or "being the clearinghouse for information," closely echoing the perceptions of clinicians.

Care and assistance providers

Medical care providers were either co-located in one physical setting (e.g., a hospital or a clinical center) or had private offices in other locations, sometimes in a patient's home community. A list of professional roles of care givers participating in coordination activities, as described in the interviews, is in **Table 1**. Those affiliated with the same care network shared an EHR system although some records such as emergency department visits, discharge summaries or specialty visit notes may have been stored on non-integrated systems and therefore available only to some clinicians. Primary care physicians (PCP) provided routine care and either communicated with a coordinator or assumed the responsibilities of a medical home. Nurses provided care at clinics, patient homes (visiting services) or at schools, therapy centers or other facilities and organizations. Providers not directly affiliated with the network generally maintained their own notes and gave occasional updates to a PCP. There were several coordinators of clinical care and of social services who worked at state agencies (e.g., Departments of Social Work, Early Intervention, Families and Children, etc.) or at community support centers, schools and other institutions. Their contact with PCPs and hospital-based care coordinators and social workers was primarily through email and telephone. Several parents noted that their PCP was not central to coordination and mostly "was there for your typical colds and well visits." However, this arrangement required them to do more information forwarding and updating as "all specialists refer you back to your PCP."

Communication and information technology

Coordinators and providers relied primarily on email, phone and mail to reach others across institutional boundaries. These forms of communication were often inadequate, laborious and time consuming, requiring extra and duplicative effort to manually update electronic records and private notes. For example, information circulated by mail was not always received or reviewed in time to effectively support face-to-face conversations and maintain situational awareness. As one coordinator explained, "if I have a question, I'll call the PCP and he says that we didn't receive a letter from you although I sent it the day after I saw the patient. They may be too busy to follow up with us." Specialists who did not have access to a shared EHR had difficulties forwarding information such as visit notes and records of prior care. They often scanned documents to a PDF and emailed them or used a fax or the mail.

Delays and other complications resulting from slow communication sometimes affected care decisions. Several parents whose children received immunizations, sick and other routine care from PCPs in their community and more extensive chronic and specialty care at hospital network centers decided to transfer all their care to the centers even if it entailed more travel and time. However, coordination and communication between the providers became less problematic as a result. A coordinator noted that "this may not be the case if we were able to communicate better."

All interviewed professionals expressed a strong preference for a common electronic platform that would accommodate automated updates of such basic information as contact records and also clinical and other personal data for the entire team, including patients and their families. According to one social worker, "just knowing who is

Technology	Limitations
EHR System	Limited to affiliated providers only. Duplicate sets of data need to be maintained separately by other parties.
Email	Bridges institutional boundaries but is often a poor workflow fit. Group conversations are difficult to maintain and security and privacy is a concern.
Mail	Substitutes for lack of interoperability. Extra work is needed to abstract data for coded entries into an EHR.
Telephone	Poor support for group decision making. Synchronous conversations do not often fit into workflows.
Fax, PDF	Documents are mostly not searchable and need to be abstracted for EHR use.

Table 2 Communication technology

on the team is a huge challenge for parents and also for the vast army of involved people, the folks that help the children at home, physical therapists, early intervention people, the school nurse, probably the teachers and others." This frustrating situation was described by a PCP who noted that "reaching out to the community side, agencies responsible for approving equipment the patient needs, finding out what letters need to go with that and recording it is enormously laborious." A list of the most frequently used communication technology and its limits for documentation exchange, as described in interviews, is in **Table 2**.

Direct access to selected information by patients and family members was considered essential for reducing their dependence on mediation by clinicians, coordinators and social workers. For minors, elderly patients and those without computer skills, "care planning still has to be done the old fashioned way, either in person or on the phone," according to one PCP. However, many others would benefit from direct access and could update or correct information and be more engaged in their own care.

Care coordination

Activities directly associated with care coordination that were discussed in the interviews were, a) *aggregation* of documents and information, b) *abstraction* of key events from charts for EHR entry, c) *interpretation* of medical findings and advice for patients, d) scheduling and tracking of referrals and *visits*, e) maintaining long-term *situational awareness*, f) development and updates of *care plans* and g) maintenance of medication, allergy and problem *lists*. The associated tasks and activities are summarized in **Table 3** and described in detail in the sections below**.**

Aggregation

Before the first visit of a new or a transfer patient, administrators, nurses and coordinators collected all available electronic, paper and scanned documents they were able to obtain from prior and current providers. Patients or their families were central to this process as they often provided names, contact information and key dates although missing information still had to be identified and added. Some PCPs were able to "off-load much of this information gathering" to staff but for very complex patients or those with cognitive impairments clinicians had to "figure out who in their life are key people who could help with information." However, this process was time consuming and incurred considerable delays as information arrived through the mail as paper documents or printouts of electronic notes from systems that were not interoperable. As one coordinator noted, "you may have seen the child a couple of times before the records arrive and that is incredibly frustrating for parents."

Abstraction

Physicians often had to abstract, compare and summarize from multiple records core health problems, patient history, prior treatment and procedures and to identify information that is missing or requires updates and more detailed description. This work had to be completed by clinicians (often the PCP) rather than support staff. As a coordinating physician noted, "it means that I have to hunt through hundreds of pages and find what's important although summaries may exist for some patients in a one or two page document that has all the key information in one spot." Direct phone conversation with identified referring physicians can sometimes clarify "what the key issues, hospitalizations and problems are" if the record is too extensive to review.

Patients receiving complex care had a large volume of records of which only a fraction – those generated by providers from the same care network – were electronic. Very frequently, documents were forwarded across settings in scanned PDF files that could be "hundreds of pages long" and from which "immunization record,

Activity	Description
Aggregation	Compile records generated by clinicians, professionals and family across all settings.
Abstraction	Identify and summarize key points from extensive records of prior care.
Interpretation	Collate and interpret visit notes for patients, clarify results, requirements and next steps in care.
Visit planning	Initiate and track appointments, test, consultations and referrals.
Situational Awareness	Maintain awareness of new events, care progress, upcoming and missed visits.
Care Plan	Individual course of interventions, monitored for situational awareness.
Medications, Problems	Reconciliation of lists at hospital and community locations.

Table 3 Activities associated with care coordination

operative, advanced directive and other things I will need to search for" had to be extracted and the information entered in a coded form into the EHR. PDF documents were mostly not searchable automatically. Paper charts, forms, documents and even printouts from systems that were not interoperable were routinely scanned and entered into EHRs as attachments to clinical note entries.

Abstraction and comparison of documents from multiple providers was also done to reduce redundant or conflicting recommendations. A physician noted that "I want to know what the social worker has done so that I'm not asking the same battery of questions and recommending the same services already put in place." As the notes were coming from many different sources "it took a long time to do our homework – there is no one place to read all of this."

Interpretation

Clinicians and coordinators compiled and interpreted for patients and their families important findings and instructions from numerous specialist and procedures that were not well understood during the visit, either for their complexity or the sheer volume of information. Coordinators typically talked to a patient after three or more medical visits have taken place (often on a single day), "collating who they saw and what that person wanted to happen next" and to "look at the big picture of who recommended what, and to synthesize information." This overview often generated further queries to the providers to confirm particular requirements or to get more detailed instructions. For example, a coordinator noted that "I can address any mismatches or ask questions – if the cardiologist said this patient needs to have a catheterization I can make sure that the patient understood it." Patients or parents often asked coordinators to follow up with questions they did not raise during visits for lack of opportunity or understanding. A coordinator noted that "it took a lot of emailing behind the scenes to get a sense. For example, if a kid was started on new epileptics, are they supposed to be seizure-free or are we hoping for a 50% decrease? What's the time period that you would expect this, what are the side effects and can he continue to go to school?"

Visit planning and referrals

Patients typically needed to schedule and keep regular appointments with four or more sub-specialists, therapists, social workers and other assistance providers. Coordinators helped patients navigate new environments such as large hospital and clinical centers so that they did not get "overwhelmed by the volume of information and moving parts" and called and emailed multiple providers to arrange visits so they to co-occurred on the same day in one location. In some cases, complex scheduling was arranged by a dedicated service center available at the hospital. When patients did not have access to a coordinator, their PCPs had fewer opportunities to call and schedule multiple visits, so often "the parents were pretty much on their own for making those appointments." Knowing which appointments needed to take place, what procedures or tests will need to be done and coordinating planned care was essential for avoiding redundant visits or missing some due to conflicts.

Tracking visits and their outcomes was difficult, required careful planning and active outreach to providers to find out whether an appointment did take place and what were the next follow-up steps. For example, a coordinator noted that "a large proportion of all referrals never actually get completed and without proactively keeping track I can't assume that they were done." The EHR was considered to be only marginally useful in displaying scheduled and missed appointments as many specialists did not share the same system. A physician described the task as "actively, manually tracking what are the specialties and clinics, who needs to follow up with them and when."

Reasons for missing appointments varied widely but coordinators felt that they could effectively intervene to prevent some of those if they were sufficiently informed of the situation. For example, a social worker noted that families needing transportation services from a state agency may "miss six appointments in a day if the required paperwork did not go through." A more effective, integrated communication system may allow coordinators to notice such gaps and contingencies and act in time to make sure that medical care can be provided or rescheduled.

Situational awareness

Awareness of activity and communication taking place in a socio-technical system is essential for understanding how events and actions of all agents – human and technological – affect current personal and common objectives and future plans. The design of most EHRs generally does not allow clinicians to maintain this level of activity awareness, especially across different institutions.[27]

A physician described updating others on the same care team about planned actions: "After a patient visit I usually talk to their various other providers to make sure we're all on the same page." There was no central place to "see the care team and when they want to see the patients next and what the active issues are." The physician had to recreate the sense of the current state of care from "looking at notes," a time-consuming task of aggregating and reading through electronic and paper records and emails. Coordination of tasks for very complex patients required in-person meetings and appointments with patients to "go through all of the specialty visits, answer questions, think ahead and troubleshoot any care deficits."

Workarounds were devised by coordinators to keep them informed about unscheduled events, underscoring the inadequate support for this task by their EHRs. For example, the IT department developed by request an email service to automatically generate an alert message when patients on a tracking list were admitted, discharged or seen at the emergency department or same-day surgical units. Events and visits taking place outside of the hospital network, however, could be tracked only if patients reported them. Another repurposing of email to remind about future events, as reported by a physician, was to set up messages to arrive on days of important follow-up appointments and to "call afterwards to make sure that they went." However, this strategy was not suitable for tracking a large number or frequent visits. The physician in such cases required patients to "come back and figure out if they've missed things – but I can't be calling." Sub-specialists would sometimes notify PCPs by email about missed care and then "somebody would reach out to the family."

Poor situation awareness and event tracking may contribute to adverse events resulting from missing or delayed care. A physician reported that "one of my patients didn't go to catheterization because they didn't understand the conversation with the cardiologist and I didn't know about it, then the patient got hospitalized." Tracking abnormal radiographs and tests was also seen as a recurring problem with many opportunities for omission.

The lack of reliable, integrated information sharing between medical and non-medical care providers confounded situation awareness for the entire team. Most communications took place through email, over the phone or by postal service, requiring secondary manual documentation and updates. Non-medical events with the potential to negatively affect care were difficult to discover in time to intervene. For example, a physician may learn about a change in a patient's social services or about serious incidents "only if the family informs me or if someone from DPH tries to contact me, but that hasn't really happened." In cases where children are "moved to a different home or foster parents" their local records may not be transferred with them. As one PCP noted, "when they come back in 2 or 3 years, I don't know what happened, did they get their vaccines, etc. It takes digging into records and talking to the social workers and getting all that information."

Coordinators or primary care physicians usually received information from social and community workers only "on demand" and in case of serious problems when "they will reach out to us to come up with a plan together." Several clinicians thought that "sometimes folks in the community feel very disconnected from teams at the hospital" and that better access to shared records would benefit all involved and allow them to make better decisions. Some questions from patients could be directly answered by the most appropriate person rather than being mediated by physicians. For example, families now approach their primary care providers for "community related things like they need to know how to get a wheelchair, ramp, accessible housing, transportation, a better day program and a host of other non-medical things. Finding out which case worker needs to know that a request for equipment was filed is very time consuming."

Communicating with patients is often done through secure email and then documented by "copying all the emails to the chart." Some physicians felt that it is currently difficult to extend or route these conversations to those who can provide the best answers such as "administrators, nurses, social workers and the other people on the team."

Parents noted that they often need to actively monitor their own care and recognize that certain events should take place which could be difficult for very complex treatment plans. An automated system of reminders about "appointments or blood draws that they have to do through email or something outside of their own scrap paper" was regarded as highly desirable.

Care plan

As the number of problems, providers, and sites increase, overall coordination of different provider-specific treatment plans, discipline-specific plans of care, and the production of a master care plan are needed.[28] Physicians providing primary and specialty care as well as therapists develop their own respective plans that need to be monitored for progress and completion of goals and adjusted if necessary. Core information from plans such as due dates, missed and achieved goals and new or unplanned events are important components of situational awareness. Coordinators needed to review and monitor this data periodically in order to take appropriate action. This function was approximated by reviewing notes to "figure out if they haven't seen some person, are due for something, or there are upcoming things involving patient care." The process had almost no support by information technology. A physician reported that "I can put one concrete piece of plan in a clinical message to my nurse and the nurse can call but if my plan is complex and evolves through a discussion then I need to do that myself." Email and phone communication, however, was ineffective for team discussions and real-time updates as "there are so many points where any little thing can slip up that will erode that communication and it can be disastrous sometimes."

Increased situational awareness may improve appointment scheduling, event prioritization and better responsiveness to emergent care needs. Such system, however, would need to bridge the existing electronic boundaries delineated by limited access to shared electronic information and replace the often poor workflow fit of email and phone communication with more appropriate forms. A parent noted that they felt as if long-term planning was not prioritized or well maintained by anyone. For example, "I don't feel that there is a management system for the long term things; critical care is well managed and followed up, but it's up to me to figure out what needs to be taken care of next and send emails so that it doesn't fall off the radar screen."

Medication and problem lists

Lists of currently active and prior medications and problems were usually maintained by the primary care physician on an EHR. Care providers in the community such as visiting and school nurses administered the medications and kept their own records. Although dosing changes and substitutions need to be approved by a physician, central updates and reconciliation of lists was usually done at intervals coinciding with visits that sometimes substantially lagged behind actual events. As reported, many reconciliations took up patient visit time that would be otherwise focused on care issues. For example, patients may have a medication list from visiting and community nurses and the primary care physician "hopes that when they come in, it matches mine, but most of the time it doesn't so the majority of the visit is spent reconciling medications." Reconciling over the phone may "take 20 minutes or so" and fax communication may never close the confirmation loop if not returned in time.

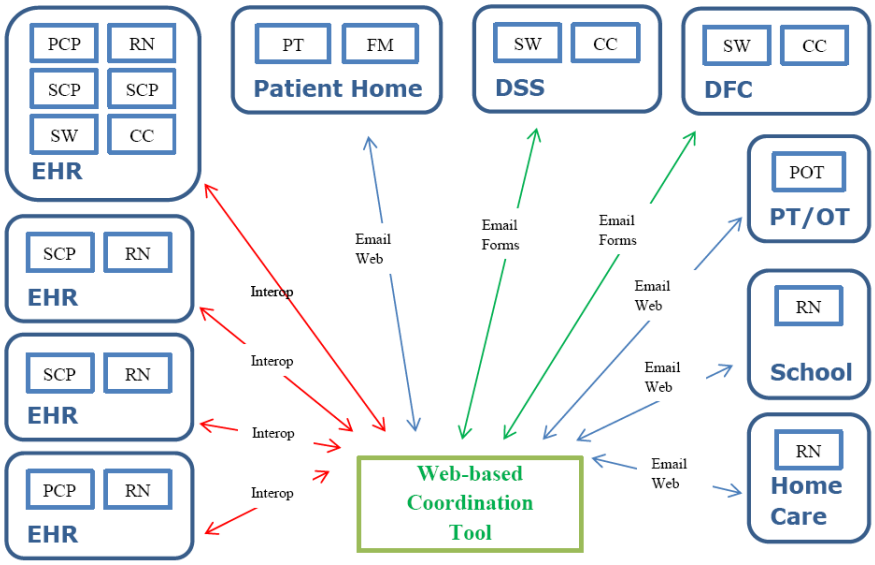
Parents acknowledged that from their experience, reconciliation was a laborious and often imprecise process. For example, one parent reported that "it's painful; they ask you to verify it and I see the same things that we told them the last time to take off the list. Something was prescribed two years ago by a specialist and then it just sits there on the list of your standard medications."

Differences in separately maintained lists could be substantial if a primary care physician sees a patient once every 2-3 months. Updates accumulated over time between home, school and hospital discharge lists may take several hours to complete, especially with patients who may be taking 40 or more medications. A more efficient way to update lists electronically and to propagate changes to all when they occur was a common request by all interviewed coordinators and physicians.

Care coordination model

A web-based tool would need to support, at minimum, the following tasks and activities: increased situation awareness by allowing plan, visit and event tracking, allowing medication and problem list updates via secure email

and embedded links, provide a two-way patient-facing portal for care instructions and requests and integrating updated information such as medication changes with the primary EHR after a reconciliation and confirmation process. We hypothesized that while it was unreasonable to expect that all participants would be willing to change their preferred workflows to navigate to a common website in order to communicate with others or update shared lists, they would likely respond to email prompts with one-click links to give and receive information updates that would be valuable for their own work.



Legend

PCP Primary Care Provider, **SPC** Specialty Care Provider, **RN** Registered Nurse (Clinic, Home, School), **CC** Care Coordinator (Clinic, Agencies), **SW** Social Worker (Clinic, Agencies), **POT** Physical/Occupational Therapist, **PT** Patient, **FM** Family member, **Iterop** Two-way interoperable connection, **Web** Access through web interface, **Forms** Electronic forms

Figure 1 Model of possible integration with existing technology

A model of possible integration of a coordination tool with EHRs at separate institutions and with direct access from patients is in **Figure 1**. EHRs at hospitals and clinics can implement basic interoperability of medication, problem and contact updates through email links, state agencies may exchange electronic forms, community and visiting nurses can update medications through email automation and patients can review care instructions and educational materials online. Participants may copy emails to others on a shared server and update their contact information. Electronic forms can support semi-automated entry of coded information through text processing and confirmation.

Discussion

The core themes about barriers to effective care coordination that emerged from the interviews showed that lack of integration and system interoperability within and primarily across institutions and professions adds considerable effort to the work of most clinicians and providers of services. Real-time situation awareness that is indispensable for the coordination of actions and reduction of duplicative effort is difficult to derive from existing fragmented and mixed use of electronic and paper documentation. Compiling, tracking and following up on events across the medical-social services boundary was the most challenging part of coordination as the EHR – virtually the only shared electronic resource – does not extend outside of a network of affiliated clinics and offices.

Clinicians and assistance services managers disproportionately relied on paper or scanned documents when information had to be sent to others on the care team. However, it seems that many transactions (e.g., requests for

standard social services or medical equipment approvals) used paper forms and collected fairly structured information that could be easily replicated in electronic form. We also noted that some tasks described as having no electronic support could in fact be accomplished with existing systems although some clinicians were not aware of the possibility.

Our results correspond with previous reports about inadequate EHR support for team-based chronic care management and that monitoring, plan feedback and assurance that plan components were completed are core requirements for effective coordination.[29] Our model, however, outlines the necessity to integrate non-medical professionals with the care team more closely. We also propose that improvement of EHR design alone can benefit only those clinicians who have access and that bridging communication gaps to care and service providers outside the network may require innovative approaches that integrate several technologies such as web and email-based forms and include interoperable connections between different EHR systems.

We formulated design advice and proposed a model primarily to support the work of care coordinators serving as liaisons between patients requiring complex care and the care team. Primary care physicians with responsibilities to maintain a medical home may also benefit from a similar system that is tightly integrated (e.g., as a module) with their EHR. They often take care of young adults who have transitioned from specialized pediatric care centers but may require as much coordination as before without adequate staff support.

Limitations

This study was conducted on a scale that may not be sufficiently large for generalization or validation of the coordination model as a comprehensive and predictive construct. It was intended, however, to give adequate understanding of the needs of patients, clinical and other professionals involved in the process of chronic care as a pre-requisite for the design of an advanced electronic coordination tool. Studies of similar size are commonly done in the first, planning stage of a user-centered design process and give informaticians and designers insights into unmet needs, communication and information flow complexities and requirements that the software will need to support and accommodate. Our methods and results should serve as valid reference and a comparison points to investigators and developers engaged in the design of similar electronic tools and looking for guidance.

Acknowledgment

This work was supported by a grant from Partners-Siemens Research Council. We thank the clinicians, coordinators and parents who shared their expertise and experiences with us and generously gave us their time to be interviewed.

References

1. Bodenheimer T. Coordinating care: a major (unreimbursed) task of primary care. Ann Intern Med. 2007 Nov 20;147(10):730-1. PubMed PMID: 18025448.
2. Mehrotra A, Forrest CB, Lin CY. Dropping the baton: specialty referrals in the United States. The Milbank quarterly. 2011 Mar;89(1):39-68. PubMed PMID: 21418312. Pubmed Central PMCID: 3160594.
3. Bodenheimer T. Coordinating care--a perilous journey through the health care system. The New England journal of medicine. 2008 Mar 6;358(10):1064-71. PubMed PMID: 18322289. Epub 2008/03/07. eng.
4. Berry LL, Rock BL, Smith Houskamp B, Brueggeman J, Tucker L. Care coordination for patients with complex health profiles in inpatient and outpatient settings. Mayo Clinic proceedings Mayo Clinic. 2013 Feb;88(2):184-94. PubMed PMID: 23290738. Epub 2013/01/08. eng.
5. Schoen C, Osborn R, How S, Doty M, Peugh J. In Chronic Condition: Experiences of Patients with Complex Health Care Needs, in Eight Countries. Health Affairs [Internet]. 2008:[w1-w16 pp.].
6. Forrest CB, Majeed A, Weiner JP, Carroll K, Bindman AB. Comparison of specialty referral rates in the United Kingdom and the United States: retrospective cohort analysis. Br Med J. 2002 Aug 17;325(7360):370-1. PubMed PMID: 12183310. Pubmed Central PMCID: 117891.
7. Pham HH, Schrag D, O'Malley AS, Wu B, Bach PB. Care patterns in Medicare and their implications for pay for performance. The New England journal of medicine. 2007 Mar 15;356(11):1130-9. PubMed PMID: 17360991. Epub 2007/03/16. eng.
8. O'Malley AS. Tapping the unmet potential of health information technology. The New England journal of medicine. 2011 Mar 24;364(12):1090-1. PubMed PMID: 21428764.

9. O'Malley AS, Grossman JM, Cohen GR, Kemper NM, Pham HH. Are electronic medical records helpful for care coordination? Experiences of physician practices. Journal of General Internal Medicine. 2010 Mar;25(3):177-85. PubMed PMID: 20033621. Pubmed Central PMCID: 2839331. Epub 2009/12/25. eng.
10. Vawdrey DK, Wilcox LG, Collins S, Feiner S, Mamykina O, Stein DM, et al. Awareness of the Care Team in Electronic Health Records. Appl Clin Inform. 2011;2(4):395-405. PubMed PMID: 22574103. Pubmed Central PMCID: 3345520. Epub 2011/01/01. Eng.
11. Office of the Secretary. Health Information Technology: Standards, Implementation Specifications, and Certification Criteria for Electronic Health Record Technology, 2014 Edition; Revisions to the Permanent Certification Program for Health Information Technology. In: Health And Human Services, editor. March 7 - Proposed rules ed. Washington, D.C.2012. p. 13832-85.
12. Metzger J, Zywiak W. Bridging the Care Gap: Using Web Technology for Patient Referrals. Oakland, CA: California HealthCare Foundation, 2008.
13. Craig C, Eby D, Whittington J. Care Coordination Model: Better Care at Lower Cost for People with Multiple Health and Social Needs. Cambridge, MA: Institute for Healthcare Improvement, 2011.
14. Antonelli RC, McAllister JW, Popp J. Making Care Coordination a Critical Component of the Pediatric Health System: A Multidisciplinary Framework. The Commonwealth Fund, 2009.
15. McAllister JW, Presler E, Cooley WC. Practice-based care coordination: A medical home essential. Pediatrics. 2007 Sep;120(3):e723-33. PubMed PMID: 17766512. Epub 2007/09/04. eng.
16. Rich E, Lipson D, Libersky J, Parchman M. Coordinating Care for Adults With Complex Care Needs in thePatient-Centered Medical Home: Challenges and Solutions. Rockville, MD: Agency for Healthcare Research and Quality, 2012 January 2012. Report No.: Contract No.: AHRQ Publication No. 12-0010-EF.
17. Rudin RS, Bates DW. Let the left hand know what the right is doing: a vision for care coordination and electronic health records. J Am Med Inform Assoc. 2013 Jun 19. PubMed PMID: 23785099.
18. Bates DW. Getting in step: electronic health records and their role in care coordination. J Gen Intern Med. 2010 Mar;25(3):174-6. PubMed PMID: 20127195. Pubmed Central PMCID: 2839327.
19. Schumacher RM, Lowry SZ. NIST Guide to the Processes Approach for Improving the Usability of Electronic Health Records. Washington, D.C.: National Institute of Standards and Technology, 2010 NISTIR 7741.
20. The Office of the National Coordinator for Health Information Technology. Health Information Technology Patient Safety Action & Surveillance Plan. Washington, DC: Department of Health and Human Services, 2013.
21. Institute of Medicine. Health IT and Patient Safety: Building Safer Systems for Better Care. Washington, D.C.: The National Academies Press; 2011. 197 p.
22. QSR International Pty Ltd. NVivo qualitative data analysis software, Version 10. 2010.
23. Bernard HR, Ryan GW. Analyzing qualitative data: Systematic approaches. Los Angeles Calif.: SAGE; 2010. xxi, 451 p. p.
24. Maykut PS, Morehouse R. Beginning qualitative research : A philosophic and practical guide. London ; Washington, D.C.: Falmer Press; 1994. xii, 194 p.
25. Berg M. Patient care information systems and health care work: A sociotechnical approach. International Journal of Medical Informatics. 1999 Aug;55(2):87-101. PubMed PMID: 10530825.
26. Clegg CW, Frese M. Integrating organizational and cognitive approaches towards computer-based systems. Behaviour and Information Technology. 1996 Jul-Aug;15(4):203-4. PubMed PMID: 1997-07773-001.
27. Lawler EK, Hedge A, Pavlovic-Veselinovic S. Cognitive ergonomics, socio-technical systems, and the impact of healthcare information technologies. International Journal of Industrial Ergonomics. 2011;41(4):336-44.
28. Dykes PC, Samal L, Donahue M, Greenberg JO, Hurley AC, Hasan O, et al. A patient-centered longitudinal care plan: vision versus reality. J Am Med Inform Assoc. 2014 Jul 4;In Press. PubMed PMID: 24996874.
29. Dorr DA, Jones SS, Wilcox A. A framework for information system usage in collaborative care. Journal of Biomedical Informatics. 2007;40(3):282-7.

Effect of Obesity and Clinical Factors on Pre-Incision Time: Study of Operating Room Workflow

Narges Hosseini[1,2], PhD, M. Susan Hallbeck[1,2], PhD, Christopher J. Jankowski[3], MD, Jeanne M. Huddleston[1,2,4], MD, Amrit Kanwar[1], Kalyan S. Pasupathy[1,2], PhD
[1] Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery
[2] Department of Health Care Policy & Research
[3] Department of Anesthesiology
[4] Department of Internal Medicine
Mayo Clinic, Rochester, MN

Abstract

As the obese population is increasing rapidly worldwide, there is more interest to study the different aspects of obesity and its impact especially on healthcare outcomes and health related issues. Targeting non-surgical times in the operating room (OR), this study focuses on the effect of obesity along with clinical factors on pre-incision times in OR. Specifically, both the individual and combined effect of clinical factors with obesity on pre-incision times is studied. Results show that with the confidence of 95%, pre-incision time in the OR of obese patients is significantly higher than those for non-obese patients by approximately five percent. Findings also show that more complex cases do not exhibit significant differences between these patient subgroups.

1. Introduction

The International Obesity Task Force estimated that 300 million people worldwide are obese and 750 million more are overweight. Thirty percent of adults in the United States are obese (Source: CDC National Center for Health Statistics), and 20.6% of U.S. healthcare expenditures are coming from additional medical spending as a result of obesity[1]. These expenditures account only for the direct costs of medical spending as a result of medical visits, and prescription drugs. The indirect and hidden costs that obesity imposes on healthcare systems need to be further analyzed and investigated. The operating room (OR) is considered to be the most expensive resource to keep open and the most profitable resource in hospitals; mainly related to the expensive staffing of these resources. Studies reveal that duration of surgery and length of stay of obese patients is longer when compared to non-obese patients ([2], [3]). However, there is less research focusing on effect of obesity on non-surgical times in the OR.

It is important to better understand the effect of obesity on non-surgical times in the OR as this helps to (re)design workflow in the OR and to accurately predict OR times for surgical scheduling. Extensive amounts of time stamp data and clinical factors are collected in the OR which provides the opportunity for understanding trends and patterns. Using this data, this research focuses on the duration of pre-incision times in the OR, and the impact of obesity along with other clinical factors on these times. Difficulties with the positioning of obese patients as well as difficulties with anesthesia process for these patients are some of the problems reported for obese patients [4],[5] by OR staff during pre-incision times. However, it is not clear whether these difficulties make a difference in the duration of pre-incision times in OR. For instance, spine surgery involves extensive amount of tasks around patient positioning. Further, the Agency for Healthcare Research and Quality reported spine surgery as the most costly operating room procedure performed in U.S. hospitals. They also ranked spine surgery sixth in terms of frequency of the surgeries that are performed in the U.S. [6]**.** For these reasons and to ensure a homogenous group of surgeries, this paper will focus on pre-incision times of spine surgeries. Specifically, this research will address two questions: 1) how does obesity impact pre-incision times in OR for spine surgery? and 2) how do clinical factors combined with obesity influence pre-incision times of spine surgery?

The answer to these questions allows better understating of the impact of clinical factors on workflow, and help OR managers better plan staffing and surgical scheduling, thereby reducing delayed surgical starts and overtime. This also addresses some of the indirect costs that obesity may levy on healthcare. The reason for denominating the costs of obesity is not to stigmatize obese population, rather to inform clinicians and administrators to better plan for such patients and reduce unsafe surgeries, delays and excessive cost of overtime.

The rest of the paper follows with a background section with a brief review of the literature, description of workflow during pre-incision time, and methods and results sections. Next, a discussion and conclusion follows with future research direction.

2. Background

The prevalence of the obesity epidemic is increasing worldwide along with the health problems associated with obesity. A BMI higher than 30 kg/m^2 is generally regarded as obese [7]. There have been several causative factors associated with obesity such as lack of exercise, genetics, poor lifestyle choices, and poor diet [8]. Obesity is associated with several adverse clinical conditions including cardiac disease, diabetes, arrhythmias, increased all-cause mortality, high stress levels, lack of sleep, and increased cancer incidence [8]. Additionally, obese women who have early-stage breast cancer have been shown to have worse survival rates than their non-obese counterparts [9]. Complications[10], mortality[11], outcomes[12], readmissions[13], duration of surgery[14], levels of infection[14], and financial costs[13] have all been shown to be higher in obese populations undergoing surgery.

The impact of obesity on certain aspects of spine surgery has been studied. Bederman et al., [10] studied social structure, demographics, personal, health beliefs, medical need, and community resource factors impacting rates of knee, hip, and spine surgery, and found that obese patients had higher levels of surgical complications. Kalanithi et. al., [11] studied obese patients undergoing spinal fusion and discovered that obese patients had higher rates of mortality. Fang et. al., [12] studied obese patients who had undergone spinal procedures which had become infected after surgery and identified risk factors. Fang et. al., [12] found obese patients to have worse outcomes than their non-obese counterparts. Silber et. al., [13] examined obese patients and specifically studied financial and medical outcomes that are associated with surgery and found that obese patients have higher numbers of readmission and greater financial costs than their non-obese counterparts. Mehta et. al., [14] studied the role of weight in post-operative infections and found that obese patients have longer surgical durations and higher levels of infection than non-obese patients. There are also reports for higher chance of adverse events on obese patients [8].

Obesity is becoming more prevalent among patients who are considering orthopedic surgery and spine surgery [7]. However, there is still some controversy regarding obesity's association with failed reconstructions, complications, and reoperations after spine surgery is performed [7]. Also the relationship between obesity and the duration of non-operative times in the OR is unknown. The aim of this study is to examine how obesity and certain clinical factors specifically affect preparation times that occur within the OR. Next, a brief background on the workflow of various activities in the OR along with a description of clinical factors for spine surgeries is provided.

2.1. Clinical Workflow during Pre-incision times

Surgery related activities on the day of surgery consist of two major set of tasks. The first set of tasks occur outside of the OR, and the second set of tasks is performed inside the OR. The OR-related times (from the time the patient enters the OR to the time the patient exits the OR) is divided into three common time periods: time to incision (TtI), skin to skin (StS), and closure to exit (CtE) (Figure 1). As the names imply, TtI is from the time the patient enters the OR until the time the incision starts; StS time is from the incision start to closure; and finally CtE is the time after closure to the time the patient is taken out of the OR. The focus of this study is on TtI in spine surgeries. There are several events that take place during TtI, some of which depend on the type of surgery and the institution where the surgery is performed (Figure 2). The order of these events however may vary based on the approach taken for the surgery (anterior or posterior or staged, to be discussed shortly) and the anatomic location of incision. As the patient is moved to the OR on the cart, the staff verifies patient's information, discuss the procedure with the patient and answer last minute questions. The patient is then prepared for intubation. After intubation, IV (intravenous) lines, other lines (e.g. central venous catheter or CVC) are placed as needed for intraoperative administration of fluids, medications and anesthesia, as well as to draw blood. Then the patient is positioned on the OR bed for the approach of surgery to be performed (anterior, posterior, combination of anterior and posterior which is called staged) dependent upon the type of surgery and targeted levels of the spine. The most commonly used positions in spine surgery are supine, prone, and sitting. In a supine position, the patient is slid from the cart to the OR table and remains face up. This position allows the surgeon adequate access to the anterior aspects of the chest, abdomen and pelvic region. In this position, staff should ensure that the head and cervical spine are aligned along the midline. There are also guidelines for protection of the heels, elbows, and upper arms with soft padding to prevent pressure ulcers from long surgeries. The prone position is usually achieved by intubation on the patient's stretcher and movement after anesthesia onto the patient's abdomen by either manually logrolling or using a specifically made bed

which rotates the patient on a table that supports the head, upper chest, hip area and legs, to leave the center of the chest without pressure. Care must be given to avoid excessive pressure on the eyes, ears, nose, breasts, and male genitalia [15]. In addition special attention must be given to the knees and ankles, head and to the arm position to avoid injury. Less commonly used position in spine surgery is the seated position. Positioning of the patient can require several members of the OR team, and require a long time to perform, contributing to high variability in TtI time.

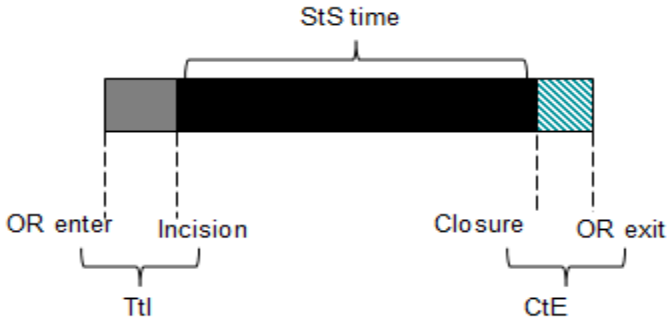


Figure 1. Time durations in the operating room

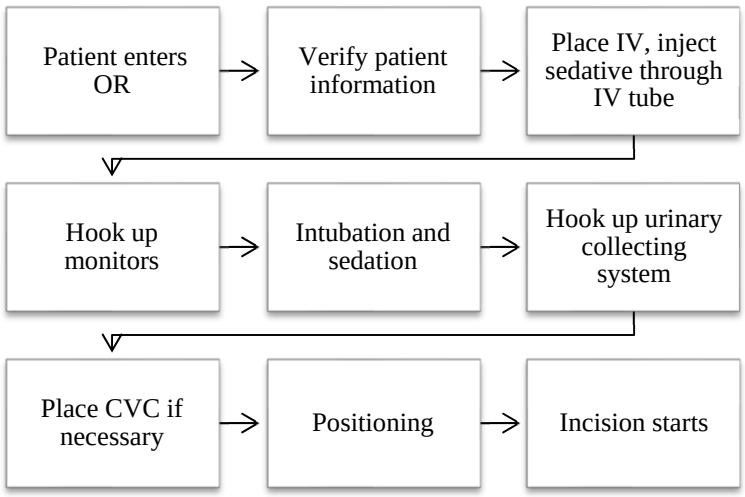


Figure 2. Pre-incision time workflow

3. Methods

3.1. Sample and Assumptions

This study is based on records of 3,677 adult patients aged 30 years or older (gathered from Surgical Information Recording System, SIRS) which had planned spine surgeries in a large tertiary academic medical center in the mid-west, from the start of 2010 through the end of the 2012 (3 years). This data includes a total of 2,611 procedures performed by neurosurgeons and 1,066 performed by orthopedic surgeons. Outliers were excluded where TtI was greater than 2.3 hours which was the highest one percent of the distributed times.

3.2. Measures

The primary outcome variable studied is time to incision (TtI) which is measured in minutes, and one of the independent variables is obesity measured as body mass index (BMI). The other independent variables include ASA category, number of spine levels, approach and fusion, and from here forward, these four will be referred to as clinical factors. Table 1 shows the data spread for spine patients over the four BMI categories defined by Centers for Disease Control & Prevention. As could be seen, the majority of spine surgery patients are overweight or obese.

Table 1. Patients in BMI categories

	Underweight	Normal	Overweight	Obese
% of Total	0.8%	19.4%	36.1%	42.8%
N	31	713	1329	1574

Since the primary focus was on obese patients, and the underweight patients accounted for less than one percent, these underweight patients were dropped from further analyses. Further, with significance of 95% there was no difference in TtI between the normal and overweight patients based on statistical hypothesis testing (difference of less than a minute on average), hence, they were lumped together under non-obese patients and compared against obese patients. In the final sample, 43% of the patients were obese and the rest were normal or overweight.

The American Society of Anesthesiologists (ASA) adopted a five-category physical status classification system; a sixth category was later added. These are:

1. Healthy person.
2. Mild systemic disease.
3. Severe systemic disease.
4. Severe systemic disease that is a constant threat to life.
5. A moribund person who is not expected to survive without the operation.
6. A declared brain-dead person whose organs are being removed for donor purposes.

The ASA physical status classification is used for assessing the clinical condition of patients before surgery. Only patients from ASA categories one through four were present in the sample.

The number of levels as the name implies refers to the number of spine levels in the procedure. Level 0 is used to represent cases where no spines are worked on, rather the procedure involves just a washout. As the number of levels of spine increases from one through five, TtI increases rapidly. Then for levels six through eight, TtI flattens and again starts to increase slowly for levels greater than nine through 19. However, these cases are very rare. Specifically, there are only 57 cases with spine levels greater than eight. Based on these facts, the sample includes levels 0 through five, 6-8 and 9+, thus a total of seven groups.

Spine surgeries are performed using three different approaches, anterior, posterior, and staged. Anterior approach is applied when spine is accessed from an anterior position whereas in posterior approach, spine is accessed from a posterior position. Staged approach is the combination of anterior and posterior where both of these approaches are used in one surgery and with separate incisions. The patient is usually positioned supine for anterior approach and prone or rarely seated for posterior approach. In staged approach, the positioning is done based on the order in which anterior and posterior approaches are used, and the patient is repositioned after the first incision is closed. Thus the variable approach can be one of anterior, posterior or staged.

The spinal column is made of individual vertebrae separated by discs. Fusion surgery removes much or the entire disc, replacing it with bone grafts or hardware (screws and rods). Hardware keeps the spine stabilized while the bones grow together or fuse. This variable is binary in nature representing the presence or absence of fusion.

3.3. Analysis

Two types of analyses were conducted, 1) regression analysis to study the effect of the five factors on TtI, and 2) t-tests to study the combined effect of the factors (obesity and ASA category, obesity and spine levels, obesity and approach, and obesity and fusion). TtI is normally distributed, and hence the student t-test was selected. We also studied the individual effect of each of the factors on TtI. The results of this analysis are described in the next section.

4. Results

4.1. Regression and Individual Effects

Based on the regression analysis, all five factors have highly significant effect on TtI. This could be seen in Table 2.

Table 2. Regression analysis test results

Factor	Sum of Squares	F Ratio	P-Value
Obesity	6578.75	11.1679	<.0001
ASA category	22996.32	26.0253	<.0001
Number of levels	105712.6	21.1123	<.0001
Approach	12426.1	21.0942	<.0001
Fusion	96427.46	327.3851	<.0001

When TtI is compared for obese and non-obese patients, the time is significantly different (slightly higher for obese patients by three minutes on average). Among cases with ASA categories of one through four, as the ASA category increases, TtI also increases. Figure 3 shows how TtI is changing for different ASA categories. As for approach, staged surgeries have on average the longest TtI among the three approaches, taking 3.5 minutes longer than those surgeries with an anterior approach, which in turn takes on average 15 minutes longer than those with posterior approach. The sample includes 80% of cases with posterior approach, 12% anterior, and only 8% of cases with staged approach. With the fusion, the sample includes 47% of spine cases with fusion. Cases with fusion have TtI which on average is 20 minutes longer than cases without fusion.

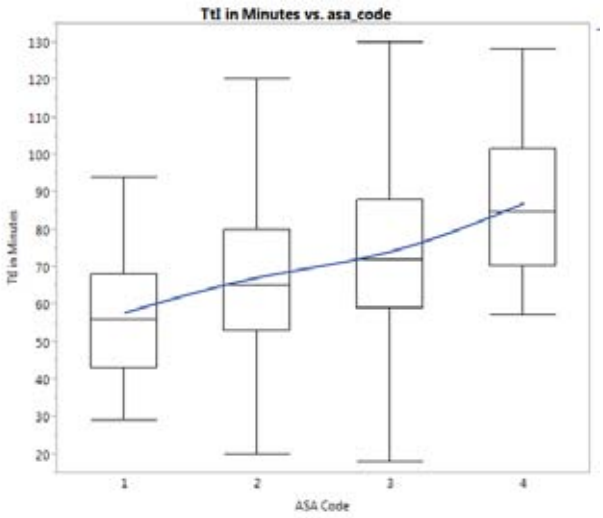


Figure 3. ASA categories and time to incision

4.2. Combined effect

When the combined effect of obesity and the clinical factors were studied, some of the factors were significant. Table 3 shows the results of the statistical results from ANOVA for effect of four clinical factors combined with obesity on TtI. These results indicate that TtI of obese and non-obese patients for any given ASA category is not statistically different. As for the effect of spine levels combined with obesity on TtI, surgeries done on one or two levels of spine, there is a significant difference between TtI of obese and non-obese patients (that is for 68% of cases). For single spine level surgery, TtI of obese patients is on average four minutes longer than non-obese patients. For two levels, this difference is three minutes on average. For higher number of levels of spine, the difference between obese and non-obese patients is not significant. The comparison of averages of TtI for obese and non-obese patients with different number of levels of spine are shown in Figure 4.

Table 3. ANOVA results for combinational effect of factors with obesity

Factor	Factor Value	P-Value
ASA category	1	0.3647
	2	0.3362
	3	0.0781
	4	0.2465
Spine levels	0	0.8095
	1	<0.0001
	2	0.0183
	3	0.4752
	4	0.3819
	5	0.4567
	6-8	0.7371
	9+	0.8456
Approach	Posterior	<0.0001
	Anterior	0.0271
	Staged	0.3985
Fusion	Yes	<0.0001
	No	<0.0001

The statistical results show that TtI for obese patients with posterior and anterior approach is significantly different from TtI for non-obese patients for the same type of approach. With the staged approach however, there is no significant difference between TtI of the two groups of obese and non-obese patients. Our analysis indicates that TtI of obese patients within each approach takes few minutes longer than non-obese patients (four minutes for posterior and anterior approaches, and three minutes for staged approach). Based on the results of ANOVA shown in Table 3, fusion has a significant effect on TtI of obese and non-obese patients. Obese patients with fusion have on average TtI which is two minutes longer than TtI on non-obese patients. This difference for patients without fusion tends to be five minutes on average.

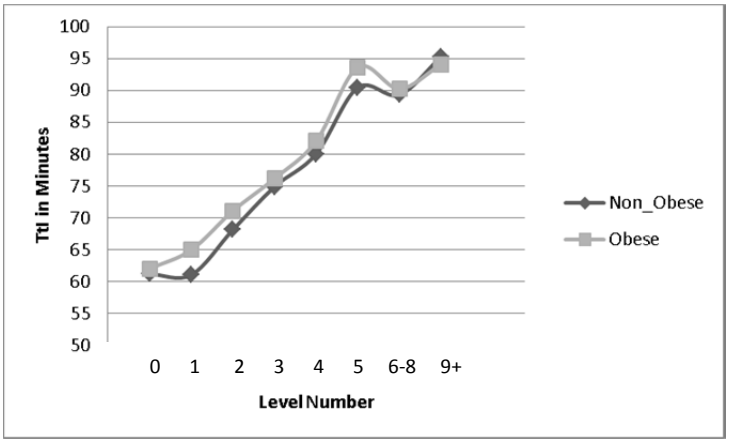


Figure 4. Average time to incision by level

5. Discussion

Obesity is considered as a factor that influences tasks that take place during TtI. The primary goal was to study the individual and combined effect of obesity and clinical factors on TtI. As for individual effect, all five clinical factors are significant. Obesity makes for a low difference of just three minutes, however the clinical factors account for much higher differences (19 minutes for approach, 20 minutes for fusion and up to 30 minutes for ASA category). This increase might be because patients with higher ASA category need more prep time in the OR. This is partly due to the complexity added as a result of the existing condition of the patient which may require additional equipment to be attached to the patient. With patients having higher ASA categorie**s**, there also may be difficulties with patient positioning due to increased frailty caused by illness. Results show that surgeries using an anterior approach have an average TtI which is 15 minutes longer than that for the posterior approach, despite the prone and seated positioning used in the posterior approach typically being more complex and taking longer time. Input from the OR team suggests that increased time is not patient-related, but OR staffing related; for the surgery done using anterior positioning (in lower levels of spine), the presence of a non-orthopedic surgeon (usually a vascular surgeon) is often required for incision to start. This is mainly due to the fact that accessing spine from anterior position requires a second surgeon to prepare the abdomen for the greatest access as possible to the front of the spine while keeping the large vessels safe, leading to extra preparation time. Also, since the presence of a second surgeon is required, the orthopedic team may need to wait until the second surgeon arrives from their primary surgery or other tasks. This factor seems to be the main reason for TtI of anterior cases to be significantly longer than posterior cases. However, it is noted that posterior approach is more common among the spine surgeries in our data. The effect of obesity could be explained due to the difficulty with intubation and positioning of obese patients. Obesity and some other factors such as limited neck mobility and poor mouth opening accounts for two thirds of all contributing factors for difficulty in intubation [16]. Obesity is also related to other contributing factors for difficult intubation such as short and fleshy necks. Obesity may also be related to difficulty in line placement, due to difficulty palpating landmarks for central lines, or peripheral lines [17]. Other difficulties as a result of patient obesity during preoperative tasks in the OR include problems with positioning of obese patients including difficulty lifting the patient, size of the bed and ensuring patient safety by requiring extra personnel. These difficulties are often assumed to require extra time to get resolved.

As for the combined effect, there is a significant difference for surgeries with one or two levels, but not for higher levels. When effect of approach is combined with obesity, for approaches anterior and posterior, there is a significant difference between obese and non-obese patients in terms of TtI, but for staged approach the difference is not significant. Overall, obesity is shown to be a significant factor affecting TtI. However, when the cases are more complex, obesity does not make a significant difference in TtI of patients. For instance, surgeries, with higher ASA category, higher number of spine levels, staged-approach, do not exhibit significant differences between obese and non-obese groups of patients. This might be due to the fact these complex surgeries are uncommon and the sample size is extremely small. To increase sample size, multi-year surgery data can be combined to better study multivariate effect of factors. Such an approach also poses the problem of changes in culture, practice and technology over time. Hence, multi-institution data need to be warehoused in registries to address this lack of sufficient sample size for more complex cases.

6. Conclusion and Future Work

This study focused on the effect of obesity and clinical factors on TtI of spine surgery. Both, individual effects and combined effects (of clinical factors with obesity) were studied. Obesity shows to be significantly affecting TtI as an individual factor and in combination with the factors such as fusion, number of levels and approach. With more common surgeries such as one or two levels of spine, and anterior or posterior approaches, and for both fusion and non-fusion cases, obese patients have longer TtI. For less common cases such as those with higher ASA categories, higher levels of spine, and a staged approach, there is no significant difference between obese and non-obese patients. However, complex cases need to be further studied. As for obesity, our study was based on BMI and does not focus on extremely tall patients. Height is also a factor that affects positioning of patients, including challenges with instruments, bed size, etc. The next phase is to focus on including height and weight as separate factors and study the effect on TtI.

References

1. Cawley, J., Meyerhoefer, C., *The medical care costs of obesity: an instrumental variables approach.* Journal of health economics. **31**(1): p. 11.
2. Gu. GF., H., SS., Ding, Y., Jia, JB., Zhou, X., *The effect of body mass index on the outcome of minimally invasive surgery for lumbar spinal stenosis complicated with lumbar instability.* Chin J Spine Spinal Cord, 2012. **22**: p. 4.
3. Hardesty CK., P.-K.C., Son-Hing JP., Thompson GH., *Obesity negatively affects spinal surgery in idiopathic scoliosis.* Clin Orthop Relat Res, 2013. **471**: p. 5.
4. Nielsen KC, G.U., Steele SM, et al., *Influence of obesity on surgical regional anesthesia in the ambulatory setting: an analysis of 9,038 blocks.* Anesthesiology, 2005. **102**: p. 6.
5. Rocke DA, M.W., Rout CC, Gouws E., *Relative risk analysis of factors associated with difficult intubation in obstetric anesthesia.* Anesthesiology, 1992. **77**: p. 6.
6. *Healthcare Cost and Utilization Project, statistical brief,* in *Weiss A and others*2014.
7. Jiang, J., et al., *Does Obesity Affect the Surgical Outcome and Complication Rates of Spinal Surgery? A Meta-analysis.* Clinical orthopaedics and related research, 2014. **472**(3): p. 968-75.
8. Ramsay, M.A., *The chronic inflammation of obesity and its effects on surgery and anesthesia.* International anesthesiology clinics, 2013. **51**(3): p. 1-12.
9. Ampil, F., et al., *Morbid obesity does not disadvantage patients with in situ or early-stage carcinoma undergoing breast-conserving surgery.* Anticancer research, 2013. **33**(9): p. 3867-9.
10. Bederman, S.S., et al., *Drivers of surgery for the degenerative hip, knee, and spine: a systematic review.* Clinical Orthopaedics and Related Research®, 2012. **470**(4): p. 1090-1105.
11. Kalanithi, P.A., R. Arrigo, and M. Boakye, *Morbid obesity increases cost and complication rates in spinal arthrodesis.* Spine, 2012. **37**(11): p. 982-988.
12. Fang, A., et al., *Risk factors for infection after spinal surgery.* Spine, 2005. **30**(12): p. 1460-1465.
13. Silber, J.H., et al., *Medical and financial risks associated with surgery in the elderly obese.* Annals of surgery, 2012. **256**(1): p. 79-86.

14. Mehta, A.I., et al., *2012 Young Investigator Award winner: The distribution of body mass as a significant risk factor for lumbar spinal fusion postoperative infections.* Spine, 2012. **37**(19): p. 1652-1656.
15. MC, V.L.D.I.R., *Positioning of patients for operation. Stiefel RH, Electricity, electrical safety, and instrumentation in the operating room.* 1993: p. 43.
16. Williamson JA, W.R., Szekely S, Gillies ER, Dreosti AV, *The Australian Incident Monitoring Study. Difficult intubation: an analysis of 2000 incident reports.* Anaesth Intensive Care, 1993. **21**: p. 5.
17. Juvin, P.M., PhD; Blarel, Anne; Bruno, Fabienne; Desmonts, Jean-Marie MD, *Is Peripheral Line Placement More Difficult in Obese Than in Lean Patients?* Anesthesia & Anangesia, 2003. **96**(4).

Enabling Locally-Developed Content For Access Through the Infobutton By Means of Automated Concept Annotation

Nathan C. Hulse, PhD[1,2], Jie Long, PhD[1], Xiaomin Xu, MS[1] Cui Tao, PhD[3]
[1]Intermountain Healthcare, Salt Lake City, UT; [2]Department of Biomedical Informatics, University of Utah, Salt Lake City, UT; [3] University of Texas School of Biomedical Informatics, Houston, TX

Abstract

Infobuttons have proven to be an increasingly important resource in providing a standardized approach to integrating useful educational materials at the point of care in electronic health records (EHRs). They provide a simple, uniform pathway for both patients and providers to receive pertinent education materials in a quick fashion from within EHRs and Personalized Health Records (PHRs). In recent years, the international standards organization Health Level Seven has balloted and approved a standards-based pathway for requesting and receiving data for infobuttons, simplifying some of the barriers for their adoption in electronic medical records and amongst content providers. Local content, developed by the hosting organization themselves, still needs to be indexed and annotated with appropriate metadata and terminologies in order to be fully accessible via the infobutton. In this manuscript we present an approach for automating the annotation of internally-developed patient education sheets with standardized terminologies and compare and contrast the approach with manual approaches used previously. We anticipate that a combination of system-generated and human reviewed annotations will provide the most comprehensive and effective indexing strategy, thereby allowing best access to internally-created content via the infobutton.

Introduction

Information needs and infobuttons

Information gaps at the point-of-care have been well documented in the medical literature, and they pose a real risk in ensuring that current best practice is reflected the care that patients receive[1,2]. Several approaches to addressing these information needs have been presented in the informatics literature[3,4]. One of the more notable means for addressing this issue is the infobutton, a context-aware linking resources that points users to relevant clinical reference materials at the point of care, typically accessed in clinical workflows while users are engaged with routine tasks in electronic health records[5,6]. The profile of the infobutton has been elevated in recent years through the development and refinement of international standards and implementation guides in Health Level Seven (HL7) as well as its inclusion in the Meaningful Use criteria, as part of the United States government incentive program for broadening the uptake and usage of key features within electronic health records[7].

Figure 1 illustrates the flow of information within a typical infobutton exchange, as prescribed by the HL7 standard. In this type of exchange, a request is manifest from the electronic health record (EHR), passing several key parameters including 1) the primary (and secondary if applicable) concept of interest, the age and sex of the patient, the role of the intended recipient, the care setting, and even the user's current task in the EHR. As information moves to the infobutton manager, key tasks including logging, selection rule logic, translation of internal codes to reference terminologies (like SNOMED, LOINC, ICD-9, etc), conversion to appropriate URLs to request further information take place. As these requests propagate outward to content repositories, including externally licensed and internally developed content, each repository is tasked with identifying relevant content and responding with metadata that can allow the users at the point of care to access that relevant content. Most content providers do so by building upon a combination of intelligent text parsing and indexing materials with relevant concepts from standardized terminologies. The approach and challenges faced by content providers in dealing with these terminology issues and aligning their content and services with these systems have been well-documented by Strasberg et al[8].

Although the infobutton standard allows for flexibility by enabling requests to pass free text in the place of coded concepts, anecdotal evidence has consistently shown that infobutton performance improves greatly when content providers index their materials appropriately using relevant terminologies, instead of just relying fully upon free text queries. This poses some unique challenges for content providers in that they must index content using appropriate terminologies and granularity, as well as deal with the important issue of 'rank' among results that come back from these queries. It also poses a unique challenge for institutions who have infobuttons, but also want to implement access to local content repositories from the infobutton. These organizations must address the infrastructure necessary

for centralizing, searching, and providing standards-based responses against their own content in order to enable this type of access from the infobutton. Open-source efforts like the Infobutton Responder within the OpenInfobutton open-source project are reflective of this need[9]. Local content owners need the ability to centralize their content, assign appropriate metadata, and include keywords indices from relevant reference terminologies in order to make their local content visible from these frameworks.

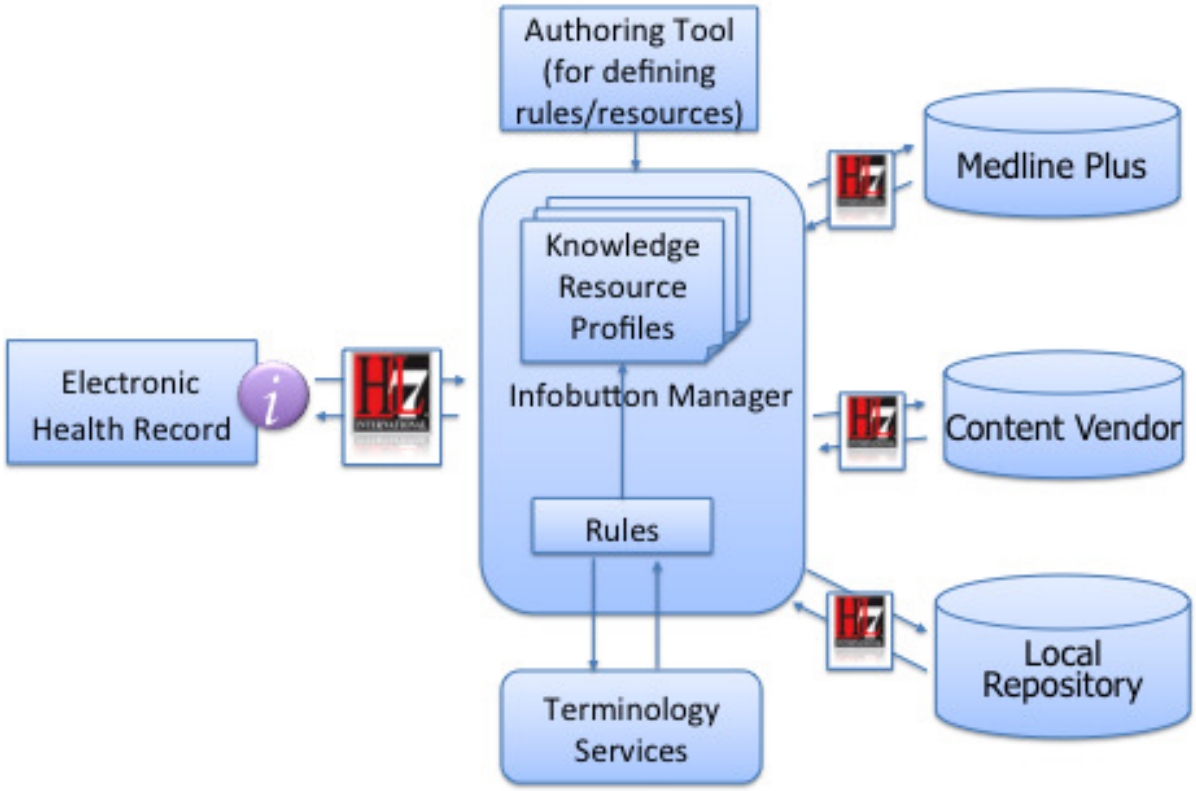


Figure 1- Flow showing data exchange and processing through infobutton manager. Request is passed from EHR to infobutton manager, where matching, code translation, and request output send queries to content providers. Content providers receive requests, process them using standardized parameters, and return best results by means of search algorithms and indices using standardized terminologies.

Background

Infobutton use at Intermountain Healthcare

Intermountain Healthcare is a not-for-profit integrated healthcare delivery system based in Salt Lake City, Utah. It provides healthcare for the entire state of Utah and parts of southeastern Idaho. Intermountain maintains 22 inpatient hospitals (including a children's hospital, an obstetrical facility and a dedicated orthopedic hospital), more than 185 outpatient clinics, and 18 community clinics serving uninsured and low-income patients. Intermountain provides primary and specialty care for approximately half of the residents of the state of Utah. Its clinical information systems are internally built and maintained, dating back several decades. They include inpatient and outpatient EHRs, a decision support framework, an infobutton manager platform, and a patient portal. They have been widely-regarded as 'state-of-the-art' and are well recognized in the literature for supporting best care practice with clinical decision support interventions[10]. Intermountain's infobutton manager (a key component of this effort) is used regularly and its development and uptake have been detailed previously[11-13].

Infobuttons have been in use at Intermountain Healthcare for over 14 years. They have been integrated in two separate clinical systems, including usage from 4 major modules within these systems including medication ordering, lab result review, problem list, and microbiology. Usage has steadily increased over the years, with over 1,700 unique monthly

users, accounting for over 18,000 infobutton sessions per month. They have also recently been made available for patient use from our patient portal in the same modules. At present, we are transitioning our production infobutton manager environment from a legacy, internally-built system over to a local implementation of the OpenInfobutton initiative.

At Intermountain Healthcare, we have maintained an enterprise clinical knowledge repository designed for centralizing, versioning, and distributing internally-developed clinical content for the past 15 years[14]. In the past, a very limited subset of these materials have been made available through our infobutton manager, but the process involved to do so was fairly arduous. Terminology engineers, knowledge engineers, and content experts had to work in concert with one another to assign specific terminology concepts to the metadata for these documents in our knowledge repository, as well as in corresponding domains in our healthcare data dictionary. The tasks involved were such that a content expert would be ill-equipped to index and maintain content without substantial technical help. Furthermore, the content authors who create and maintain many of these content resources may or may not be well-versed in the relevant terminology standards for indexing that content, and as such, may not be best suited for that type of task. As a result, only small portions of Intermountain's internal content library have been accessible via the infobutton.

Current needs and strategic directions

As part of a broader effort at Intermountain to make internally-developed content more accessible to users through channels like the infobutton, we have recently engaged with a group of content authors who maintain a library of patient education materials. The collection includes over 1,600 locally-authored patient education handouts, worksheets, forms, and we are exploring means for automating the annotation process described above. In current processes, content authors can and do assign free-text keywords to their content to facilitate local search, but have never had need to index the content using standardized terminologies. We envision an environment in which automated content analysis and suggested codes for annotation accompany the normal content publication and versioning processes typical in knowledge management. As such, we hope to draw upon the unique strengths of both computer-based and human annotators in indexing our local content libraries with relevant, appropriate metadata.

Methods

Terminology selection

We prefer to use an annotation approach that is as generalizable as possible, and be capable of producing ICD-9 and SNOMED-CT codes. Our previous efforts at annotating and exposing content through our infobutton manager have been based on our local terminology, the Healthcare Data Dictionary. This terminology, however, is not intended to be an independent reference standard. Our contracts with external providers have shown that these are the two terminologies that have the most support among content providers who index their content accordingly. Since our content of choice was most applicable for requests from the problems domain, we deemed that these were appropriate choices for the effort.

Tools and Services

We analyzed several annotation engines for possible inclusion in our effort, including Metamap and the NCBO BioPortal Annotator. Although both have their strengths, we opted to use the BioPortal Annotator, which is based on the Mgrep concept recognizer[15]. It has been shown to have high precision in a study that used it for recognizing disease terms in descriptions of clinical trials (87%) and has been characterized as a lightweight concept recognizer that can provide fast, good-enough concept recognition in clinical texts[16-17]. It also already contains lexicons for both SNOMED and ICD-9, an important feature in the overall effort we were pursuing.

We selected a convenience sample of 59 documents (from a subcollection of the library) to test our approach. In our content repository, all documents consist of two artifacts, a standardized XML header containing uniform metadata about each document, as well as the document body itself. For the annotation effort itself, we decided to focus strictly on the body of the document. Although the metadata in the document header could be useful for indexing, we felt that annotating the corpus of the knowledge itself would be a more generalizable approach, with fewer dependencies on the local approach to metadata management.

All of the 59 documents were stored in PDF format. In order to extract and preserve the content of these documents for analysis and annotation, we used the open-source Apache Tika software, a toolkit that detects and extracts metadata and structured text content from various documents using parser libraries[18]. After processing these documents through Tika, the resultant text streams files were stored to memory and streamed into a local instance of the NCBO Annotator.

The annotation engine produces XML files for each document analyzed, containing a rank-ordered list of identified concepts, links back to the NCBO website for more detailed views about each concept, character mappings for each concept identified and match types associated with each reference. Figure 2 illustrates the flow of information from the documents, through TIKA, the NCBO Annotator and the resultant output. After processing each file and creating an XML file with the annotations data, we extracted the relevant concept terminologies, identifiers, and text representations and consolidated them into a text file for analysis. All 59 documents were able to successfully be processed and annotated. The NCBO Annotator found one or more matching concepts for all of the documents in either SNOMED-CT or ICD-9.

Figure 2- Illustration depicting flow of information from PDF files through Tika software, the resultant text stream feeding through the NCBO annotator, and the outputted XML annotations files

Separately, we extracted all of the human-derived keywords for each document from our clinical knowledge repository. As mentioned earlier, each document has a corresponding header document which contains metadata about the document, including a keywords section that accommodates both free text and coded keywords. Although most of the keywords were free text, a subset of these were ICD-9 codes. For purposes of comparison, we normalized all the concepts to free text and organized them for analysis.

Results

Keyword Overloading by authors in manual keyword entry

One of the early results of our analysis is that our authors have found ways of 'overloading' or misusing the existing metadata fields in the header. They have developed an internal identification pattern and have used the keyword field as if they served as an 'alternate ID' field. Presumably, they have done so in ways that allow them to search and retrieve content using these identifiers. For example, a document about depression would contain expected keywords like "clinical depression", "mental health", "major depression" but also internal codes such as "CPM019". This did prove problematic in the analysis of the data, so we removed these data points from consideration.

Dealing with 'noise' concepts in the annotation space

One of the interesting results of processing the document set through the NCBO toolset was that raw processing of the documents against a concept space like ICD-9 or SNOMED resulted in large concept sets. Collectively, annotating the 59 documents against ICD-9 resulted in over 1138 ICD-9 codes (roughly 19 per document) and over 35,000 SNOMED-CT codes (338 per document). Some of the ICD-9 codes were deemed to be irrelevant in the context of concepts relevant to the problem list (the proposed point of integration via infobutton for this content). Concepts such as 'process', 'hospital facility' and 'time' were clearly concepts found in the document, but were deemed as less relevant in the context of the specific task of indexing content for linking from the problem list. This number was cut somewhat by further processing the data to remove category or 'container' concepts in ICD-9 (e.g. T052, or 'Activity').

One of the reason for which there were a surprisingly large number of concepts in the SNOMED-CT annotation space is that the annotator included retired concepts in the output. We were able to consolidate this list by programmatically reducing the search space by collapsing identical 'concept term' entries. We also leveraged an existing mapping table to our local terminology space to identify active SNOMED terms. This reduced the overall SNOMED-CT concept space by more than 50%.

Given that the practicalities of annotating content have shown that assigning that many identifiers to the content is not useful in making search results precise, we employed other methods to further narrow the concept space. We

preprocessed the documents by first annotating the content using a MedLine Plus ontology available in the NCBO annotator, and then subsequently annotated that concept space using the ICD-9 and SNOMED-CT lexicons. This was particularly useful in that it focused the output concept space to be relevant to concepts in the MedLine Plus ontology, one that is more tightly focused on diseases and conditions. This further reduced the concept space of the SNOMED-CT codes by almost 80%.

Collectively, after accounting for retired concepts in SNOMED-CT, removing some of the less-relevant categorical ICD-9 codes and pre-processing the annotation space against the MedLine Plus ontology, we derived a unique concept set of 117 ICD-9 concepts and 741 SNOMED-CT concepts, some of which that were identified in multiple documents in the set we tested.

Similarity between human-identified keywords and annotated concepts

The analysis of similarity between the two concept spaces (human-identified keywords and machine-derived codes) made for a somewhat difficult process. Automatic transformation of free-text entries to coded elements resulted in poor translation and very scant results overall. Translating codes back to text representations at least allowed for each identified concept to have a representation, but made the comparison between concept spaces somewhat difficult, due to the inherent nature of free text. The differences in the overall number of concepts identified per document between the human and computer raters, combined with the non-categorical data made kappa analysis not a good option. In the end, we opted to convert coded entries to free text and present some metrics of similarity based off some computed and some human-analyzed metrics.

For purposes of illustration, we present our analysis of a random sample document that we reviewed manually, aligning and matching computer-derived and human-derived concepts to compare and contrast the concepts identified by each respective sets. This document focused on the management of low back pain, including assessment, treatment options and goals. Table 1 contains keywords identified by the content authors, ICD-9 and SNOMED-CT. Although we present data specific to this particular document, the observations hold true across multiple documents in the set.

Human coded keywords

Free-text captured keywords often have special characters. It is often hard for machine to analyze and match to terms. The data not normalized and standardization of capitalization, and the approach to using nouns vs. adjectives is often haphazard. As noted earlier, most of the documents had keyword entries that serve as internal identifiers, such as CMP009d. Human assigned identifiers include no rank of relevance and appear in apparently random order.

Another interesting observation is the inclusion of concepts derived from knowledge outside the space of the concept domain. The content authors referenced a "start back protocol" while annotation engine missed the term completely. The referenced concept is the Keele STarT back tool that was developed by the Primary Care Centre at Keele University, with funding from Arthritis Research UK. It is well validated for low back pain, but is also being used widely for other spinal pains. Other documents contained similar examples, including references to instruments like the PHQ-9 for mental health.

Coded concepts

The ICD9 code list provided a much smaller concept space. In an extreme example, the ICD-9 annotation provided 20 concepts while the SNOMED-CT annotation gave 2087. The ICD-9 annotations often don't have as much detail as the SNOMED-CT codes, in part due to the granularity of the terminologies. The ICD-9 codes sometimes included concepts from unexpected domains, (including activities like 'yoga' that were found) but the tools allowed us the flexibility to include or restrict those as we saw fit. The pre-annotation against the MedLine Plus ontology helped significantly with the large numbers of concepts we derived when using SNOMED-CT. We opted not to index the content directly against the MedLine Plus ontology because these codes aren't typically used as reference terminologies in infobutton implementations.

In all three sets of concepts, the overarching concepts of 'pain' and 'chronic pain' are present in all three, but the SNOMED-CT matching offered more specificity by listing 'low back pain' as its top concept. The ICD-9 annotation found 'cauda equina syndrome' which was not found in the SNOMED-CT set, but was present in the

human keyword space. Overall the major, expected concepts were present in most annotation sets.

Table 1- Human derived keywords, and computer-derived ICD-9 and SNOMED-CT codes for a patient education sheet dealing with low back pain.

Human Keywords	ICD9CM	ICD9 text	SNOMEDCT	SNOMEDCT text
lumbar	338-338.99	PAIN	279039007	LOW BACK PAIN
radiculopathy	338.2	CHRONIC PAIN	62482003	LOW
mechanical back pain	780-789.99	SYMPTOMS	161891005	BACK PAIN
start back	733.0	OSTEOPOROSIS	22253000	PAIN
cauda equina	782.3	EDEMA	82423001	CHRONIC PAIN
stenosis	756.12	SPONDYLOLISTHESIS	90734009	CHRONIC
disc	800-829.99	FRACTURES	257733005	ACTIVITY
sacroiliac	344.61	CAUDA EQUINA SYNDROME	363679005	IMAGING
facet			7396004	DIAGNOSTIC IMAGING
CPM009 and CPM009a			261004008	DIAGNOSTIC
CPM009d				
CPM009b and CPM009c				
Low back pain				
Low back pain, chronic				
Low back pain, acute				
DJD - Degenerative joint disease				
Degenerative L-S Spine Disease				
Degeneration of intervertebral disc				
Spinal osteoarthritis				
Degenerative Arthritis-Low Back				
Degeneration of lumbar intervertebral disc				
Degenerative spondylolisthesis				
Herniated intervertebral disc L3, L4				
Herniated intervertebral disc L4,L5				
Herniated intervertebral disc L5,S1				
Spondylolisthesis, grade 1				
Spondylolisthesis, grade 2				
Spondylolisthesis, grade 3				
Herniated Disc				
Sacroiliitis				

Discussion

Feasibility and performance

One of the key points we established early in the effort is that we were not interested in implementing a full-fledged natural language processing environment. In order to index the content of these documents, we wanted a system that was 1) capable of identifying key concepts, 2) able to detect and present overall concept relevance in ranked order, and 3) able to connect these concepts with identifiers in relevant reference terminologies. We believe that we have succeeded in these objectives. The system was able to perform these annotations quickly, giving us some hope that this type of analysis might be feasible at runtime, when a user is publishing a new document or updated version of existing content. If this proves to be the case, we may be able to leverage this system while the users are updating other required portions of metadata, giving them a chance at the end of that session to review, and accept or reject the proposed coded concepts that may be deemed relevant in the document by the NCBO Annotator. An approach like this would still allow human review of the concepts, but would push much of the burden of identifying and embedding relevant terminology codes in the metadata.

We have not yet conducted an analysis of determining whether the machine-identified concepts are acceptable for use from our clinical partners. Our limited matching analyses presented above suggest that while there is significant overlap in the human and computer-derived concept spaces, there are key differences as well. We plan to present our findings to our clinical education partners (the owners of the content set that is desired to be 'infobutton friendly'. Assuming that our clinical domain experts agree with the approach, we anticipate that we should be able to scale the process up quickly to index the full set of 1600+ patient education sheets. For each document, we would add the coded concepts from SNOMED and/or ICD-9 to the header metadata document in the knowledge repository. With this information in place, we could configure infobuttons to point to these resources in relatively quick fashion, provided a brief update to the enabled profiles in our infobutton manager configuration center.

Limitations

We have not accounted for negation (e.g. through a NegEx type of implementation) in the early phases of our project[19]. While this is not likely to play as important of a role in concept identification for educational materials as it does in understanding the meaning behind clinical notes (where negative findings and 'rule-outs' are common) we anticipate that implementing its inclusion would likely enhance the performance of the system. We intend to explore this in subsequent phases of our implementation

We have not been able to calculate and share Kappa statistics, since the computer and the users annotated content in different ways (coded elements vs. free text). We anticipate that we could further analyze similarity using ontology-based similarity tools that match and assign probabilistic similarity metrics to concepts under comparison. We expect that as we do so we will continue to find some overlap, but also areas where both the human and computer codes differ from each other, due to their different approaches in annotating content. The differences between tacit and explicit knowledge will likely be exhibited in the differences that analysis would show.

Even with coded concepts in place for all the documents in the collection, our approach doesn't yet provide a satisfactory way of addressing the important issue of presenting 'rank' and overall relevance back to infobutton users. It is highly likely that indexing all 1600 documents in this collection will result in some significant overlap in terms of multiple documents that are indexed with common concepts like 'diabetes' and 'obesity'. While the NCBO toolset does output a rank-ordered list of concepts in rendering its output, our current approach for storing metadata for coded concepts does not allow for capturing 'relative relevance scores' or the like. We anticipate that in providing a service that is capable of presenting multiple results back to an infobutton, ranked by relevance, we will need to further capture this detail either in the metadata, a separate index file, or through other search engine types of techniques to apply weights to the various results.

Another limitation to our approach is that at present, we are not indexing the content with relevant terminologies in other domains like medications and lab orders. Often, the document owners would insert text-based keywords like amoxicillin, doxycycline and others, indicative of treatment patterns for the condition being discussed. We can look to perform similar indexing using these types of terminologies in future efforts.

Our approach to annotation also does not account for materials which are written in Spanish. This will be less of a problem going forward in our local knowledge repository, since English and Spanish versions of the same document are semantically linked in our database. As such, annotating these concepts in the English narrative could be used to

encode concepts in the corresponding Spanish documents as well. In other document repositories where these types of linkages don't exist, the alternate language documents may not be able to be processed and annotated automatically.

Conclusion

We have demonstrated the feasibility of using existing annotators in the biomedical informatics research domain to identify key concepts of patient education materials, using ICD-9 and SNOMED-CT as reference standards. Our system was able to extract and analyze concepts for all documents we analyzed in the set of 59 patient education sheets sampled. In a parsimonious effort to do so, we were able to identify more than 117 ICD-9 codes and 741 SNOMED-CT codes to index the starter set of content analyzed. When projected back to the complete set of more than 1600 documents, we anticipate that we will be able to present this content from hundreds of relevant entries in the problem list, expanding the reach of this content without the need of extensive, time-intensive human-based annotation. We expect that combining the strengths of computer-based annotation and human review will allow for the best approach for combining strengths and saving time in annotating knowledge content for better access from tools like the infobutton.

References

1. Covell DG, Uman GC, Manning PR. Information needs in office practice: are they being met? Ann Intern Med. 1985;103(4):596–9.
2. Ely JW, Osheroff JA, Chambliss ML, Ebell MH, Rosenbaum ME. Answering physicians' clinical questions: obstacles and potential solutions. J Am Med Inform Assoc. 2005;12(2):217–24.
3. Hauser SE, Demner-Fushman D, Jacobs JL, Humphrey SM, Ford G, Thoma GR. Using wireless handheld computers to seek information at the point of care: an evaluation by clinicians. J Am Med Inform Assoc. 2007 Nov-Dec;14(6):807-15. Epub 2007 Aug 21.
4. Goldbach H, Chang AY, Kyer A, Ketshogileng D, Taylor L, et al. Evaluation of generic medical information accessed via mobile phones at the point of care in resource-limited settings. J Am Med Inform Assoc. 2014 Jan-Feb;21(1):37-42. doi: 10.1136/amiajnl-2012-001276. Epub 2013 Mar 27.
5. Cimino JJ, Elhanan G, Zeng Q. Supporting infobuttons with terminological knowledge. Proc AMIA Annu Fall Symp.1997;528-32.
6. Maviglia SM, Yoon CS, Bates DW, Kuperman G. KnowledgeLink: Impact of context-sensitive information retrieval on clinicians' information needs. J Am Med Inf Assoc. 2006;13:67-73.
7. Del Fiol G, Huser V, Strasberg HR, Maviglia SM, Curtis C, Cimino JJ. Implementations of the HL7 Context-Aware Knowledge Retrieval ("Infobutton") Standard: challenges, strengths, limitations, and uptake. J Biomed Inform. 2012 Aug;45(4):726-35. doi: 10.1016/j.jbi.2011.12.006. Epub 2012 Jan 2.
8. Strasberg HR, Del Fiol G, Cimino JJ. Terminology challenges implementing the HL7 context-aware knowledge retrieval ('Infobutton') standard. J Am Med Inform Assoc. 2013 Mar-Apr;20(2):218-23. doi: 10.1136/amiajnl-2012-001251. Epub 2012 Oct 16.
9. Infobutton Responder - OpenInfobutton Project. Available at: http://www.openinfobutton.org. Accessed on March 1, 2014.
10. Clayton PD, Narus SP, Huff SM, et al. Building a comprehensive clinical information system from components. The approach at Intermountain Health Care. Methods Inf Med 2003;42:1–7
11. Reichert JC, Glasgow M, Narus SP, Clayton PD. Using LOINC to link an EMR to the pertinent paragraph in a structured reference knowledge base. Proc AMIA Annu Fall Symp. 2002;:652-6.
12. Del Fiol G, Rocha RA, Clayton PD. Infobuttons at Intermountain Healthcare: utilization and infrastructure. Proc AMIA Annu Fall Symp. 2006;180-4.
13. Del Fiol G, Haug PJ, Cimino JJ, Narus SP, Norlin C, Mitchell JA. Effectiveness of topic-specific infobuttons: a randomized controlled trial. J Am Med Inform Assoc. 2008 Nov-Dec; 15(6): 752–759. doi: 10.1197/jamia.M2725
14. Hulse NC, Galland J, Borsato EP. Evolution in Clinical Knowledge Management Strategy at Intermountain Healthcare. AMIA Annu Symp Proc. 2012; 2012: 390–399. Published online 2012 November 3.
15. NCBO Bioportal Annotator. Available at: http://bioportal.bioontology.org/annotator. Accessed on March 1, 2014
16. Shah NH, Bhatia N, Jonquet C, Rubin D, Chiang AP, Musen MA: Comparison of concept recognizers for building the Open Biomedical Annotator. BMC Bioinformatics 2009, 10(Suppl 9):S14.

17. LePendu P, Iyer SV, Fairon C, Shah NH. Annotation Analysis for Testing Drug Safety Signals using Unstructured Clinical Notes. Journal of Biomedical Semantics 2012, 3(Suppl 1):S5 doi:10.1186/2041-1480-3-S1-S5
18. Apache Tika Project. Available at: https://tika.apache.org/. Accessed on March 1, 2014
19. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform. 2001 Oct;34(5):301-10.

Disease progression subtype discovery from longitudinal EMR data with a majority of missing values and unknown initial time points

Ilkka Huopaniemi, DSc[1], Girish Nadkarni, MD, MPH, CPH[1], Rajiv Nadukuru, MS[1], Vaneet Lotay, MS[1], Steve Ellis, BS[1], Omri Gottesman, MD[1], Erwin P Bottinger, MD[1]
1. Icahn School of Medicine at Mount Sinai, New York, USA

Abstract

Electronic medical records (EMR) contain a longitudinal collection of laboratory data that contains valuable phenotypic information on disease progression of a large collection of patients. These data can be potentially used in medical research or patient care; finding disease progression subtypes is a particularly important application. There are, however, two significant difficulties in utilizing this data for statistical analysis: (a) a large proportion of data is missing and (b) patients are in very different stages of disease progression and there are no well-defined start points of the time series. We present a Bayesian machine learning model that overcomes these difficulties. The method can use highly incomplete time-series measurement of varying lengths, it aligns together similar trajectories in different phases and is capable of finding consistent disease progression subtypes. We demonstrate the method on finding chronic kidney disease progression subtypes.

Introduction

Electronic medical records (EMR) increasingly provide comprehensive clinical data collected during routine clinical care encounters. EMR has a collection of longitudinal phenotypic data that potentially offer valuable information for discovering clinical population subtypes and using them further in association studies in medical research and even in prediction of outcomes in patient care. A number of clinical parameters and laboratory tests are collected as part of routine clinical care and their results are stored in the EMR or in data warehouses. The data warehouse represents a general patient population and the data can be used for statistical analyses. The common examples of routinely collected variables are systolic blood pressure (SBP), low-density lipoproteins (LDL), high-density lipoproteins (HDL), triglycerides, hemoglobin A1C (marker for diabetes and diabetes (blood glucose) control), and estimated glomerular filtration rate (eGFR; a marker of kidney function).

There is obvious interest towards discovering groups of similar patients with similar disease progression patterns in metabolic syndrome that involves varying accumulation of obesity, hypertension, hyperlipidemia, Type 2 diabetes, coronary artery disease and chronic kidney disease (CKD). Previous research has suggested[1] that using population subtypes in association studies instead of broad disease definitions can lead to superior results. Separating differential progression patters in the phenotypic variables can potentially discover these subpopulations. Especially with chronic and progressive diseases, the crucial difference between subtypes of a disease is often differential rates of progression, and any model attempting to find subtypes in progressive diseases must be able to account for this.

We use CKD as a case study in this paper. The prevalence of CKD ranges from 10% to 15% in the United States, Europe and Asia[2]. CKD is associated with increased mortality, decreased quality of life, and increased health care expenditure. CKD is defined in most cases clinically by loss of kidney function as estimated by a glomerular filtration rate (eGFR) below a threshold of 60 ml/min/1.72kg^2 (normal eGFR range 90 to 120 ml/min/1.73kg^2) and/or persistent increased urinary albumin excretion lasting more than 90 days[3]. Untreated CKD can result in end-stage renal disease (ESRD) and necessitate dialysis or kidney transplantation in 2% of cases. CKD is also a major independent risk factor for cardiovascular disease, all-cause mortality including cardiovascular mortality[6,7]. Approximately two thirds of CKD are attributable to diabetes (40% of CKD cases) and hypertension (28% of cases) [3]. However, CKD is also characterized by variable rates of progression with a significant proportion of patients having stable kidney function over time while some patients have rapid progression. These differential rates of progression[9] lead to clinically relevant, interesting subtypes among patient populations.

We aim to develop an unsupervised machine learning approach that takes longitudinal data of one variable from all patients and clusters them to population subtypes of which some are healthy and some turn out to be disease subtypes. The aim is to be able to include as many of the samples as possible in the analysis. Using the population subtypes as disease labels in association studies may be superior to the standard approaches of assigning disease labels from EMR data. We also hypothesize that using population subtypes and their temporal progression patterns may lead to improved performance in risk prediction. Most existing disease risk prediction models are coarse case/control models (do not account for subtypes) and use only snapshots of data without considering the temporal

patterns. Examples are Framingham risk score or the kidney disease progression model [4]. Even most of the advanced time-series models remain case/control models and do not attempt to discover population subtypes.

Electronic medical records are a messy, observational data source, as opposed to randomized controlled trials used in designed disease or drug studies. In the latter, data is collected at regular intervals under tight control of the investigators and disease onset times (first time points) are clearly recorded. In statistical analysis of EMR data, however, there are two major challenges: (a) Sparse data (large proportion of missing data) and (b) Unaligned nature of the longitudinal data. For instance, the Mount Sinai BioMe Biobank program has a longitudinal data collection from a period of 11 years and our aim is to use quarterly (every three months) median values of the laboratory measurements to reach a clinically relevant resolution. However, the number of years from which there are data from an individual patient varies greatly and only a minority of patients have a full coverage of data from 11 years; extremely few when quarterly values are sought (see Figure 1). When a large portion of the data is missing, imputation or removing samples or rows with missing data are not sensible options since we would end with a very small number of samples available. An even more difficult problem in modeling longitudinal EMR data is that there is no clear initial time point (t=0). Since patients have their first visit to a certain hospital at highly varying phases of progression of a disease, the first hospital visit with recorded data cannot be used as the initial time point. We have also concluded that using diagnostic criteria (such as the first eGFR<60 measurement in CKD) to fix the initial time point does not give adequate results in subtype modelling [data not shown]. Furthermore, many patients do not yet even have any major disease but it is desirable to include all patients in the analysis. Without the start point, standard clustering algorithms cannot be used since time points do not match between patients. Consequently, most studies in EMR are restricted to using only a single snapshot from the longitudinal data: usually the first or last time point.

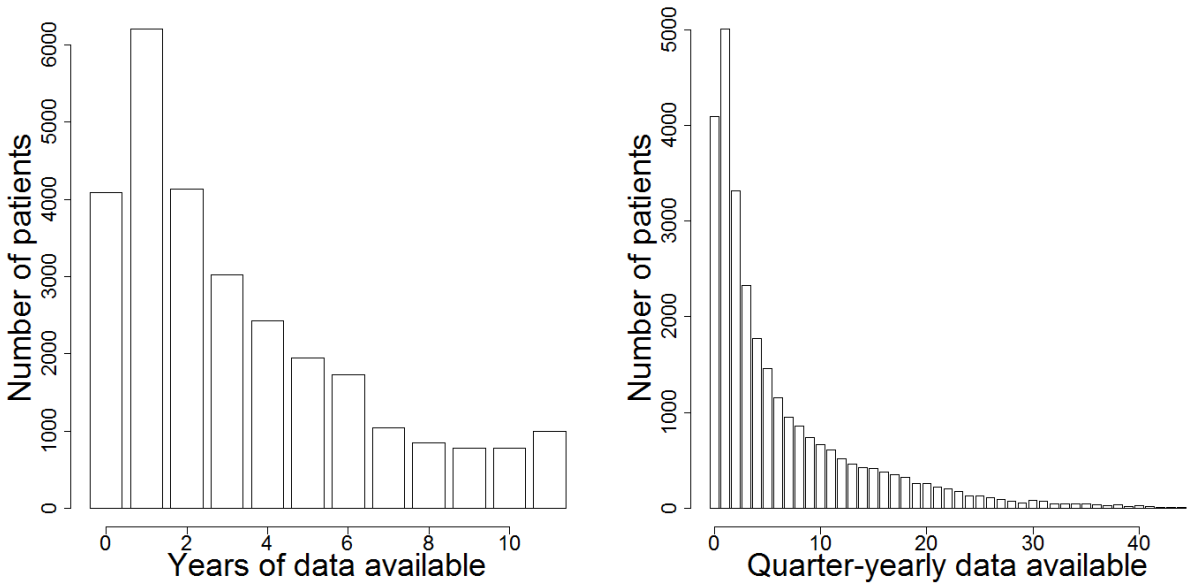


Figure 1. (Left) Most of the 27985 patients enrolled in Mount Sinai BioMe biobank have eGFR data available only from a small number of years out of a total of 11 years. The histogram shows from how many years patients have eGFR data available. (Right) Very few patients have a full coverage of 44 quarterly measurement of eGFR available. The histogram shows from how many quarterly time-points patients have eGFR data available. Multiple measurements from the same quarter-year have been converted into one median value.

In this paper, we present a Bayesian clustering and alignment model that is capable of identifying subpopulations of patients from a longitudinal dataset and overcoming the main challenges of sparse and unaligned data. The method aligns together time-series profiles in different phases of their disease progression in order to find clusters of similar progression patterns. Our generative latent variable formulation enables constructing a model that can use also samples with a large proportion of their time points missing. As a result, we can use a large proportion of the patients in the database in our modelling. A latent variable model is also a good approach for clustering short time-series, since different rates of progression can be separated easily.

One obvious purpose of clustering the longitudinal data collection in an EMR database is to visualize the progression patterns present in the entire patient population. Another application is using the cluster labels as traits in association studies with e.g. ICD9 codes, laboratory, medication or genomic data. We demonstrate that our method finds meaningful CKD progression subtypes and validate our model by showing that certain CKD-related ICD9 codes are much more common in certain disease clusters than in the rest of the patient population.

Data

The Mount Sinai BioMe Biobank Program has a collection of DNA and blood plasma from over 28000 patients linked to their full medical histories in the Mount Sinai electronic medical records database. In this paper, we use a collection of sparse longitudinal eGFR measurements from 2003-2013 from 27985 patients. We simplify the longitudinal dataset by binning each variable to quarter-yearly median values, which results in 44 time points. The eGFR has been estimated from measured serum creatinine. Though proteinuria is included in the KDIGO definition of CKD; in real-world practice, it is rarely collected and there is significant variability in the measurement tools. Also there are recent data indicating that neither microalbuminuria nor proteinuria is a significant predictor of decline in kidney function[5].

We have transformed the collection of all ICD9 diagnosis codes into a binary matrix that indicates whether patient has had a certain diagnosis code observed. Furthermore, an ICD9 is also considered observed if any more specific ICD9 code in the hierarchy has been observed (For example, 250 is considered observed if 250.1 or 250.03 has been observed. 250.1 is considered observed if 250.11 has been observed).

Methods

Clustering and alignment model

We have constructed a Bayesian generative model for the task of clustering and alignment of a longitudinal dataset. In this paper, we concentrate on one variable only. The combination of clustering and alignment of longitudinal data is an active machine learning research topic with many application areas[8]; here we consider the case of a large proportion of missing values and apply it to EMR disease progression data.

Clustering is a standard statistical method that partitions observations (patients) to sets of similar observations (clusters). This is accomplished by iterating between assigning the observations to clusters and updating the cluster centers. The number of clusters to be sought is defined *a priori* as a model parameter, but there are procedures for determining the optimal number of clusters. Clustering time-series data is a well-studied problem in the case where clear start points are known. As there are no well-defined start points in EMR data (first visit to the hospital is not a valid start point), we have to learn the start points (iteratively) from the data as well. Aligning the start point of each patient's trajectory in the cluster trajectory (cluster center) is an extra step in the iterative model. The start point parameter does not have an exact interpretation (such as disease onset), but it enables the alignment of the unaligned time-series so that coherent progression patterns can be found (Figure 2).

Each patient i ($i = 1{:}I$) comes with a data vector x_i of T time points so that the first element is the first visit to the clinic, and in general most elements are missing (Figure 1). In this paper, $T = 44$, $I = 10539$. The clustering model is essentially a multivariate mixture of Gaussians with two modifications. Firstly, as the data have missing values, cluster assignments of the samples (patients) are sampled such that the likelihood of the sparse time-series with respect to the corresponding cluster center trajectory is evaluated using only the time points with non-missing data. Secondly, the longitudinal data vectors need to be temporally aligned and we allow M different starting points in each cluster; as a result, each cluster center is of length $(T + M - 1)$, using $M = 20$. The alignment is done jointly with clustering by additionally evaluating the likelihood of the time-series in each possible start point in each cluster. The reason why we use a Bayesian generative model to tackle our problem is that when sampling the cluster assignments and alignments of time-series of varying lengths and with many missing time points, some of the time points of the cluster trajectories may not have any data currently assigned to them. In that case, priors determine the values of those cluster trajectory points.

By following the Bayesian formalism, we assume a generative model that has generated the observed data. The model can then be used to learn the model parameters from the data; the relevant model parameters here are cluster assignments k and learned start points m for each patients and the cluster trajectories (centers) θ_{kt} that can be viewed as average progression patterns.

The generative model is

$$x_{it} = N(\theta_{k(t+m-1)}, \sigma),$$
$$k \sim multinomial(\pi),$$
$$m \sim multinomial(\beta),$$
$$\pi \sim Dirichlet(\alpha),$$
$$\theta_{kt} \sim N(H, \sigma_2).$$

We thus assume that the observed data has been generated by the following mechanism: patient i comes from cluster k that is randomly chosen from a multinomial distribution of cluster weights π and the patient has the first visit to a hospital at phase m in the cluster trajectory, randomly chosen from a multinomial distribution of prior weights β. The data points in the time-series x_{it} are generated from a Gaussian distribution, where the cluster trajectory point $\theta_{k(t+m-1)}$ is the mean and σ is the standard deviation. Cluster weights π are determined by a Dirichlet distribution with a base measure α. The cluster centers θ_{kt} come from a Gaussian distribution with hyperpriors H and σ_2.

The σ is here a fixed parameter set to a tight value $\sigma = 1$ to get coherent clusters. We set H as the average of all eGFR measurements in the dataset. The $\sigma_2 = 30$ is set as a loose value to enable the modelling of a wide range of cluster trajectories, $\alpha = 1$. We set the first five and last five values of the prior weights of the alignments β to a low value and all the middle values to a uniform high value in order to improve the mixing in the sampling of the model (that trajectories would not get stuck in the beginning or end).

When a clustering configuration has been reached, the cluster assignments can be used for making inference of the data. The progression patterns can be visualized by plotting the data divided into clusters together with the alignments (Figure 2). Gibbs sampling was used for approximate inference (iteratively). It is straightforward to derive the Gibbs sampling equations from the generative model (see [19]). The method was implemented using the R statistical software. The analysis took 20 hours using a single Intel Core i7-2600 3.40GHz processor, but the computation can be made significantly faster by parallelization.

Validation of clusters by association studies

The population subtypes (cluster labels) are used in an association study where we ask whether a certain ICD9 disease diagnosis code is more common in a certain population subtype compared to the rest of the patients. We use Fisher's exact test and we run the association test between all disease subtype - ICD9 code pairs. When the association tests are run over 10000 ICD9 codes and 9 clusters, the Bonferroni multiple correction rate is $p = 10^{-7}$. Ordering the obtained p-value matrix by rows and columns gives information on what are the most distinctive subtypes and what are the most interesting disease diagnoses enriched in these subtypes. The maximum enrichment of selected relevant ICD9 codes can be used as a criterion for determining the optimal number of clusters. With $K = 9$, a 100% enrichment of ICD9 code 585 (Chronic kidney disease) was found in one cluster. The same statistical testing procedure is used to study the enrichment of males and self-reported ethnicities in the clusters.

Patient selection criterion

As patients have different numbers of data points available (Figure 1), we need a criterion for deciding which patients to include in the clustering analysis. It is clear that patients with zero or one eGFR measurements are not useful in finding longitudinal trajectories; patients with two or three measurements contain some information on the progression, but the measurements may be noisy and a large number of very short time-series may result in less coherent progression patterns. On the other hand, we aim to include as large a proportion of the available patients as possible in the analysis and the more stringent the selection criterion, the fewer patients fulfill it. We will compare the progression trajectories obtained by different selection criteria. The quantity to compare is the number of years from which patients have at least one data point available. The years do not need to be consecutive.

We construct a metric to evaluate the goodness of the learned trajectories. As differentiating disease progression rates between clusters is an important aspect of our modeling, we evaluate the difference of the eGFR slopes of individual trajectories compared to the slope of the cluster trajectory they have been assigned to. The slopes are calculated simply by fitting a regression line. Furthermore, as it turns out that some trajectories are non-linear (see Results section) and patients may have their available data from different parts of the non-linear trajectory, fitting a linear curve to a non-linear trajectory is not an optimal solution. We alleviate this problem by fitting a “local slope”, i.e. fitting the curve only to the part of the cluster trajectory from which the patient has data available and has been aligned to, and compare the individual slope to the local slope.

Results

We demonstrate our method on finding CKD progression subtypes from eGFR measurements. As can be seen in (Figure 1), only a small fraction of the total 27985 Biobank patients have eGFR data from the full period of 11 years and very few have a full coverage of 44 quarter-yearly measurements that would correspond to fully observed dataset (no missing values). As explained in the Introduction, even such full coverage data would not be readily usable since patients are in highly different phases of their disease progression and there are no clear start points. By using our Bayesian clustering and alignment approach, we can, however, use a significant portion of this heavily incomplete dataset.

Patient data criteria and evaluation

We now evaluate how many eGFR measurements are required for patients to be included in the clustering as a tradeoff between patient attrition and model accuracy. In (Table 1) we compare the criteria from how many years the patients need to have at least one measurement available (each year has been divided into four quarters). The number of available patients decreases with tighter criterion, with the benefit of better model accuracy. The slope error is the difference of the slope of an individual trajectory compared to the slope of the cluster trajectory. Please refer to the methods section for the definition of the accuracy of the model.

Table 1. Sample size and median error for different number of years

Selection criterion (Years)	2	3	4	5
Number of patients with data available	17672	13558	10539	8117
Median slope error	1.66	1.35	1.24	1.18

As can be seen from (Table 1), the number of patients with a sufficient amount of data available to meet the inclusion criterion drops rapidly when tightening the criterion. In the same time, the accuracy of the model increases, as there are a smaller number of short, potentially inaccurate time-series worsening the clustering result. We choose to include patients with eGFR measurement from at least 4 different years. Using this selection, we get very coherent progression subtypes yet have a large number of patients (10539) available.

We show in (Figure 2) the eGFR progression patterns for 9 clusters, representing the entire BioMe Biobank subcohort with at least 4 years of eGFR data. We have chosen 9 as the number of clusters as we have empirically observed it to be the minimum number that finds all the clinically meaningful main progression patterns and at least one cluster (C8, lowest eGFR values) with 100% enrichment of the ICD9 code 585 (Chronic kidney disease). As can be seen from the images, there is considerable noise in the data since eGFR measurements are inherently noisy and the trajectories from 10539 patients have been forced to 9 clusters. This noise could be reduced by using yearly medians instead of quarterly medians (with the cost of clinically important time resolution); even more coherent clusters could be sought by increasing the number of clusters.

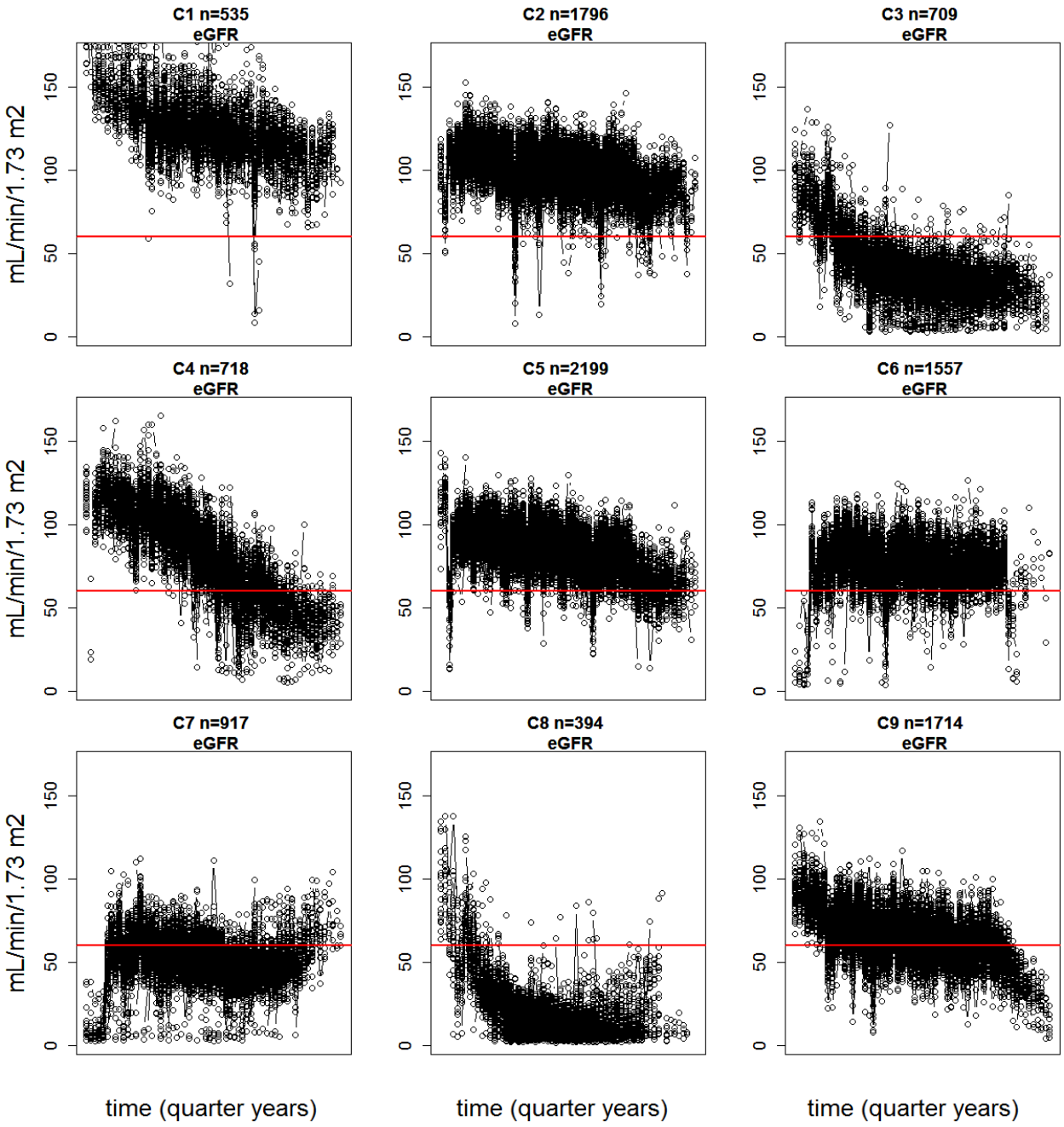


Figure 2. Many distinct coherent eGFR progression patterns can be found from the 10539 patients that represent the entire hospital cohort. The figure shows clustering and alignment results for eGFR using 9 clusters; each cluster in the figure consists of eGFR trajectories of all the patients in that cluster that have been aligned together. These trajectories have highly varying lengths (see Figure 1 and Table 2) and varying numbers of missing values. The time span corresponds to 16 years; each patient has data from 4-11 years (up to 44 quarter-yearly time points) and 20 possible start points are allowed. The n indicates the number of patients in each cluster; C indicates the cluster index.

In (Table 2) we demonstrate the median and interquartile range of the first and last time points of the eGFR of patients in each cluster, the mean duration (years) of data available, and the average slope of progression. Columns 4-7 show the values of the first and last points of the cluster trajectories (cluster centers) and the slope that has been fitted to the cluster trajectories. The values are in accordance with one another and with (Figure 2). Note that the median of the first values of the individual trajectories is different from the first point of the cluster trajectory since the patients in a cluster have their first time point (first visit to the clinic) at varying stages of the cluster trajectory (this also applies to the last time points). The accordance of the slopes of individual trajectories in a cluster with cluster trajectories is further visualized in (Figure 3).

Table 2. Summary of the eGFR progression patterns

	Mean (S.d.) years eGFR data	Median first eGFR [IQR]	Median last eGFR [IQR]	Average ΔeGFR per year	Cluster center first eGFR	Cluster center last eGFR	Cluster center ΔeGFR per year
C1	7.5(2.3)	127.6[18.1]	120.3[17.1]	-1.6	142.9	86.9	-2.9
C2	7.6(2.4)	105.7[13.4]	98.7[14.8]	-0.9	105.9	91.2	-1.7
C3	6.9 (2.5)	50.3[28.1]	33.2[15.7]	-3.4	78.1	31.4	-3.4
C4	6.4(2.6)	108.0[18.0]	76.9[38.3]	-5.2	118.2	43.2	-7
C5	6.9(2.6)	93.8[12.4]	83.9[15.6]	-1.5	94.5	63.7	-2
C6	6.5(2.6)	80.9[14.4]	77.5[12.6]	-0.6	79.6	73.4	-0.7
C7	6.9(2.7)	58.2[15.8]	50.3[12.4]	-0.8	50.5	47.8	-1.4
C8	6.5(2.4)	26.9[31.9]	9.9[8.6]	-3.6	51.8	13.7	-2.5
C9	6.9(2.64)	74.2[16.4]	62.4[13.4]	-2.1	89.2	41.2	-1.6

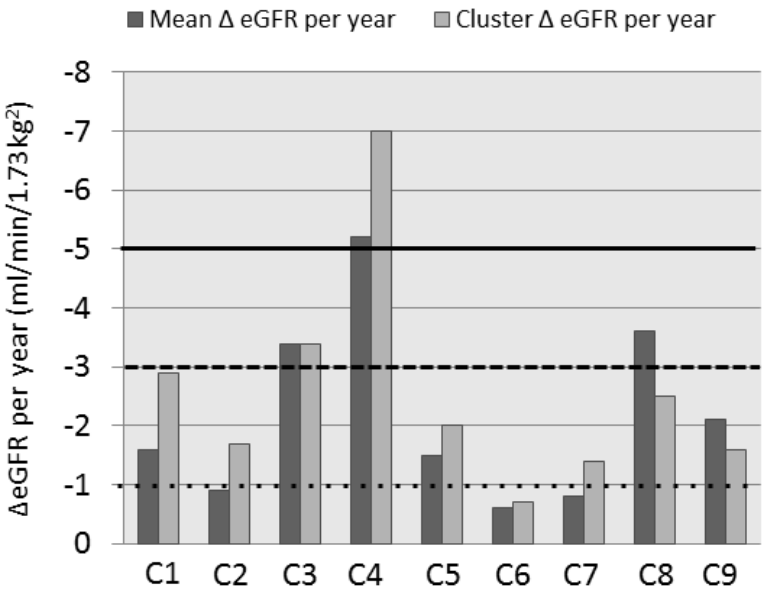


Figure 3. Bar graph of mean of eGFR change (ΔeGFR) per year (dark grey) and cluster center ΔeGFR (light grey) for patients in clusters C1 to C9. Lines indicate usual thresholds for nonprogression (dotted line), moderate progression (dashed line), and rapid progression (solid line).

In order to assess the clinical applicability and relevance of this clustering method, we hypothesized that demographic and disease patterns that were seen in longitudinal studies with similar patterns of eGFR progression would replicate independently in these clusters. In (Table 3) we show the mean and standard deviation of age, percentage of males and self-reported ethnicities (European ancestry (EA), African ancestry (AA), Hispanic/Latino (HL), Others) in each cluster. The star denotes clusters were the enrichment of a certain ancestry or gender was statistically significantly higher than for all the other patients using Fisher's exact test (Bonferroni rate $p = 10^{-7}$).

Each cluster had a statistically significantly different mean age compared to the patients in the other clusters using t-tests.

Table 3. Demographic characteristics of clusters, of all the patients in the analysis and all the patients in the biobank

	Age [Sd]	Males (%)	EA (%)	AA (%)	HL (%)	Other (%)
C1	36.9[10.7]*	32	5.5	59.3*	32.1	3.1
C2	50.1[10.6]*	36	13.1	34.3*	46	6.6
C3	71.7[12.2]*	40	24.4	28	42	5.6
C4	49.7[13.3]*	32	14.1	44.3*	34.1	7.5
C5	57.8[11.0]*	37	22	26.1	44.9	7
C6	62.5[11.6]*	40	28.6*	22.4	42.2	6.9
C7	70.3[11.7]*	40	26.3*	25.3	41.6	6.8
C8	62.9[13.9]*	52*	14.9	40.7*	36.4	8
C9	66.1[11.1]*	37	25.3*	25	42.8	6.9
All	59.1[14.4]	38	21	30.2	42.1	6.7
ALL in Biobank	53.7[17.2]	41	30.8	24.3	35.3	9.6

Table 4 shows the percentage of patients in each cluster with a diagnosis of selected ICD9 codes (or a more specific ICD9 code in the same hierarchy). The star denotes clusters were the enrichment of ICD9 codes is statistically significantly high compared to all patients in the other clusters (pooled). The Bonferroni multiple correction rate is $p = 10^{-7}$.

Table 4. Distribution of ICD9 codes among clusters, of all the patients in the analysis and all the patients in the biobank

		C1	C2	C3	C4	C5	C6	C7	C8	C9	All	Biobank
585.xx	CHRONIC KIDNEY DISEASE (CKD) (%)	1	2	89*	21	4	6	55*	100*	21	21	12
585.6	END STAGE RENAL DISEASE (%)	0	0	14*	2	0	1	7	77*	1	5	2
v45.1x	RENAL DIALYSIS STATUS (%)	0	0	6	1	0	0	3	64*	0	3	1
403.xx	HYPERTENSIVE CHRONIC KIDNEY DISEASE (%)	0	1	67*	14	2	3	35*	93*	13	15	8
401.xx	ESSENTIAL HYPERTENSION (%)	43	62	95*	63	70	72	88*	97*	81*	73	52
250.xx	DIABETES MELLITUS (%)	31	36	65*	45	39	38	55*	65*	44	43	28
410.xx	ACUTE MYOCARDIAL INFARCTION (%)	1	2	12*	4	3	4	8	19*	6	5	3
414.xx	CHRONIC ISCHEMIC HEART DISEASE (%)	5	12	53*	20	18	22	41*	68*	31	25	20
428.xx	HEART FAILURE (%)	8	9	41*	18	10	10	28*	60*	18	17	10
584.xx	ACUTE KIDNEY FAILURE (%)	3	5	58*	22	5	8	30*	67*	17	16	8
285.xx	OTHER AND UNSPECIFIED ANEMIAS (%)	43	38	73*	42	31	30	50	92*	41	42	24

These demographics and ICD9 codes present an independent clinical validation of the relevance and applicability for the clustering patterns. For example; cluster 1 represents a group of patients that start at a high eGFR with the median eGFR being more than 120 ml/min/1.73m^2. Clinically, this represents a group of patients who have glomerular hyperfiltration (a precursor to developing kidney injury with elevated eGFR above 120 ml/min/1.73m^2) which usually happens in younger patients who are usually African-American and occurs in the very early stages of diabetes mellitus and hypertension and thus might not have a confirmed diagnosis of them[10,11,12]. As demonstrated in (Table 3 and 4); patients in cluster 1 are significantly younger than those in other clusters with a mean age of 36.9 years and have a lower prevalence of diabetes mellitus and hypertension as compared to the other clusters.

Clusters 3 and 8 provide more evidence for this validation. As shown in (Figure 2), these are clusters where patients starting from a CKD stage 3/4 with a mean eGFR of 50 and 27 ml/min/1.73m^2 progress rapidly to a low eGFR (mean eGFR of 33 and 10 ml/min/1.73m^2 respectively). These clusters have the highest prevalence of an ICD9 code for acute kidney injury (AKI), heart failure and anemia amongst the clusters. As shown in multiple studies, AKI[13,14], heart failure and anemia[15] are very significant risk factors for both CKD progression and end stage renal disease (ESRD) development. This is further validated within these clusters since cluster 8 that has a higher prevalence of acute kidney injury, heart failure and anemia compared to cluster 3, also has a higher proportion of ESRD and dialysis and a lower final eGFR. Cluster 2 is an example of healthy patients with normal eGFR and they do not have many CKD diagnoses.

Thus we demonstrate that this automated machine learning approach organizes sparse and non-aligned data into coherent and clinically meaningful subtypes based on disease progression and this finds further independent validation after comparing demographics and ICD9 code enrichment.

Conclusions

We have demonstrated the use of clustering and alignment modelling for finding disease progression subtypes from highly incomplete EMR laboratory data. We have shown that using this type of modelling, we can use a large portion of a longitudinal dataset that has irregular time series of varying lengths and a high proportion of missing data. In particular, we have shown how to deal with the fact that there are no clear initial time points in the time-series; the solution is to align similar trajectories together. Our method was successful in finding from the data meaningful CKD progression patterns that correspond to known disease subtypes and stages.

The generative Bayesian modelling formalism is a flexible approach that allows for the construction of models that take into account all the necessary aspects of the modelling problem. In our case, clustering longitudinal data, alignment and dealing with missing data could all be done within a single unified model. We also successfully validated our clusters by association studies between the clusters, demographics and ICD9 diagnosis codes.

There are many potential applications for this approach. For instance, although novel genetic associations with eGFR have been reported, there are other potential genetic associations that explain the differential rates of CKD in different ethnic populations[16,17]. However most genetic association studies are cross-sectional in nature and longitudinal studies require the resources of clinical cohorts. This clustering approach could be applied to evaluating genetic associations with longitudinal disease progression especially in institutions which have EMR linked biobanks. This is of special importance with national consortia such as the Electronic Medical Records and Genomics (eMERGE) Network, a NHGRI funded consortium tasked with developing methods and best-practices for the utilization of the Electronic Medical Record (EMR) as a tool for genomic research[18]. Also, since this approach can be deployed at multiple sites with EMR, a large number of patients can be used for modeling purposes that would not be possible in conventional longitudinal cohort studies.

In this paper, we considered the clustering of only one longitudinal variable, however, our model can be directly used for multiple variables. One can, for instance, cluster and align longitudinal eGFR, SBP and hemoglobin AIC data together in order to find clusters with similar progression in multiple variables. Adding more variables and increasing the number of clusters in the analysis can lead to discovering ever more specific clinical subtypes, critical in the future direction of personalized treatment decision support. Finally, though we used CKD as an example the opportunities for examining distinct disease progression subtypes and making innovative discoveries are endless in any disease area depending on available data in the EMR

Acknowledgements

The eMERGE Network is funded by NHGRI, with additional funding from NIGMS through the following grants: U01HG004438 to Johns Hopkins University; U01HG004424 to The Broad Institute; U01HG004438 to CIDR; U01HG004610 and U01HG006375 to Group Health Cooperative; U01HG004608 to Marshfield Clinic; U01HG006389 to Essentia Institute of Rural Health; U01HG04599 and U01HG006379 to Mayo Clinic; U01HG004609 and U01HG006388 to Northwestern University; U01HG04603 and U01HG006378 to Vanderbilt University; U01HG006385 to the Coordinating Center; U01HG006382 to Geisinger Clinic; U01HG006380 to Icahn School of Medicine at Mount Sinai; U01HG006830 to The Children's Hospital of Philadelphia; and U01HG006828 to Cincinnati Children's Hospital and Boston Children's Hospital.

References

1. Warde-Farley D, Brudno M, Morris Q, Goldenberg A. Mixture model for sub-phenotyping in GWAS. Pacific Symposium on Biocomputing .2012; 17:363-374.

2. Hallan SI, Matsushita K, Sang Y, Mahmoodi BK, Black C, Ishani A, Kleefstra N, Naimark D, Roderick P, Tonelli M, Wetzels JF, Astor BC, Gansevoort RT, Levin A, Wen CP, Coresh J; Chronic Kidney Disease Prognosis Consortium. Age and association of kidney measures with mortality and end-stage renal disease. JAMA. 2012 Dec 12; 308(22): 2349-60.

3. Kidney Disease: Improving Global Outcomes (KDIGO) CKD Work Group. KDIGO 2012 Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease. Kidney inter., Suppl. 2013; 3: 1-150.

4. Tangri N, Stevens LA, Griffith J, Tighiouart H, Djurdjev O, Naimark D, Levin A, Levey AS. A predictive model for progression of chronic kidney disease to kidney failure. JAMA. 2011 Apr 20; 305(15): 1553-9.

5. Perkins BA, Ficociello LH, Roshan B, Warram JH, Krolewski AS. In patients with type 1 diabetes and new-onset microalbuminuria the development of advanced chronic kidney disease may not require progression to proteinuria. Kidney Int. 2010 Jan; 77(1): 57-64.

6. Chronic Kidney Disease Prognosis Consortium, Matsushita K, van der Velde M, Astor BC, Woodward M, Levey AS, de Jong PE, Coresh J, Gansevoort RT. Association of estimated glomerular filtration rate and albuminuria with all-cause and cardiovascular mortality in general population cohorts: a collaborative meta-analysis. Lancet. 2010 Jun 12; 375(9731): 2073-81.

7. Gansevoort RT, Matsushita K, van der Velde M, Astor BC, Woodward M, Levey AS, de Jong PE, Coresh J; Chronic Kidney Disease Prognosis Consortium. Lower estimated GFR and higher albuminuria are associated with adverse kidney outcomes. A collaborative meta-analysis of general and high-risk population cohorts. Kidney Int. 2011 Jul; 80(1): 93-104.

8. Mattar M, Hanson A, Learned-Miller E. Unsupervised joint alignment and clustering using Bayesian Nonparametrics. Conference on Uncertainty in Artificial Intelligence (UAI) 2012.

9. Li L, Astor BC, Lewis J, Hu B, Appel LJ, Lipkowitz MS, Toto RD, Wang X, Wright JT Jr, Greene TH. Longitudinal progression trajectory of GFR among patients with CKD. Am J Kidney Dis. 2012 Apr; 59(4): 504-12.

10. Palatini P. Glomerular hyperfiltration: a marker of early renal damage in pre-diabetes and pre-hypertension. Nephrol Dial Transplant. 2012 May; 27(5): 1708-14.

11. Chaiken RL, Eckert-Norton M, Bard M, Banerji MA, Palmisano J, Sachimechi I, Lebovitz HE. Hyperfiltration in African-American patients with type 2 diabetes. Cross-sectional and longitudinal data. Diabetes Care. 1998 Dec; 21(12): 2129-34.

12. Palatini P, Mormino P, Dorigatti F, Santonastaso M, Mos L, De Toni R, Winnicki M, Dal Follo M, Biasion T, Garavelli G, Pessina AC; HARVEST Study Group. Glomerular hyperfiltration predicts the development of microalbuminuria in stage 1 hypertension: the HARVEST. Kidney Int. 2006 Aug; 70(3): 578-84.

13. James MT, Ghali WA, Knudtson ML, Ravani P, Tonelli M, Faris P, Pannu N, Manns BJ, Klarenbach SW, Hemmelgarn BR: Associations between acute kidney injury and cardiovascular and renal outcomes after coronary angiography. Circulation 123: 409–416, 2011.

14. Parikh CR, Coca SG, Wang Y, Masoudi FA, Krumholz HM. Long-term prognosis of acute kidney injury after acute myocardial infarction. Arch Intern Med 168: 987–995, 2008.

15. Virani SA, Khosla A, Levin A. Chronic kidney disease, heart failure and anemia. Can J Cardiol. 2008 Jul; 24 Suppl B: 22B-4B; Rossert J, Froissart M. Role of anemia in progression of chronic kidney disease. Semin Nephrol. 2006 Jul; 26(4): 283-9.

16. Köttgen A, Glazer NL, Dehghan A, Hwang SJ, Katz R, et al. (2009) Multiple loci associated with indices of renal function and chronic kidney disease. Nat Genet 41: 712–717.

17. Parsa A, Kao WH, Xie D, Astor BC, Li M, Hsu CY, Feldman HI, Parekh RS, Kusek JW, Greene TH, Fink JC, Anderson AH, Choi MJ, Wright JT Jr, Lash JP, Freedman BI, Ojo A, Winkler CA, Raj DS, Kopp JB, He J, Jensvold NG, Tao K, Lipkowitz MS, Appel LJ; the AASK and CRIC Study Investigators. APOL1 Risk Variants, Race, and Progression of Chronic Kidney Disease. N Engl J Med. 2013 Nov 9.

18. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, Sanderson SC, Kannry J, Zinberg R, Basford MA, Brilliant M, Carey DJ, Chisholm RL, Chute CG, Connolly JJ, Crosslin D, Denny JC, Gallego CJ, Haines JL, Hakonarson H, Harley J, Jarvik GP, Kohane I, Kullo IJ, Larson EB, McCarty C, Ritchie MD, Roden DM, Smith ME, Böttinger EP, Williams MS; eMERGE Network. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet Med. 2013 Oct; 15(10): 761-71.

19. Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian Data Analysis (2nd edition). Chapman & Hall/CRC, Boca Raton, FL, 2003.

Piloting a Deceased Subject Integrated Data Repository and Protecting Privacy of Relatives

Vojtech Huser MD, PhD[1], Mehmet Kayaalp MD, PhD[2], Zeyno A. Dodd PhD[2], James J. Cimino, MD[1,2]

[1]Laboratory for Informatics Development; National Institutes of Health, Clinical Center
[2]National Library of Medicine, Bethesda, MD, USA

Abstract

Use of deceased subject Electronic Health Records can be an important piloting platform for informatics or biomedical research. Existing legal framework allows such research under less strict de-identification criteria; however, privacy of non-decedent must be protected. We report on creation of the decease subject Integrated Data Repository (dsIDR) at National Institutes of Health, Clinical Center and a pilot methodology to remove secondary protected health information or identifiable information (secondary PxI; information about persons other than the primary patient). We characterize available structured coded data in dsIDR and report the estimated frequencies of secondary PxI, ranging from 12.9% (sensitive token presence) to 1.1% (using stricter criteria). Federating decedent EHR data from multiple institutions can address sample size limitations and our pilot study provides lessons learned and methodology that can be adopted by other institutions.

Introduction

The use of electronic health records (EHR) for research is becoming commonplace. Patient privacy and ethical oversight by institutional review boards (IRBs) present practical and administrative challenges for scientists who need to access Big Data.[1] De-identification (elimination of the protected sections of health information) can allow EHR records to be considered "not human research subject data", thus reducing regulation.[2] Another approach is to restrict research to the use of deceased patients' records.[3] Although data analyses that use only deceased patient records may have inadequate power due to small sample size, they can provide pilot results whether particular EHR-based data extraction or analysis is feasible. Data analyses in multi centric clinical studies that use data inputs from varying EHR software vendors or varying user groups may benefit from an administratively simpler pilot experiment first. Such a pilot study on deceased subjects' records can prevent situations in which an investigator tediously obtains IRB approval for a multi-site study that uses more complete record sets (such as de-identified or even identified data) only to discover that the records lack the data necessary for the study.

In a previous perspective paper,[3] we outlined several challenges to establishing a *deceased subject integrated data repository* (dsIDR), such as proving death status, involvement of next of kin (surviving spouse or children), possible deletion of deceased patient data by the organization (to limit the data warehouse size after the minimum record-keeping period expired), and patients' perceptions of the value of dsIDRs. Creation of dsIDRs using only terminology-coded information, such as structured medication, laboratory results and diagnostic or procedural data that do not require de-identification is relatively easy to accomplish. However, additional challenges are encountered when a dsIDR contains narrative text clinical documents. In such a repository, personal identifiers defined in the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule (Safe Harbor Method) as well as "personally identifiable information" (PII) defined in the Privacy Act must be detected and redacted properly.

One particular challenge with narrative text document inclusion in the dsIDR, identified in our prior analysis, was handling the PHI of people other than the patient. Because research use of private health data of living and deceased is regulated differently, we need to distinguish two types of PHI/PII: *primary* (about the primary patient in question) and *secondary* (about individuals other than the primary patient, such as relatives or caregivers). We use the term *secondary PxI* to refer to both secondary PHI and secondary PII. We also use the term PHI to denote removal of the 18 specified identifiers (e.g., primary or secondary phone number), rather than removal of health data. The regulatory provisions for decedent EHR research do not mandate specific de-identification method (as opposed to the clear definition of 18 PHI identifiers in the Safe Harbor Method). The relaxed approach to de-identification in a dsIDR does not extend, however, to secondary PxI of living relatives, which must be entirely removed. This removal may be more difficult since the typical approach of using the primary patient's name may not be helpful in identifying secondary names. The process is further complicated because primary PHI that would ordinarily be allowed in a dsIDR must be removed if it helps reveal identity of secondary persons mentioned in the record (e.g., if a primary patient last name remains in the record, it may also provide a last name for a living relative identified in

the EHR only by a first name). Based on our own legal analysis, we hypothesize that mention of dates and locations (e.g., city or town) that would have to be scrubbed in traditional de-identification for whole population IDR, may stay in dsIDR.

In this paper, we report on efforts to establish a dsIDR at the National Institutes of Health's Clinical Center (NIH CC) and our investigation of the rate of secondary PxI in deceased patients' EHR.

Background

Institutional context

The NIH Clinical Center is a 240-bed hospital dedicated to research. Since 2008, it has established an integrated data repository called the Biomedical Translational Research Information System (BTRIS). BTRIS contains data on over 500,000 research subjects seen at the NIH CC since 1976. BTRIS contains data from the current EHR (Sunrise Clinical Manager v5.5, Allscripts, Chicago, IL), a prior EHR (Medical Information System, Technicon Data Systems, Tarrytown, NY) and several ancillary clinical and research systems.[4] BTRIS offers two modes of data export: identified data extracts available to study investigators and de-identified data extracts available to any NIH investigator. Currently, de-identified data exports are limited to structured coded data and a limited number of narrative text documents (e.g., pathology reports).

Design Principles of NIH CC dsIDR (dsBTRIS)

With our dsIDR dataset and framework, we hope to offer interested researchers yet another platform for generating hypotheses and piloting informatics methods. We also hope to learn important lessons about dsIDR design and actual use from creating a single institution dsIDR. These findings can inform development of a federated dsIDR of deceased patient EHR data from multiple institutions. The design principles for our dsIDR include the following elements:

- Access is limited to researchers who obtain approval (with exemption from IRB review) from the NIH Office of Human Subject Research Protection (OHSRP)
- Researchers must certify that they will not try to identify deceased subjects or other persons in the provided data
- The EHR of deceased patients undergoes the following transformations
    - Structured data: direct patient identifiers (e.g., structured name, address or employer field) are removed since they have none or very limited value to research
    - Unstructured data: The current dsIDR implementation pilot does not contain any unstructured data; selection of the best de-identification approach is one of the goals of this study; we plan to remove primary patient name from narrative text documents (e.g., progress note and laboratory text comments); dates and locations (e.g., "Suburban Hospital") would remain in the documents; secondary PxI would also be removed

Legal framework for deceased subject research

A legal framework for use of EHR data post mortem has been explored in greater detail in our earlier perspective paper.[3] The most important legal factor is that deceased subject research does not require traditional consent and full IRB review. Decedent research is listed in Title 45 of the Code of Federal Regulation under section §164.512, titled "Uses and disclosures for which an authorization or opportunity to agree or object is not required" [45 CFR 164.512(i)(1)(iii)].[5] The enabling statue for this regulation is the Health Insurance Portability and Accountability Act (HIPAA; Public Law 104-191).

The regulation allows disclosure of PHI of decedents without the need to seek full IRB approval and has to adhere to the following conditions:[6]

(1) use is sought solely for research on the PHI of decedents;
(2) the researcher can provide, on request, documentation of the death for individuals used in the study; and
(3) the PHI is necessary for the research.

Throughout this article, we refer to this HIPAA provision as "decedent research clause".

When information is private

Of note is also the fact that health related information may become secondary PHI only after it is combined with a clear designation of a secondary person (such as full name or phone number). If a record contains the sentence "Her mother and daughter both have symptoms of multiple sclerosis", it only becomes secondary PHI when associated with a *living* person's full name. How unique and identifying is the context information provided about the

secondary person is of significance (quasi-identifiers). This masking phenomenon is well described by the concept of k-anonymity (each dataset record is indistinguishable from at least k-1 other records given a group of identifying attributes) and the concept of l-diversity (each group of identifying attributes is immune to probabilistic inference attack and has at least l well represented value).[7] There is no clear established boundary; however, the HIPAA ZIP code rules offer one potential precedence: HIPAA zip code rule (45 CFR 164.514) permits revealing 3-digit ZIP codes as long as the 3-digit ZIP code covers an area populated by more than 20,000 people, as this is considered to be sufficient "masking" of the individual. The masking principle is important in redacting or preserving a sentence, such as, "Patient has a 9-year-old daughter" in a document that otherwise contains unmodified dates and locations, but does not contain the primary patient name.

Fall back policy

In addition to the dsIDR use policy that forbids subject re-identification, NIH CC has a general institutional policy for situations when identifiers are discovered in a dataset that has been thought to be de-identified. The existing NIH CC process generating de-identified data starts with production of an *initial de-identified data set*. An NIH employee not affiliated with the research (an "honest broker") reviews any BTRIS-derived dataset and determines whether it is free of PHI. If the initial data set is certified as sufficient, it is provided to the researcher. If PHI is found, the honest broker creates additional rules and data transformations to produce a *revised de-identified data set*. Ultimately, the researcher receives the initial or revised de-identified data. If the researcher encounters PHI that were missed by the honest broker, NIH OHSRP policy requires him to discontinue the use of the data and obtain IRB approval (typically, with waiver of consent) in order to continue the research. This process has two weak points. First, the review by the honest broker is lengthy, labor intensive, and has to be repeated for each project. Second, the researcher runs the risk of project interruptions and the requirement for additional review if PHI is discovered during the data analysis. The goal of the dsIDR is to produce a resource that does would not require additional human review and could be re-used for multiple projects.

Methods

Establishing the dsIDR

We used existing BTRIS death status data to establish the size of the dsIDR and explored the proportion of structured and narrative-text data available on deceased subjects. We looked at how many data rows are excluded by existing PHI filters (developed for the purpose of current de-identified BTRIS exports). We compared the size of the dsIDR with the full BTRIS repository in terms of number of patients, laboratory results, medication data, diagnoses and clinical documents.

Analysis of narrative text clinical documents

We used National Library of Medicine's (NLM) Scrubber software to parse dsIDR narrative documents.[8] Due to data size, requirements for computing time, and focus on recent and active documents, we limited our analysis to documents authored between October 1st, 2011 and October 1st, 2013). We also limited the analysis to the most frequent documents that accounted for 99% of all documents, ignoring rarely used document types. We obtained ethical approval for our deceased records research from NIH OHSRP office.

In order to correctly detect and remove sensitive patient information, we define the term *a priori PHI* as patient identifiers that have limited or no research value, that are known prior to the initiation of the PHI scrubbing process, and that definitely have be removed from clinical documents. In other words, our term "a priori PHI" refers to primary patient identifiers that are known from patient demographic data entered in structured form within an EHR. A priori PHI elements in our study were first, middle and last name and medical records number (MRN) of the primary patient. Of note is that patient identifiers do not equate Protected Health Information. Patient identifiers are Personally Identifiable Information, which meet the definition of PHI for living patients but not technically PHI for deceased patients.

Our work also relies on the assumption that any document containing secondary private information must indicate by at least one token who that person is (e.g., husband). We refer to this as the "token presence assumption". We defined a set of secondary person tokens that may indicate presence of secondary PxI. We used a regular expression search with the following word elements (including plural versions, combinations with applicable prefixes and suffixes, and other variations): mother, father, spouse, parent, wife, husband, daughter, son, sister, brother, children, child, mom, mommy, mama, dad, daddy, papa, sibling, aunt, uncle, "grand-*", "*-in-law", "in[-]law", friend, supervisor, employer, employee, girlfriend, boyfriend, roommate, partner, boss, collaborator. We sought to discover the frequency of secondary person tokens in clinical documents.

In addition to general occurrence, we used the document type (e.g., discharge summary vs. pathology report) as the key document parameter to investigate the occurrence of secondary PxI. We assumed that patterns of secondary PxI occurrence would be similar within a given document type; hence, our analysis is structured by document type. However, we also understand that this assumption may not be fully relied upon, as users may differ in their choice of document type and may differ in how they use those types. For each document type, we measured the frequency of occurrence (in dsIDR), and the frequency with which they contained a priori PHI, any personal name or a secondary person token.

We used NLM Scrubber to detect human names, alphanumeric identifiers (IDs) and addresses. Personal names are detected by multiple methods with the most discriminative method being the computation of the likelihood ratio of name to non-name of a given word. To identify sensitive alphanumeric identifiers, NLM Scrubber uses a two phase process where it first tries to identify laboratory values (that should be preserved) based on a set of hard coded patterns and in the subsequent second phase marks the remaining alphanumeric strings as sensitive tokens. Addresses are recognized mostly via "shapes" component of a specialized part-of-speech tagger (called dTagger) that looks for address-like patterns. For each analyzed document, NLM Scrubber outputs a set of tagged tokens classified by the identified type (e.g., address or personal name) appended to the end of the document or stored within a database. Marked up documents can also be browsed and edited in an associated editor called Visual Tagging Tool (VTT). Because of our focus on secondary PxI, we did not use other features of NLM Scrubber such as de-identification of ages (over age 89) and dates. On the other hand, we added new features, such as the ability to detect a set of secondary person tokens and the ability to analyze sensitive tokens that are in proximity to a secondary person token (e.g., address near "daughter"). We used NLM Scrubber's token-level output to detect frequency of various identifiers by tag. A full description of NLM Scrubber is available in a prior published study on name detection[8] and in an internal NLM report (non-name tokens).[9]

Manual review of secondary PxI instances

In addition to software-assisted analysis, we employed manual review of a subset of clinical documents to improve our understanding of different types of secondary PxI, to improve our definition of a true positive secondary PxI, and to inform our final scrubbing methodology. Due to a large volume of documents, we picked a single frequent secondary person token and a single frequent document type. We then reviewed all those document instances in a single calendar year (2013) and classified the secondary PxI into classes that could be later used in PxI removal efforts.

Table 1: Size of structured EHR data (full and decease subject repository)

Parameter	Full repository (BTRIS)	Deceased subject repository (dsBTRIS)
# patients (with laboratory data)	264,885	23,962 (9.1%)
# laboratory tests results	213.3 M	52.0 M (24.4%)
# medication administrations	16.5 M	5.2 M (31.5%)
# diagnoses	1.9 M	0.3M (15.8%)
# clinical documents	21.3M	5.6 M (26.3%)

Results

Our analysis resulted in the creation of a pilot deceased subject data repository. NIH researchers interested in using the dsIDR can seek ethics approval using an NIH OHSRP form that has pre-filled elements applicable to dsIDR research projects. This form (with pre-filled elements indicating the use of dsIDR platform) is available as an online appendix A at http://dx.doi.org/10.6084/m9.figshare.924793). This form is relatively short and dsIDR projects require less detailed description compared to a full IRB submission.

Size of dsIDR dataset

As of March 2014, the NIH BTRIS data repository contained coded data on 23 962 deceased patients. Table 1 shows side by side comparison of the full BTRIS repository versus deceased subject subset (numbers reflect laboratory test results and medication administrations, not orders). In comparing the number of patients (first row in table 1), we only considered patients with at least one laboratory result due to a significant number of patients that are administratively entered but lack a critical number of clinical diagnostic or treatment data (e.g., trial screening or patient's relatives). Existing BTRIS de-identification filters excluded 1.4M laboratory test results and 45 644 medication records where associated comments include PHI (patient name or MRN).

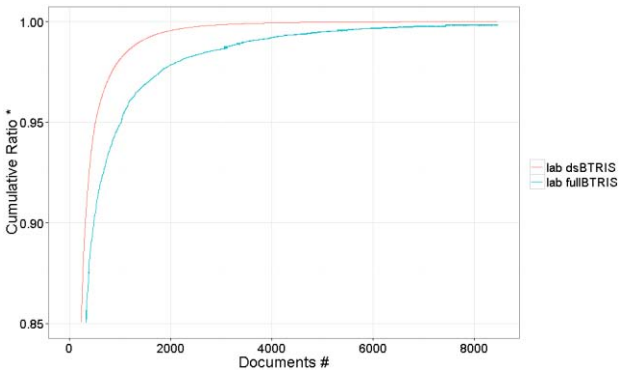


Figure 1: Pareto chart of lab tests for BTRIS and the dsBTRIS. # 8470 dsBTRIS laboratory types; 5000 full BTRIS test types (0.14% of all test instances) are not shown. * Cumulative portion of the respective repository that is accounted for by the test terms on the horizontal axis; the fullBTRIS line stops at 99.86% of all tests covered by all 8470 dsBTRIS tests).

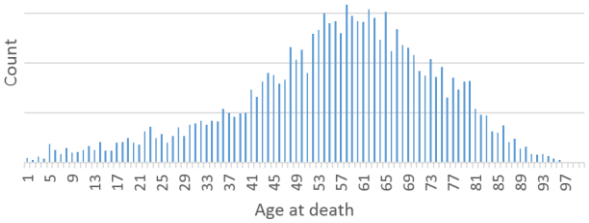


Figure 2: Patient distribution by Age at death (years 2009-2013)

(vertical axis legend (count) is not shown to mask exact counts)

A sub-analysis of laboratory data showed that whereas BTRIS contains 13,470 distinct laboratory test types (e.g., plasma folic acid level), dsBTRIS has 8,470 types. However, looking at the most frequent tests[10], we found that dsIDR contained 99.85% (1997 test types) of the 2000 most frequent test types (see Figure 1 for cumulative percentage graph).

To help researchers seeking recent data within a particular age group, we produced histograms of age of death for different time periods. An example, for deaths occurring during 2009-2013, is shown in Figure 2.

Analysis of narrative text clinical documents (by document type)

Within the documents from the last 2 years, we identified over 379 thousand clinical documents in deceased patients that were of 391 distinct document types. Excluding 1% of rare documents, the number of distinct document types reduced to 177 document types (376 051 documents). Each analyzed document type in the final set had at least 80 instances. A small number of document types capture the majority of the data. We found that 28 document types accounted for 80% of all documents and 59 document types accounted for 90% of all documents. An overview of all document types is available as appendix B at http://dx.doi.org/10.6084/m9.figshare.924793.

NLM Scrubber took 105 hours running in 4 parallel processes on a Linux workstation (8 CPU, 16GB RAM) with most computing time spent on part-of-speech tagging. Results were stored in a MySQL database to allow subanalysis by document type, tag or token. NLM Scrubber parsed 112 million tokens overall and marked 1.9 million as sensitive, with 53.4% being tagged as alphanumeric ID, 42.6% as personal name, 2.3% as address, and the remaining 1.7% as a combination of multiple tags. On average, a document contained 5.3 tagged tokens (ranging from 0 to 324) with 5.3% of documents containing no sensitive tokens. No list of local provider names was provided to NLM Scrubber.

Presence of a priori PHI: The presence of a priori PHI discovered by the NLM Scrubber varied widely by document type, from 0% to 100%. Table 2 presents document types that contain a priori PHI in more than 50% of document instances. Six document types (e.g., Speech Language Pathology Document) contain a priori PHI in more than 75% of all instances. The table shows that a priori PHI occurrence also differs by user group (e.g., National Institute of Allergy and Infectious Disease [NIAID] Intramural Research Program). A priori PHI presence data on all document types are available in the online appendix B.

On the other hand, we found 17 document types (9748 documents in total) that neither contained a priori PHI nor secondary person tokens. Analysis by document type revealed that 9 of these 17 document types always contain the same text telling the clinician to refer to a second document. For example, "Ophthalmology Consult (National Eye Institute)" document type consisted of 350 documents with the text: "NEI Clinical Documentation - Please click the camera icon in the bottom left corner of the window to see this document". Six document types were very brief forms related to admission, blood transfusion or anesthesia (e.g., Pre-Operative Anesthesia Record, Anesthesia Perioperative Record, and Post Anesthesia Care Unit Record), and often consisted of administrative content. Such "medium value" documents often repeated patient data such as research protocol assignment or patient gender. Only one document type ("Point of Care Testing Document") contained more variable clinician-entered data, such as glucose bed side test results.

Table 2: Document types that frequently contain a priori PHI

Document Type	Total Count	% of instances with a priori PHI	% of instances with secondary person token
NIAID Progress Note - Study Coordinator Document (CC, CRIS)	154	100%	15%
Speech Language Pathology Document (CC, CRIS)	98	99%	53%
Transfusion Medicine Note Document (CC, CRIS)	88	89%	27%
Dermatology Consult (CC, CRIS)	237	86%	39%
SOAP Progress Note Document (CC, CRIS)	1052	83%	28%
Genitourinary (GU) Oncology Free Text (CC, CRIS)	230	80%	24%
Bereavement Intake Progress Note (CC, CRIS)	121	79%	99%
Radiation Oncology HPE Document (CC, CRIS)	123	78%	96%
Radiation Therapy Summary Document (CC, CRIS)	112	75%	21%

Pooling all documents (regardless of the document type), we found that 4.61% of clinical documents (17 340 out of 376 051) contained a priori PHI.

Presence of a secondary person tokens: Table 3 presents the frequency of secondary person tokens across all documents identified by NLM Scrubber. The most frequent secondary person tokens were: mother, spouse, parent, wife and father. Secondary tokens with less than 0.5% relative occurrence were collapsed into two bottom items "other-family" and "other-non-family" to reduce the size of the table. Data on all tokens are available in a separate sheet of online appendix B.

When analyzed by document type, occurrence of secondary person tokens ranges from 0% to 99%. For example, in "Bereavement Intake Progress Note" the occurrence was 99%, while the occurrence in "Tumor Measurements and Response Document" was 0%. Again, online Appendix B also contains data on secondary person token presence for all 177 analyzed document types and Table 2 includes secondary person token column for selected documents.

To produce a conservative estimate and assuming that every document with a secondary person token includes secondary PxI, then NIH CC's IDR incidence of secondary PxI is 12.94% (48,679 out of 376,051 documents). Using a two-fold condition of simultaneous presence of secondary person token and presence of a priori PHI (primary patient name or MRN), we found that 1.14% of all documents (4,276 out of 376,051) satisfy this stricter condition.

Table 3: Absolute and relative occurrences of individual secondary person tokens across all document instances.

Secondary Person Token	Occurrence	% of All Occurrences	Secondary Person Token	Occurrence	% of All Occurrences
MOTHER	17,979	14.21%	PARENTS	3,196	2.53%
SPOUSE	16,942	13.39%	MOM	2,695	2.13%
PARENT	13,825	10.93%	SIBLING	2,583	2.04%
WIFE	11,858	9.37%	FRIENDS	2,540	2.01%
FATHER	8,959	7.08%	FRIEND	2,528	2.00%
DAUGHTER	6,460	5.11%	CHILD	1,872	1.48%
HUSBAND	6,124	4.84%	SIBLINGS	898	0.71%
SISTER	5,359	4.24%	DAUGHTERS	754	0.60%
BROTHER	5,313	4.20%	DAD	743	0.59%
SON	3,867	3.06%	OTHER (FAMILY)	6,477	5.12%
CHILDREN	3,742	2.96%	OTHER (NON-FAMILY)	1,780	1.41%

Regarding the detection of personal names, we found that a majority of document types include the names of the authoring clinicians (complete data available in Appendix B). Without prior knowledge about local clinician names or document structure clearly tagging electronic signatures, the detection of personal names was not an optimal signal for detecting instances of secondary PxI due to too many false positives.. However, proximity search

measures (personal name near a secondary person token) could potentially result in improved detection of true instances of secondary PxI.

Manual review

For manual review, we selected the "daughter" secondary person token since this was the most frequent secondary person token belonging to an offspring which, in turn, is the generation that is likely to be surviving the primary patient. For the reviewed document type, we selected the highly variable document type "Progress Note – Free Text Document", authored by physicians (rather than nursing staff). This selected document contained a secondary person tokens in 16% of instances. A pilot study of two other pairs of secondary person token and document type did not yield significant findings.

Based on review of all sentences containing the selected token, we classified the secondary PxI instances into three categories listed in Table 4 together with de-identified examples. We found:

1. *Health related secondary information:* e.g., secondary person diagnosis. Note that such information technically constitutes *protected* (in HIPAA sense) health information only if it is explicitly or implicitly combined with a secondary person first name (most often), full name or other identifier in addition to the secondary person token presence. Otherwise, it can be considered de-identified (meets criteria for the Safe Harbor Method).
2. *Non-health secondary information with an identifier (secondary PII)* (e.g., secondary name, secondary phone number or secondary age)
3. *Non-health secondary information without a secondary identifier* (e.g., description of care situations involving a family member)

Table 4: Classification of secondary PxI and examples identified during manual review

Class	Information Domain	Example
Health related secondary information	Diagnosis	Her *mother* and *daughter* both have symptoms of *multiple sclerosis*.
		Patient started on antibiotic post exposure prophylaxis for *pertusis* (daughter tested positive)
Non-health secondary information with an identifier (secondary PII)	Name	I had a frank 2 hour discussion with Ms Doe and her husband and her daughter *Alice*.
		She was accompanied in our meeting by her husband and her daughter *Alice*
	location	His daughter lives in Richmond.
	phone	Mrs. Doe daughter *Alice Carol Smith (123-123-1234)* contacted research nurse Alice Jones, R.N., and requested that I call her.
		I attempted to reach *Mrs. Smith (cell 123-123-1234)* this morning to convey our condolences on Mr. Smith passing earlier this morning.
	age	He is married and has a 9-year-old daughter.
	combination	Mr. Doe young adult daughters (*ages 18 and 23*) as well as his sister and sister-in-law are anticipated to arrive this evening from *Cleveland* around 8:30pm.
		SH: ... Happily married. One son lives in *Annandale – active duty Army*. *Daughter* in *Pennsylvania*. *3 grandkids*.
		Mrs. Doe was accompanied by her daughters *Alice* and *Carol* who drove her from *Pennsylvania* for the procedure.
Non-health secondary information without a secondary identifier	n/a	Pt did ROM exercises in bed with her *daughter*.
		We discussed in detail with patient, *spouse* and *daughter* the progression of disease based on radiographic findings and worsening symptoms
		The *daughter-in-law* will be returning with Ms Smith between 10am - 12pm tomorrow for dressing change, supplies, and d/c instructions. (data note: Smith is the primary patient)
		She stated that his *daughters* did arrive to the hospital but the *sons* had not made it.

Discussion

This study is the first to report on an informatics repository and platform created specifically around deceased subjects. The dsIDR contains a smaller number of patients and observations compared with the full IDR; however, the most frequent items are well represented. For example, 99.9% of the top 2000 the full IDR laboratory tests are also present in the dsIDR. Table 1 shows that while the dsIDR consisted of only 9% of patients (24 000 out of all 265 000 IDR patients with any laboratory data), it accounted for 24% of all laboratory tests and 32% of all medication records. This can be explained by the disproportional care provided to older or sicker patients[11] and also by the NIH Clinical Center casemix bias due to its research hospital status. In building the dsIDR, we were able to

re-use existing processes for structured coded data (e.g., laboratory results) that remove results where comments contain patient name or medical record number. However, due to costs and available resources, robust de-identification techniques for unstructured data (narrative clinical documents) are only now being developed.

Narrative text clinical documents

Since valuable clinical information is often stored in unstructured narrative text clinical documents and the dsIDR's purpose is to serve as a pilot platform, inclusion of narrative-text clinical documents within the dsIDR is highly desirable. In order to protect the identity of secondary persons, removal of primary patient name is highly advisable. For example, if a record mentions the diagnosis and the first name of a family relative, the presence of the primary patient last name helps reveal the full name of the relative. Although the Privacy Act does not apply to dead people, for non-research use, HIPAA protects health records of decedents for the period of 50 years after death as if they were living.

In our effort to remove a priori PHI (more specifically, primary patient identifiers), we found that their presence in documents differs widely (from 0% to 100%). For document types where all or almost all instances contain primary patient name and medical record number the reason may be due to data integration processes for external ancillary system. If an external system is used, re-entering patient IDs ensures proper data integration of data authored in the external system to the primary system or common data viewing platform. Although some document types do contain with high frequency a priori PHI identifiers (Table 2), the overall rate across all documents is only 4.61%. This implies that in majority of narrative clinical documents, the patient identity is not repeated within the text. Therefore, if a document mentions a family relative by relationship (e.g., daughter), 95% of documents will not provide the context of primary patient full name to identify the secondary person. This is also demonstrated by the considerably lower estimate (1.14%) that uses the two-fold criterion as opposed to mere secondary token presence (12.94%).

The review of alphanumeric identifiers found by NLM Scrubber also showed that date of birth is frequently used as safety check within EHR forms and could potentially be another a priori PHI element type. However, including date of birth in this set (together with name and MRN) may be meaningless because dsIDR already contains structured data on date of birth.

Study limitations

Our pilot exploration of secondary PxI in a deceased subject repository has several limitations. Our goal was not to arrive at a final method capable of removing all secondary PxI. Our set of secondary person tokens, although designed to be broad, may not contain all keywords of importance at other institutions. For example, the presented results reflect an older version of this token set that did not contain the keyword "fiancé". Also, we have only used a single de-identification tool and data from a single institution.

Our work relies heavily on the secondary person token presence assumption. It was out of scope of this pilot study to formally evaluate and measure, most likely by manual validation, the percentage of excluded documents (based on explicit presence of a secondary person token) truly contain secondary PxI (true positive vs. false positive). Similarly, we did not validate that included narrative documents, lacking secondary person token, are truly free of any implicit secondary PxI (false negative). This evaluation would improve our estimate of secondary PxI presence which should be interpreted as a machine-based upper bound approximation rather than an exact determination.

Lessons learned and future work

During our pilot creation of the dsIDR we found that removing secondary PxI differs significantly from general de-identification tasks that focus on the 18 identifiers specified by the HIPAA Safe Harbor method. While the Safe Harbor research route is well characterized in the regulation, the regulatory guidance for decedent clause research is less precise. We can apply the Safe Harbor Method guidance on any secondary PxI, but it is less clear on how aggressive the scrubbing must be for the primary deceased patient. The classification shown in table 4 (manual review of secondary PxI instances) shows several examples of varying level of sensitive context ranging from family phone number to a simple age fact). In our interpretation, we assumed that all primary patient information can stay, unless it provides context to reveal secondary PxI of a living person. For example, whereas traditional de-identification within a whole population IDR would require removal of locations and dates, they are permitted to stay in the dsIDR (unless related to a secondary person, which is much less frequent). It is also important to note that the recent regulatory change in decedent PHI protection (shortening to 50 years instead of indefinitely) does not apply to the research context; the decent research clause provides research access to decedents' records immediately after death.

Another useful lesson was about importance of empty (un-filled) EHR forms. We found that many documents are semi-structured and clinicians enter only short entries into this pre-populated structure (e.g., anesthesia type). Some

of those semi-structured document fields are even pre-populated by software (e.g., protocol number, gender) and never edited by the user. Documenting this as metadata at the document level (list of all document questions/sections, software field pre-population, editability) provides useful prior knowledge for scrubbing. In our future revisions, we may consider measuring Levenshtein distance, which is a string metric for measuring the difference between two text sequences as number of single-character edits (empty EHR form vs. complete document). Such an approach is suitable for documents with minimal clinician-entered content (e.g., Post Anesthesia Care Unit Record).

Alternatively, for a limited set of non-dictated documents authored within the modern EHR form paradigm, the optimal dsIDR unit of information is a clinical document section or question rather than the full document. In a separate study,[12] we analyzed problem list sections in dsIDR as well as full IDR. Due to the OHSRP requirement for de-identification, we performed a manual review to verify that truly only problem list information is entered into the problem list document section. We found that clinicians may occasionally forget to hit <tab> or otherwise advance to the next form field and enter social history or family history into the problem list section resulting in PHI being present in the dataset. Another observed problem related to the issue of empty forms and form structure was the presence of a false positive secondary person token (e.g., children) in the form questions or pre-populated entry (e.g., clinical trial title "Continuing Treatment for *Children* and Adults in the Center for Cancer Research; clinicaltrials.gov: NCT00001295).

Regarding the detection of phone numbers as the most frequent alphanumeric identifiers, we found that local hotels, local pharmacies and internal department pagers were often detected as false positive secondary PxI identifiers. Scrubbing software could take advantage of a prior list of permissible phone numbers. Detecting the nature of the phone number by an Internet search engine is another strategy that can distinguish publicly known phone number from phone numbers intentionally kept private. If an EHR contains a structured and well used field that provides primary patient's phone number, it can be used as a priori PHI element to improve the scrubbing accuracy. Similarly, names of investigative drugs could be extracted from clinical trial descriptions and white-listed as allowed alphanumeric identifiers.

Table 5: Summary of lessons learned

- Lack of clear regulatory guidance for de-identification of decedent records for research use (in contrast to the Safe Harbor Method for general de-identification within a whole population IDR)
- Need for metadata on document structure (empty EHR forms; pre-populated fields, secondary person tokens as part of the form question; plain free-text vs. structured provider e-signatures)
- Phone number as the most sensitive secondary ID and possible disambiguation methods for phone numbers (local hotel or pharmacy vs. secondary person private number)
- Provision of prior information to scrubber about local provider names or permitted alphanumeric IDs, such as experimental drug names.
- Importance of accurate deceased status using local institutional data or external data from State Vital Statistics Administration (research context only) or federal death index data

We plan to incorporate some of the above mentioned issues (see Table 5 for summary) in the future versions of our overall processing pipeline and NLM Scrubber feature set. Our pilot study and presented experience offer some guidance to other institutions interested in establishing their own dsIDR in terms of how to characterize the size of the dsIDR, narrative document lessons learned and the value of collecting and verifying death status data. NLM Scrubber is planned to be released freely prior to the AMIA Annual Symposium in November 2014 and could provide to interested institutions as a no-cost software tool with potentially additional experimental features for detecting secondary person tokens. Federation of multiple dsIDRs can address the sample size limitations of any single repository. Central Intelligence Agency Factbook estimates that each month about 221 thousand deaths occur in the USA and 66.5% of those deaths occur outside the home (in hospitals, nursing homes or other facilities; according to a 2009 Medicare study).[13] Internal records and first-hand primary death data are important because national death registries, such as the Social Security Death Master File contained 0.03% false positive deaths and since 2011 omits records from several US states.[3,14]

Conclusions

The HIPAA decedent clause represents an important and often overlooked alternative to using de-identified EHR records. Use of deceased subject records does not require full IRB review or patient consent. The 18 patient identifiers listed in the HIPAA Safe Harbor Method do not have to be removed in dsIDR; however, in order to protect the privacy of secondary living persons, we argue for removal of those that have limited research value. Removal of only secondary PxI results in less distorted research data compared with general de-identification used in whole population IDR. Our study is the first to suggest a basic method for de-identification of decedent records and provides the first estimate of the rate of secondary PxI in EHR data (12.9% using only the secondary person token presence criterion and 1.1% using stricter criteria). With potential increased use of decedent EHR data, clearer guidelines addressing the issue of secondary PxI and what constitutes acceptable proof of death could facilitate such research. We also recommend modifications of existing de-identification programs (such as NLM-NS) to offer a dsIDR scrubbing mode in addition to existing modes where all 18 PHI identifiers are being targeted. In our exploration of secondary PxI, we found secondary person phone numbers to be most frequent and most sensitive secondary person identifiers. Federated dsIDRs can address the issue of limited sample size and make them even primary research platform for certain hypotheses.

Acknowledgments: This work has been supported by intramural research funds from the NIH Clinical Center and the National Library of Medicine. The opinions expressed in this article are authors' own and do not reflect the view of the National Institutes of Health, or the Department of Health and Human Services.

References

1. Kushida CA, Nichols DA, Jadrnicek R, Miller R, Walsh JK, Griffin K. Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. *Med Care* 2012;50 Suppl:S82-101.
2. Meystre SM, Friedlin FJ, South BR, Shen S, Samore MH. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC medical research methodology* 2010;10:70.
3. Huser V, Cimino JJ. Don't take your EHR to heaven, donate it to science: legal and research policies for EHR post mortem. *J Am Med Inform Assoc* 2013.
4. Murphy EC, Ferris FL, 3rd, O'Donnell WR. An electronic medical records system for clinical research and the EMR EDC interface. *Investigative ophthalmology & visual science* 2007;48:4383-9.
5. Electronic Code of Federal Regulations: Title 45: §164.512 Uses and disclosures for which an authorization or opportunity to agree or object is not required. http://www.ecfr.gov/cgi-bin/retrieveECFR?SID=0c83756f3a487d6d70dd3232e084c0a0&n=45y1.0.1.3.78.5&r=SUBPART&ty=HTML#45:1.0.1.3.78.5.27.8.
6. HIPAA Privacy Rule and Its Impact on Research. 2013. http://privacyruleandresearch.nih.gov/pr_08.asp.
7. Ninghui L, Tiancheng L, Venkatasubramanian S. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In: Data Engineering, 2007 ICDE 2007 IEEE 23rd International Conference on; 2007 15-20 April 2007; 2007. p. 106-15.
8. Kayaalp M, Browne AC, Callaghan FM, Dodd ZA, Divita G, Ozturk S, et al. The pattern of name tokens in narrative clinical text and a comparison of five systems for redacting them. *J Am Med Inform Assoc* 2013.
9. M K, AC B, Z D, P S, CJ M. Technical Report to the LHNCBC Board of Scientific Counselors: Clinical Text De-Identification Research 2013.
10. LOINC Top 2000+ Lab Observations (US version) 1.1. 2014. http://loinc.org/usage/obs/loinc-top-2000-plus-loinc-lab-observations-us.csv/view.
11. Gielen B, Remacle A, Mertens R. Patterns of health care use and expenditure during the last 6 months of life in Belgium: differences between age categories in cancer and non-cancer patients. *Health policy* 2010;97:53-61.
12. Huser V, Fung KW, Cimino JJ. Natural Language Processing of Free-text Problem List Sections in Structured Clinical Documents: a Case Study at NIH Clinical Center. *AMIA Summits Transl Sci Proc (accepted)* 2014.
13. Teno JM, Gozalo PL, Bynum JP, Leland NE, Miller SC, Morden NE, et al. Change in end-of-life care for Medicare beneficiaries: site of death, place of care, and health care transitions in 2000, 2005, and 2009. *JAMA* 2013;309:470-7.
14. Death Master File: Important Notice. 2011. http://www.ntis.gov/pdf/import-change-dmf.pdf.

TextHunter – A User Friendly Tool for Extracting Generic Concepts from Free Text in Clinical Research

Richard G. Jackson MSc[1], Michael Ball MSc[1], Rashmi Patel BMBCh[1], Richard D. Hayes PhD[1], Richard J.B. Dobson PhD[1], Robert Stewart MD[1]

[1]King's College London (Institute of Psychiatry), London, UK

Abstract

Observational research using data from electronic health records (EHR) is a rapidly growing area, which promises both increased sample size and data richness - therefore unprecedented study power. However, in many medical domains, large amounts of potentially valuable data are contained within the free text clinical narrative. Manually reviewing free text to obtain desired information is an inefficient use of researcher time and skill. Previous work has demonstrated the feasibility of applying Natural Language Processing (NLP) to extract information. However, in real world research environments, the demand for NLP skills outweighs supply, creating a bottleneck in the secondary exploitation of the EHR. To address this, we present TextHunter, a tool for the creation of training data, construction of concept extraction machine learning models and their application to documents. Using confidence thresholds to ensure high precision (>90%), we achieved recall measurements as high as 99% in real world use cases.

Introduction

The increasing use of electronic health records (EHR) provides potentially transformative opportunities for clinical research in the breadth and depth of data contained within them. However, unstructured clinical notes are often the most valuable source of phenotypic/contextual information because of limitations in the scope and acceptability of structured fields. In response to this challenge, Natural Language Processing (NLP) has been employed to extract appropriate assertions in a structured format amenable to the needs of researchers[1]. While significant success has been achieved in many areas, the demand for ever more variables to be extracted from the clinical narrative is currently bottlenecked by the limited supply of technical skills[2]. For example, rule based approaches to novel problems are often effective, but require a certain degree of technical knowledge and experience, which can be too time-consuming and thus expensive to produce in high volume. Proposed solutions include ontological or dictionary mapping techniques, which are appropriate where there are well-constructed resources; however, the standards imposed by these may not be easy to adapt to real world clinical sub-languages[3], or where there is controversy about the appropriate use of clinical language[4]. Machine Learning (ML) approaches are an increasingly popular means of circumventing rule based systems, but require even more technical expertise and are limited by the availability and ease of creating appropriate training data[5,6]. Finally, although progress has been made in the development of publicly available corpora for evaluating different clinical NLP methodologies[7], these offer no guarantee that the performance obtained by models trained on such data will provide a generalizable solution (for example, for work on EHRs in different medical domains, dialects, languages, or work cultures)[8].

These issues form barriers to progress for groups who have access to unstructured clinical data, but do not have sufficient technical capabilities to trial the wealth of information extraction techniques on offer. In recent years this has prompted the development of tools such as Arc[9] to democratize access to generic information extraction capabilities. However, there are currently no free tools available that offer a full end-to-end solution for concept level extraction, including the principle tasks of:

1) Extracting instances of concepts from a database or large collection of documents

2) Creating sufficient training data specific to a concept to enable a machine learning approach

3) The configuration and testing of an (ML) algorithm for the given concept

4) The application of the model to the entire document set of interest, and the subsequent export of results into a familiar format

In order to make concept extraction technologies accessible to groups without informatics support, we have developed the TextHunter tool to address these tasks.

Methods

Data: The South London and Maudsley mental health case register

The South London and Maudsley NHS Trust (SLAM) is the largest mental health organization in Europe, and is a virtual monopoly provider of mental health services to 1.2 million individuals within its geographical catchment area (Lambeth, Southwark, Lewisham and Croydon boroughs in South London). In 2007-08, funding from the British National Institute for Health Research supported the development of the Clinical Record Interactive Search (CRIS) database. CRIS operates as a pseudonymized version of SLAM's EHR system, accessible for researchers via its distinctive, patient-led information governance model[10]. CRIS houses more than 230,000 de-identified patient records, which in turn represent over 20 million free text documents. The CRIS system continues to grow at a rate of approximately 170,000 free text documents per month. Clinical information documented in unstructured text is of particular value in mental health research where there is an increasing emphasis on using dimensional symptom scales to define mental illness rather than discrete diagnostic categories[11–14]. While CRIS also has large amounts of data contained within structured fields, the development of TextHunter was precipitated by the needs of many disparate groups of researchers who require access to the wealth of additional information contained within the clinical narrative.

TextHunter System Description

TextHunter is a program that guides a user through all of the required processes to create and apply a concept extraction model for a selection of documents from start to finish. It performs six important tasks, the end result of which delivers a structured representation of a concept. Its intended use case is typically phenotype cohort identification, although it can be employed for more generic purposes. The program is built from open source libraries, and uses the GATE library as its core NLP engine[15]. The ML element uses the Support Vector Machine (SVM) based 'Batch Learning' plugin supplied with GATE[16]. In consideration of the rigorous information governance requirements of clinical data, TextHunter is designed to operate as a standalone 'offline' program on desktop hardware, although its multithreaded design enables its deployment on more powerful workstations/virtual machines to handle larger datasets. It is capable of connecting to commercial database environments such Microsoft SQL Server to process massive datasets, but for succinctness, only its standalone operation mode is described here.

The underlying principle of TextHunter is 'Find, Annotate, Build, Apply' - respectively addressing the four key problems described above. The integration of these concepts into a single system creates the possibility of providing lay users with access to more advanced ML techniques, such as active learning. Each phase of the TextHunter pipeline is described below:

1. Search Phase

This phase addresses task 1). The first stage of the TextHunter pipeline requires a user to define a list of keywords, regular expressions and/or phrases to describe their concept of interest. The user then directs the program to a directory holding the text files of interest. Upon executing the 'search' phase, each document is scanned for mentions of the user's expressions. When a mention is identified, a short section of text consisting of multiple sentences, including the sentence where the concept mention was found, and up to two sentences either side of the sentence of interest is extracted. This is stored in an embedded file based database, along with a copy of the underlying document. Deconstructing documents in this way facilitates the downstream management of text instances for annotation and classification.

2. Annotation Phase

This phase addresses task 2). The user is directed to TextHunter's annotation interface, which has been specifically designed for the rapid annotation of concept instances. We define an instance as a group of one to five sentences centered on a concept keyword, and its classification as defined below:

i) Positive – the example is a relevant hit and is an appropriate positive example of the user's concept

ii) Negative - the example is a relevant hit and is an appropriate negated example of the user's concept

iii) Unknown - the example is a relevant hit but the user is unable to ascertain the correct classification, or the example is irrelevant

In this phase, the user is required to produce a 'test' corpus for model validation (typically of 100-300 instances), which are randomly selected from all instances in the document set. This is followed by the production of a 'seed' corpus to be used in training models. This also numbers about 100-300 instances, but is enriched by ensuring no identical instances are present. In real world clinical datasets, the required semantic context that enables the classification of a concept instance may cross sentence boundaries. To ensure appropriate features are available for training, the user can specify the required 'context' (up to two sentences before and two sentences after) needed to make the classification, centered on the sentence containing the concept keyword. These boundaries are arbitrarily chosen by the GATE sentence splitter module, although we expect that only in very rare cases will more than five sentences be required to express medical concepts as they are normally found in EHRs.

3. Feature selection/Model Building Phase

This phase addresses task 3). Here, TextHunter builds and evaluates a range of models against the task, using different features and SVM parameters each time. The default feature vector used by TextHunter is a classic bag of words using part-of-speech tags and token stems from the user specified context around a concept. When applying a model to unseen data, TextHunter creates feature vectors from up to six different combinations of sentences around the sentence containing the concept term. The classification resulting from the feature vector producing the highest overall confidence is chosen as the result. In addition, TextHunter has a modular design that allows developments from the clinical NLP community to be integrated into its core pipeline via GATE creole plugins. Currently, TextHunter takes features of the GATE implementation of the ConText algorithm[17], which uses hand crafted rules to determine whether a concept is negated, temporally irrelevant or refers to a subject other than the patient. Stop word removal is also explored during feature selection.

Cross validation of the training data is used to mitigate the dangers of overfitting the model to a small amount of data. The model producing the best F1 score is taken forward for testing against the human labeled 'test' corpus, which is never used in model training. A range of easy to interpret output files are produced, containing estimates of 'real world' performance the user might expect.

4. Application Phase

This phase addresses task 4). This phase allows the user to apply the best performing model to all instances of text in their dataset, as captured in the search phase. As with the model building phase, combinations of sentences are tested around the concept. The classification that results from the combination with the highest confidence is chosen as the final result. Once this stage is complete, the user may export the output into several formats.

5. Active Learning Phase (optional)

Conceptually, active learning is an iterative process whereby an ML algorithm selects instances that it has difficulty classifying and presents them to a human annotator for labeling. These are then fed back into the model, with the intention that the new model arising will be better at classifying similar, difficult examples. TextHunter supports a 'simple margin' inspired method of active learning[18]. A seed model is constructed from randomly selected instances of text, as described above. This model is then applied to a large sample of the entire population of relevant text instances. For each classification the model makes, it also assigns a level of certainty, between -1 and +1. Theoretically, highly positive scores are representative of easy to classify 'positive' instances, whereas highly negative scores are representative of easy to classify 'negative' or 'unknown' instances. Instances with a certainty score close to 0 are thus 'difficult', and presented to the user for labeling in order to retrain the classifier.

Use cases

To evaluate the performance of TextHunter, we defined three real world use cases of concept extraction. Examples of search expressions and typical instances for each use case are detailed in Table 1:

Case Study 1: Cannabis Smoking

Cannabis use has been indicated as a potentially aggravating factor in patients suffering from mental illness[19]. Through the vast amount of electronic documentation generated in the course of patient care, we attempted to identify a patient's cannabis smoking status based upon reports by mental health professionals. The CRIS database contains intra-profession clinical correspondence style documents and clinical notes resulting from patient contact. Each type of document may contain references to cannabis usage by the patient. In this study, our objective was to use TextHunter to build a classifier to identify current or historical cannabis usage. We conducted a review of the most common nouns and slang terms used to describe cannabis in SLAM, to produce a list of expressions which formed the basis for finding instances to classify. A psychiatrist then produced multiple sets of annotations using the standard TextHunter procedure, making use of the active learning functionality. Although it was not possible to double annotate the training data, we adopted a restrictive manual coding strategy in order to allow as little subjectivity as possible (for example, by classifying mentions pertaining to future events, or tangential/circumstantial references into our predefined 'unknown' class).

Case Study 2: Psychosis Symptomatology

Patients suffering from psychosis can exhibit a wide range of symptoms, which in turn inform the nature of their treatment plan. Common tools to quantify symptomatology in psychosis include such instruments as the Positive and Negative Symptom Scale and the Clinical Assessment Interview for Negative Symptoms[12,13]. These depend on an assessment of the patient's presentation in regard to a wide range of possible symptoms. Our previous work to capture some of these from clinical notes with ML approaches has been described[20,21]. In this case study, we used TextHunter to capture two additional symptoms: delusional symptoms and evidence of hallucinations, using the standard TextHunter workflow. The annotated data for 'delusions' were generated by a clinical informatician, with a random sample checked for accuracy and consistency by a psychiatrist. In the case of 'hallucinations', all annotations were generated by a public health physician. In both cases, the restrictive coding strategy as described above was employed.

Table 1: Search expressions and examples of instances. Theoretical patient identifiers masked by ZZZZZ.

Case Study	**Examples of subword patterns (case insensitive) for search phase**	**Fictitious examples of instances (Parentheses indicates typical labeling by human annotator)**
Cannabis smoking	cannab hash weed pot	ZZZZZ told me that he continues to smoke cannabis only no other illicit drugs. **(positive)** ZZZZZ has no history of amphetamine or cannabis use. **(negative)**
Psychosis symptomatology	delusio hallucina	She is continuing to experience hallucinations and is becoming increasingly distressed by these. **(positive)** Staff observed him to rambling and delusional, repeating himself and his gait was abnormal and more pronounced. **(positive)**

Case Study 3: Ethnicity

Ethnicity is a key variable in many epidemiological and clinical studies. Although ethnicity can theoretically be captured via the structured elements in SLAM's EHR system, in reality, it is often not recorded in the course of routine clinical practice. However, as with many other variables, ethnicity is often referenced in clinical free text. The purpose of this case study was therefore to classify instances of text describing a patient's ethnicity, into one of 17 ethnic groups. A range of terms was selected in association with each ethnic group, and a 'positive' classification was made if the context for the term was suggestive of the patient belonging to that group. A single researcher produced the annotated dataset for training/testing, using a similarly restrictive coding strategy.

In each case study, a sample of the evaluation instances were double annotated by an individual in a related profession to generate inter-annotator agreement statistics.

Results

In all case studies, we used 10 fold cross validation for the model building phase, which took approximately one hour on a desktop computer with a Core 2 Duo E7500 processor.

In the cannabis smoking study, we used 13 terms to capture cannabis mentions. The CRIS database yielded 663,979 mentions of cannabis. For the psychosis symptomatology study, the search phase found 603,818 mentions of delusions, and 703,996 mentions of hallucinations. Each symptom was represented by a single term in the search phase. Finally, there were 3,444,435 mentions of concepts potentially related to ethnicity, resulting from 277 terms commonly used to define our 17 ethnic identities.

Traditionally, the performances of information extraction algorithms in NLP are described in terms of precision, recall and the F1 statistic. However, the high level of noise commonly associated with EHR based observational research necessitates the capture of high quality data in order to generate clearly defined cohorts. This data quality requirement restricts the use of automated concept extraction techniques to those that can be shown to have a high true positive rate, relative to the inherent predictive value of a mention of a concept. For example, a mention of a cannabis synonym will refer to a patient's current or past use 70% of the time, whereas a mention of a term denoting ethnicity will refer to a patient's actual ethnicity only 20% of the time (Table 1). A further consideration of the real world viability of a given model is the longitudinal nature of the electronic health record. A patient may have numerous contacts with a health service over a number of years, creating multiple instances of time independent concepts. For example, a patient may have multiple references to their cannabis consumption habits, especially if it is identified as a factor in their illness. Similarly, a patient's ethnicity may be described in service referral letters generated during the course of their care. Only one positive instance needs to be captured precisely for a high quality output to be achieved. However, spurious data points are more problematic. Given these factors, it is more practical to develop information extraction tools that favor precision over recall in most use cases. For this reason, in Table 2 we describe the recall statistic at two arbitrarily defined levels of precision (90% and 95%), which are identified by filtering the classified instances in the test set via the classification confidence threshold. We present Receiver-Operator Characteristic (ROC) plots for each case study in Figure 1. For brevity, we only report the highest F1 achieved without any confidence filtering (note, this is not necessarily the same model that achieves the highest recall at the 90%/95% precision threshold).

The best performance was seen in the hallucinations case study, with over 97 % recall obtained at the 95% precision threshold. The worst performance was observed in the ethnicity study, where recall reached only 9% at 90% precision, and declined with further training.

Different problems required different features in order to obtain the best overall result. In Table 3, we present the types of features that were found to be most useful in each case study.

The rate of training data production varied moderately between the studies, the slowest recorded at approximately 100 instances labeled per hour, and the fastest at roughly 230 instances per hour. Since different individuals annotated each study, further comparisons were not possible. Anecdotal reports from the annotators suggested that the process of annotating instances selected via active learning was slower than the randomly selected instances in the seed set.

Table 2: Performance statistics for TextHunter 'positive' instances ('unknown' and 'negative' instances are grouped together). [1]Observed Agreement and Cohen's Kappa. [2]Baseline precision assumes presence of keyword is a 'positive' instance (by definition, recall is 100%), and provides a measure of how predictive a mention is of a concept without any processing applied. P = precision, R = recall, F1 = harmonic mean of precision and recall. [3]Parentheses indicate count of training instances in the model building phase (subsequent active learning iterations increase the number of training instances available). [4]Recall measured at precision levels of 90% and 95%, attained by confidence filtering.

Case Study	Inter annotator agreement [1,3]	Test Instances	Baseline precision[2]	Seed data base performance[3]	Seed data Recall[3,4]		Active learning iteration 1 recall[3,4]		Active learning iteration 2 recall[3,4]		Approximate total annotator time spent creating training data
					90	95	90	95	90	95	
Cannabis smoking	88% 0.76 (211)	233	75%	P = 81% R = 95% F1 = 0.87 (478)	45	38	68	52	72	53	~10 hours
					(478)		(1 329)		(1 835)		
Delusions	95% 0.91 (110)	206	68%	P = 89%/ R = 99%/ F1 = 0.93 (708)	95	87	N/A		N/A		~4 hours
					(708)						
Hallucinations	89% 0.78 (117)	131	70%	P = 93% R = 99% F1 = 0.96 (150)	99	97	99	97	N/A		~7 hours
					(150)		(914)				
Ethnicity	97% 0.94 (201)	650	20%	P = 82% R = 75% F1 = 0.78 (396)	9	9	3	3	N/A		~3 hours
					(396)		(805)				

Table 3: Additional features used in best performing model delivering >90% precision

Case Study	Best Model ID	ConText used?	Stop words removed?	SVM Cost	SVM kernel type
Cannabis Smoking	128	No	No	0.6	polynomial
Delusions	136	No	No	0.6	polynomial
Hallucinations	88	Yes	No	0.5	polynomial
Ethnicity	24	Yes	Yes	0.7	polynomial

Discussion:

In our analysis, we used TextHunter to extract a diverse set of concepts that are typically in demand in clinical research environments. We arbitrarily set two desired precision standards, and adopted strategies to try to maximize the recall given this requirement. Three of the four test cases reached over 70% recall at the lower precision cut-off of 90%. We do not attempt to tackle the question of what constitutes acceptable performance for research applications here. Nevertheless, we have confidence that the range of case studies investigated here establishes a

proof of concept in enabling end users to create and deliver information extraction solutions independently of significant NLP expertise.

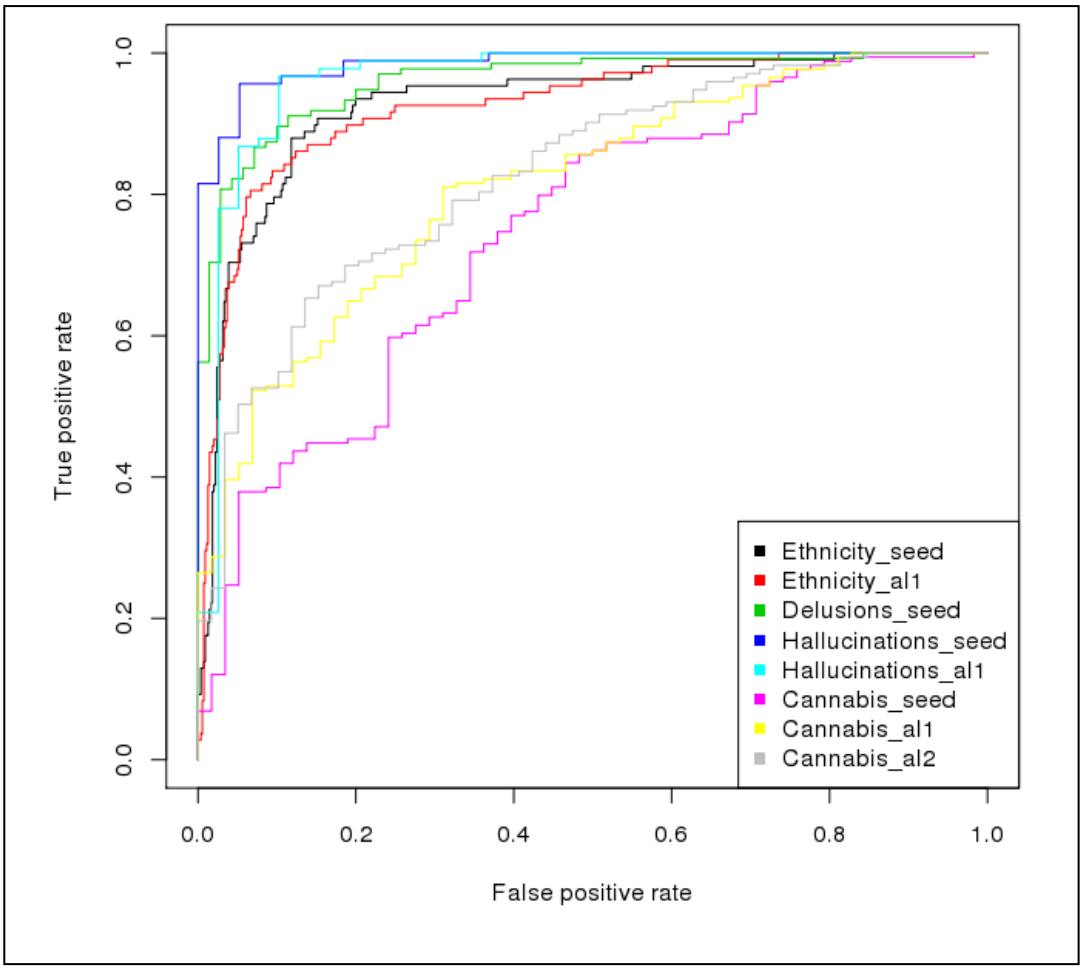


Figure 1: Receiver Operator Characteristic for TextHunter models on 'test' data, generated with SVM confidence thresholds.

Given our limited range of test cases, the SVM parameters and additional features used varied greatly, even between the two conceptually similar problems explored in psychosis symptomatology. This substantiates our approach of testing a range of models to find the best solution for a given problem. However, a predominant factor in the algorithms' ability to reach higher levels of recall is the predictive value that a simple mention of a concept produces (i.e. how likely a human annotator is to label a randomly selected mention of a concept as 'positive'). For instance, the ROC curve produced for the ethnicity study compares favorably with that of the cannabis study, and we achieved a substantial performance benefit over the baseline precision for our list of ethnicity terms. However, because of our self-imposed requirement of a minimum 90% precision, the recall for ethnicity falls very quickly as this threshold is approached. Intuitively, in high noise datasets where 'positive' mentions of a concept are rarer, the concept extraction problem is significantly more challenging. In addition, the low predictive value of ethnicity terms means the 'positive' class will be less represented than the 'negative' or 'unknown' classes in the model. Currently, TextHunter makes no adjustment for unbalanced classes, and future work could investigate mitigation strategies for this, such as using uneven margins [22]. It should also be noted that we were required to use many more terms to capture mentions of ethnicity, which may be indicative of the inherent difficulty of defining concepts that are largely social constructs.

In the case of the cannabis study, we were able to improve the model substantially by providing additional training data through active learning. We did not try to quantify the added benefit of adopting an active learning

methodology over randomly selecting new instances. However, others have previously demonstrated that active learning can accelerate the development of machine learning models in clinical NLP[18,23,24]. Active learning did not produce an additional benefit in the hallucinations case study, although the model resulting from the seed data had already produced a very high F1 statistic. Here, our application of confidence filters was not required, as the performance of the model generated from the seed annotations surpassed our precision requirement of 95%. In the case of ethnicity, adopting an active learning approach noticeably depreciated the quality of the model. To investigate, we conducted a subjective review of the instances that active learning retrieved. This revealed that many were incoherent strings of text, seemingly resulting from jumbled emails, faxes and other malformed documents. Since these were not representative of natural language, their inclusion in training the model possibly introduced more noise than benefit. Previous reports have highlighted the difficulties of applying general NLP tools on clinical text[8,25], and we suspect that this scenario is not uncommon in real world EHR systems. One possible mitigation strategy would be to employ document classification methods to filter out malformed documents and/or a more sophisticated active learning methodology, such that new training data are more representative of the instances of interest. Nevertheless, an SVM approach as implemented in TextHunter appears to be valid for simple concepts that tend be succinctly expressed - for example, if it can be defined with a relatively short list of keywords, is not over-complicated by frequent ungrammatical usage (such as in lists or questionnaire text) and has a baseline precision of at least 60%.

It was not practical to double annotate our training data fully, so we are only able to provide inter-annotator agreement (IAA) statistics for a subset of the total test set in each case study. Despite our limited set, our data suggest relatively high levels of agreement, highlighting a high degree of objectivity in the expression of concepts in clinical text. However, clinical constructs in mental illness are often subtle. Initial reports from annotators in each case study suggested that the annotation process itself influenced their own views on the interpretation of notes created by others. Specifically, the exposure to a wide range of writing styles from other clinicians may introduce unforeseeable subjectivity into the annotation process. Regardless, methods that place subject matter experts (rather than NLP specialists) in the role of defining a concept are likely to be less subjective, as any subjectivity introduced by the annotation process will likely be compounded by attempting to convey the subtleties to a non-expert third party. Any clinical subjectivity may then be mitigated by a process of iterative discussion and re-annotation to produce well defined annotation guidelines. A potentially useful future development of TextHunter may be to incorporate a model of clinical data, such as the Clinical Element Model[26]. This would encourage the re-use of standard definitions of concepts, thus promoting greater interoperability with NLP tools.

A notable shortcoming of the TextHunter methodology was the ethnicity case study, which had the highest Kappa statistic but the lowest F1 score from the seed data. This highlights the divide between human and machine interpretation, and the need for more complex reasoning systems to resolve more difficult problems.

Conclusion

The requirement to develop this software was driven by an imbalance between the demand for concept extraction and the supply of skilled individuals capable of delivering solutions to the needs of researchers. We have shown that it is feasible to package an appropriate suite of tools into a simple interface, and that this enables researchers to produce concept extraction models without input from NLP specialists. TextHunter uses a flexible SVM based algorithm as a generic, user friendly information extraction capability. We have validated the methodology with a variety of typical problems, and produced high precision and relatively high recall models. Although it is not suitable for all tasks, we argue that the ‘solve small problems quickly’ approach to information extraction is appropriate for many types of variable likely to be of interest to researchers, and offers the attractive advantage of rapidly generating models that have been trained on data sourced from the intended target. Finally, the simple annotation interface enables a rapid annotation process, with labeled data stored in a standard, reusable format. The pipeline style operation of GATE and the open source licence of TextHunter should encourage the future development of additional features to improve performance and expedite its use on more complex NLP problems.

TextHunter is available at https://github.com/RichJackson/TextHunter

Funding/Support Acknowledgement

RJ, RD and RS are part-funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. MB is supported by the BRC Nucleus jointly funded by the Guy's and St Thomas' Trustees and the South London and Maudsley Trustees. RP is supported by a Medical Research Council Clinical Research Training Fellowship. RH is funded by a Medical Research Council (MRC) Population Health Scientist Fellowship. RD and RS are joint last authors on this work.

References

1. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. Journal of the American Medical Informatics Association. 2013 Nov 7;21(2):221–30.

2. Chapman WW, Nadkarni PM, Hirschman L, D'Avolio LW, Savova GK, Uzuner O. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. Journal of the American Medical Informatics Association. 2011 Aug 16;18(5):540–3.

3. Demner-Fushman D, Mork JG, Shooshan SE, Aronson AR. UMLS content views appropriate for NLP processing of the biomedical literature vs. clinical text. Journal of Biomedical Informatics. 2010 Aug;43(4):587–94.

4. Chmielewski M, Bagby RM, Markon K, Ring AJ, Ryder AG. Openness to Experience, Intellect, Schizotypal Personality Disorder, and Psychoticism: Resolving the Controversy. J Pers Disord. 2014 Feb 10;

5. Afzal Z, Schuemie MJ, van Blijderveen JC, Sen EF, Sturkenboom MCJM, Kors JA. Improving sensitivity of machine learning methods for automated case identification from free-text electronic medical records. BMC Med Inform Decis Mak. 2013;13:30.

6. Khor R, Yip W-K, Bressel M, Rose W, Duchesne G, Foroudi F. Practical implementation of an existing smoking detection pipeline and reduced support vector machine training corpus requirements. Journal of the American Medical Informatics Association. 2013 Aug 6;21(1):27–30.

7. Zhai H, Lingren T, Deleger L, Li Q, Kaiser M, Stoutenborough L, et al. Web 2.0-based crowdsourcing for high-quality gold standard development in clinical natural language processing. J Med Internet Res. 2013;15(4):e73.

8. Patterson O, Hurdle J. Document Clustering of Clinical Narratives: a Systematic Study of Clinical Sublanguages. AMIA Annu Symp Proc. 2011.

9. D'Avolio LW, Nguyen TM, Goryachev S, Fiore LD. Automated concept-level information extraction to reduce the need for custom software and rules development. Journal of the American Medical Informatics Association. 2011 Jun 22;18(5):607–13.

10. Stewart R, Soremekun M, Perera G, Broadbent M, Callard F, Denis M, et al. The South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLAM BRC) case register: development and descriptive data. BMC Psychiatry. 2009;9:51.

11. Adam D. Mental health: On the spectrum. Nature. 2013 Apr 24;496(7446):416–8.

12. Kring AM. The Clinical Assessment Interview for Negative Symptoms (CAINS): Final Development and Validation. American Journal of Psychiatry. 2013 Feb 1;170(2):165.

13. Kay SR, Fiszbein A, Opler LA. The positive and negative syndrome scale (PANSS) for schizophrenia. Schizophr Bull. 1987;13(2):261–76.

14. Axelrod BN, Goldman RS, Alphs LD. Validation of the 16-item Negative Symptom Assessment. J Psychiatr Res. 1993 Sep;27(3):253–8.

15. Cunningham H, Tablan V, Roberts A, Bontcheva K. Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. Prlic A, editor. PLoS Computational Biology. 2013 Feb 7;9(2):e1002854.

16. Chang C-C, Lin C-J. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology. 2011 Apr 1;2(3):1–27.

17. Harkema H, Dowling JN, Thornblade T, Chapman WW. ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. Journal of Biomedical Informatics. 2009 Oct;42(5):839–51.

18. Koller D, Tong S. Support vector machine active learning with applications to text classification. The Journal of Machine Learning Research. 2001;2:45–66.

19. Moore TH, Zammit S, Lingford-Hughes A, Barnes TR, Jones PB, Burke M, et al. Cannabis use and risk of psychotic or affective mental health outcomes: a systematic review. The Lancet. 2007 Jul;370(9584):319–28.

20. Gorrell G, Jackson R, Roberts A. Finding Negative Symptoms of Schizophrenia in Patient Records. Proc NLP Med Biol Work (NLPMedBio). Hissar, Bulgaria; 2013. p. 9–17.

21. Patel R, Jayatilleke N, Jackson R, Stewart R, McGuire P. Investigation of negative symptoms in schizophrenia with a machine learning text-mining approach. The Lancet. 2014 Feb;383:S16.

22. Li Y, Bontcheva K, Cunningham H. Adapting SVM for data sparseness and imbalance: a case study in information extraction. Natural Language Engineering. 2008 Dec 18;15(02):241.

23. Chen Y, Carroll RJ, Hinz ERM, Shah A, Eyler AE, Denny JC, et al. Applying active learning to high-throughput phenotyping algorithms for electronic health records data. Journal of the American Medical Informatics Association. 2013 Jul 13;20(e2):e253–e259.

24. Figueroa RL, Zeng-Treitler Q, Ngo LH, Goryachev S, Wiechmann EP. Active learning for clinical text classification: is it better than random sampling? Journal of the American Medical Informatics Association. 2012 Jun 15;19(5):809–16.

25. Barrett N, Weber-Jahnke JH. Applying natural language processing toolkits to electronic health records - an experience report. Stud Health Technol Inform. 2009;143:441–6.

26. Wu ST, Kaggal VC, Dligach D, Masanz JJ, Chen P, Becker L, et al. A common type system for clinical natural language processing. Journal of Biomedical Semantics. 2013;4(1):1.

Analysis of Online Information Searching for Cardiovascular Diseases on a Consumer Health Information Portal

Ashutosh Jadhav, MS[*1], Amit Sheth, PhD[1], Jyotishman Pathak, PhD[2]
[1]Knoesis Center, Wright State University, Dayton, OH; [2]Mayo Clinic, Rochester, MN

ABSTRACT

Since the early 2000's, Internet usage for health information searching has increased significantly. Studying search queries can help us to understand users "information need" and how do they formulate search queries ("expression of information need"). Although cardiovascular diseases (CVD) affect a large percentage of the population, few studies have investigated how and what users search for CVD. We address this knowledge gap in the community by analyzing a large corpus of 10 million CVD related search queries from MayoClinic.com. Using UMLS MetaMap and UMLS semantic types/concepts, we developed a rule-based approach to categorize the queries into 14 health categories. We analyzed structural properties, types (keyword-based/Wh-questions/Yes-No questions) and linguistic structure of the queries. Our results show that the most searched health categories are 'Diseases/Conditions', 'Vital-Sings', 'Symptoms' and 'Living-with'. CVD queries are longer and are predominantly keyword-based. This study extends our knowledge about online health information searching and provides useful insights for Web search engines and health websites.

INTRODUCTION

Since the last decade, Internet literacy and the number of Internet users have increased exponentially. With the growing availability of online health resources, consumers are increasingly using the Internet to seek health related information[1,2]. Online health resources are easily accessible and provide information about most of the health topics. These resources can help non-experts to make more informed decisions and play a vital role in improving health literacy. According to the Center for Disease Control and Prevention (CDC) in the United States, CVD is one of the most common chronic diseases and the leading cause of death (1 in every 4 deaths) for both men and women[3]. CVD is common across all socioeconomic groups and demographics including age groups, genders, and ethnicities. Most of the CVDs require lifelong care and the patient is in charge of managing the disease through self-care (such as diet, exercise, and other health lifestyle choices)[4]. Prior studies[4,5] have shown that online resources are a "significant information supplement" for the patients with chronic conditions. As the percentage of people suffering from CVD is very high, the number of people using the Internet to search and learn about CVD is also large[4,5].

One of the most common ways to seek online health information is via Web search engines such as Google, Bing, Yahoo!, etc. According to the Pew Survey[1], approximately 8 in 10 online health inquiries start from a Web search engine. Therefore, studying CVD related search logs can help us to understand what health topics Online Health Information Seekers (OHIS) search for ("information need") and how do they formulate search queries ("expression of information need"). Such knowledge can be applied to improve the health search experience as well as to develop more advanced next-generation knowledge and content delivery systems. Although chronic diseases affect a large population, very few prior studies have investigated online health information searching exclusively for chronic diseases and especially for CVD. In this study, we address this knowledge gap in the community by performing a comprehensive analysis on a significantly large corpus of 10 million CVD related search queries. These queries are submitted from Web search engines and directed OHISs to the Mayo Clinic's consumer health information portal[6].

One of the contributions of this paper is the demonstration of the effectiveness of UMLS MetaMap[7,] as well as UMLS semantic types and concepts for customized categorization. We implemented a rule based categorization approach (with Precision: 0.8842, Recall: 0.8607 and F1 Score: 0.8723) based on the UMLS semantic types and UMLS concepts. Using our approach, we categorized 92% of the 10 million CVD related search queries into 14 "consumer oriented" health categories. Additionally, we analyzed the structural properties of the queries (length of the search queries, usage of search query operators and special characters in the search queries), types of search queries (keyword-based, Wh-questions, Yes/No questions), and misspellings in the search queries. As the linguistic structure of the search queries has implications on information retrieval using Web search engines[8], we have also analyzed basic linguistic characteristics of the CVD search queries.

[*] This work was done during author's internship at Mayo Clinic, Rochester, MN, United States.

As per our analysis, the top searched health categories for CVD are 'Diseases and Conditions', 'Vital Sings', 'Symptoms' and 'Living with'. CVD search queries are longer and are predominantly keyword based. CVD queries have few search query operators, special characters, and spelling mistakes. This study provides useful and interesting insights for online health information seeking in chronic diseases and particularly in CVD. Such knowledge gives us better understanding about how OHIS search for CVD information and their information needs, which can be utilized to improve the health information search process.

Related work

Many previous studies have investigated online health information searching behavior, primarily using focus group studies, user surveys, and by analyzing health-related Web search query logs. In the studies[9-11] based on focus group and user surveys, researchers have analyzed online health searching characteristics such as how people use the Internet for health information searching, their demographic information (age, gender, education level, etc.), devices/Web search engines used for the searching, online health searching in the case of specific health conditions, and age-groups. Although these studies provided important insights, their main limitation was the inclusion of a limited number of participants (ranging from 100 to 2000 people), which may not cover population-level diversity and represent all socioeconomic groups. Researchers have analyzed web search logs from the health domain with diverse objectives, such as health/epidemic surveillance[12], PubMed usage[13], and online health information searching[14-19]. The studies focusing on online health information searching, have studied a variety of aspects of health query logs, such as health query length, changes in the health behavior with type of disease[17], effect of device used for health information searching[18] and changes in online health search patterns with disease escalation from symptoms to serious illness[19].

Previous studies in chronic diseases have looked at different facets of the diseases such as how education, income, and occupation contribute to risk factors for cardiovascular diseases[20]; use of information technology to improve the management of chronic diseases which includes home monitoring of vital signs for patients with chronic diseases[21]. Ayer et al.[4] studied the relationship between chronic illness, use of the Internet for health information, and change in health behavior. The study suggests that the use of the Internet empowers patients' in the management of their chronic conditions resulting in an increased ability to make informed decisions about their health. Lorig et al.[22] indicated the effectiveness of Internet based chronic disease self-management. Although chronic diseases affect a large population and prior studies have highlighted the usefulness of the online health resources for patients with chronic diseases, very few studies[4] have investigated how and what OHIS search for chronic diseases and especially for CVD. A focused study on chronic diseases, such as the CVD use-case presented in this paper, helps us to study details about online health searching for chronic diseases that may not be revealed through analyzing online health searching in general. Such knowledge can be applied to improve online health information systems, to promote health literacy and to accomplish a more balanced approach to wellness and prevention of the chronic diseases.

MATERIALS AND METHODS

Data Source: In this study, we have collected CVD-related search queries originating from Web search engines (such as Google and Bing) that direct OHISs to Mayo Clinic's consumer health information portal[6] (MayoClinic.com), which is one of the top online health information portals within the United States. The MayoClinic.com portal provides up-to-date, high-quality online health information produced by professional writers and editors. Our recent Web analytics statistics indicate that the MayoClinic.com portal is on average visited by millions of unique visitors every day and around 90% of the incoming traffic is originated from Web search engines. This significant traffic to the portal provides us with an excellent platform to conduct our study.

Dataset Creation: The MayoClinic.com Web Analytics tool (IBM Netinsight on Demand) keeps detailed information about Web traffic such as input search query, time of visit, and landing page. MayoClinic.com has several CVD-related webpages that are organized by health topics and disease types. Using the Web Analytics tool, we obtained 10 million CVD-related anonymized search queries originating from Web search engines that "land on" CVD webpages within MayoClinic.com and are related to CVD. These queries are in English language and are collected between September 2011-August 2013. Our final analysis dataset consists of 10,408,921 CVD related search queries, which is a significantly large dataset for a single class of diseases.

Data Analysis: We performed the following analysis on the CVD related search queries: 1) Top search queries associated with CVD; 2) Semantic analysis: categorization of the queries into health categories using UMLS MetaMap; 3) Structural analysis: length of the search queries in number of words and number of characters, usage of search query operators (such as 'and' and 'or') and special characters in the search queries; and 4) Textual analysis:

types of search queries (keyword based, Wh-questions, Yes/No questions), misspellings in the queries, and linguistic structure of the search queries (part-of-speech analysis).

1. Top CVD search queries: We selected the 20 most frequently searched queries from the analysis dataset.

2. Categorization of the queries into health categories using UMLS MetaMap

2.1. Selection of health categories: There are many possible health categories of interest. In this work, we selected 14 health categories **(Table. 1)** that are "consumer oriented" as well as can reveal details about what OHISs generally search for in the context of CVD. Here, we define "consumer oriented" health categories as categories that are easily understandable for a non-expert, lay population. While selecting the health categories, we studied the health categories on popular health websites (e.g., Mayo Clinic, WebMD, etc.) and the types of information frequently mentioned along with CVD search queries, e.g. vital signs (blood pressure, heart rate), age groups (infants, adult, elder), etc. Note that there can be possible overlaps between some of the health categories, for example 'Drugs and Medications' can be considered as a part of 'Treatment', but in our analysis we considered both as a separate health categories in order to study search traffic for each category separately. These categories and the categorization scheme **(Table. 1)** is reviewed and verified by the Mayo Clinic clinicians and domain experts.

Table 1. List of health categories and their description with examples

Categories	Description and Examples
Symptoms	Search queries related signs and symptoms, e.g. Stroke symptoms, heart palpitations with headache, home remedies for heart murmur, heartburn vs heart attack symptoms
Causes	Search queries related to cause/reasons for various CVD conditions, symptoms, e.g. causes of an elevated heart rate, heart failure reasons, morning hypertension causes
Risks and Complication	Search queries related to risk and complications, e.g. risks of pacemaker, risk factors to hypertension, complications of bypass surgery, heart ablation surgery risks
Drugs and Medications	Search queries related to drugs and medications, e.g. dextromethorphan blood pressure, medications pulmonary hypertension, tylenol raise blood pressure, ibuprofen heart rate
Treatments	Search queries related to treatments, e.g. exercise for reducing hypertension, cardiac arrest treatments, dilated cardiomyopathy treatment, bypass surgery, cardiac rehabilitation
Tests and Diagnosis	Search queries related to tests and diagnosis, e.g. heart echocardiogram, diagnosis of vascular disease, ct scan for heart, test for cardiomyopathy, urinalysis in hypertension
Food and Diet	Search queries related to food and diet, e.g. what is cardiac diet, what foods lower blood pressure and cholesterol, red wine heart disease, alcohol and hypertension
Living with	Search queries related to control, management, cure and living with CVD, e.g. exercises to lower high blood pressure, controlling blood pressure naturally, cure for postural hypotension, lifestyle changes to lower hypertension, living with pacemaker, how to control cholesterol
Prevention	Search queries related to prevention, e.g. ways to prevent heart attack, preventing stroke, foods to avoid heart diseases, aspirin for prevention of stroke, foods that lower risk of heart disease
Side effects	Search queries related to side effects, e.g. blood pressure pills side effects, side effects of beta blockers for hypertension, coq10 bp side effects, side effects of pulmonary hypertension
Medical devices	Search queries related to medical device references, e.g. living with a pacemaker, using blood pressure cuff, pump for pulmonary hypertension, blood pressure monitor
Diseases and conditions	Search queries related to diseases and conditions, e.g. medications pulmonary hypertension, born with holes in heart, stroke tia symptoms, hypotension, heart attack in pregnancy
Age-group References	Search queries with references to age groups, e.g. cardiac defects in children, average heart rate for an adult, heart murmur in adult, hypertension in adolescents, heart murmurs in infants
Vital signs	Search queries with references to blood pressure, heart rate, pulse rate, temperature, heart beat (without high/low blood pressure as we considered them under 'Diseases and Conditions'), e.g. blood pressure 125/90, normal resting heart rate, can tylenol raise blood pressure, healthy heart rate chart, sleep and blood pressure

2.2. Mapping CVD search queries to UMLS semantic types and concepts: We performed semantic analysis on the CVD search queries by mapping all the search queries from the dataset to UMLS concepts and semantic types using UMLS MetaMap[7]. MetaMap is a tool for recognizing UMLS concepts in the text. For a given search query, MetaMap identifies one or more UMLS concepts, their semantic types, Concept Unique Identifiers (CUIs), and other details. UMLS incorporates variety of medical vocabularies and concepts, and maps each concept to semantic types. Thus using UMLS, we can understand 'semantics' or 'meaning' of the words and phrases. For example: in

search query 'heart attack red wine' semantics is used to understand that the query is about two separate phrases 'heart attack' and 'red wine' and not about 4 separate words 'heart', 'attack', 'red' and 'wine'. Moreover using semantics facilitated by UMLS, we can understand "heart attack" is a disease or syndrome (DSYN) and "red wine" is a "Food item" (FOOD). Refer to Table 2 for semantic type abbreviations used in this paper.

Table 2. List of health categories and their respective UMLS semantic types/concepts used categorization. **Abbreviations**: SOSY-Signs and Symptoms, ORCH-Organic Chemical, PHSU- Pharmacologic Substance, CLND-Clinical Drug, TOPP- Therapeutic or Preventive Procedure, FTCN-Functional Concept, CNCE-Conceptual Entity, DIAP- Diagnostic Procedure, LBPR-Laboratory Procedure, LBTR- Laboratory Test Result, FOOD-Food, MEDD-Medical Device, DSYN-Disease or Syndrome, AGGP-Age Group

Health Categories	UMLS Semantic Types (**ST**), UMLS Concepts (**CC**) and Keywords (**KW**)
Symptoms	**ST**: SOSY **CC:** symptoms, signs, heart murmur
Causes	**CC**: cause, reason
Risks & Complications	**CC**: risk, complications
Drugs and Medications	**ST**: ORCH|PHSU, CLND, PHSU **CC**: medication, medicine, drugs, dose, dosage, tablet, pill **KW**: meds (without CC: alcohol, caffeine, fruit, prevent)
Treatments	**ST**: TOPP, FTCN (treatment, surgery), CNCE (treatment), **CC**: remedy, remediate (without CC: prevention and 'Drugs and Medication' queries)
Tests and Diagnosis	**ST**: DIAP, LBPR, LBTR **CC**: Test, diagnosis (without ST: DIAP| TOPP, CC: alcohol, blood caffeine)
Food and Diet	**ST**: FOOD **CC:** caffeine, recipe, meal, menu, diet, eat, breakfast, lunch, dinner, alcohol, drink
Living with	**CC:** control, manage, reduce, lower, coping, cure, recover **KW:** living with, bring down, low down
Prevention	**CC:** prevent, avoidance, low risk
Side effects	**CC:** side effect **KW:** side effect
Medical devices	**ST**: MEDD
Diseases and conditions	**ST**: DSYN **CC:** arrhythmia, avascular necrosis, enlarged heart, hypotension, blood pressure low **KW**: heart damage
Age-group References	**ST**: AGGP
Vital signs	**CC:** blood pressure, heart rate, pulse rate, temperature, Heart beat (without high/low blood pressure as we considered them under 'Diseases and Conditions')

2.3. Categorization Approach (**Table 2**): We categorized the search queries into 14 health categories as following:

1) UMLS has 140 semantic types and some of them are directly mapped to health categories that we selected; for example, 'AGGP' (Age-group) semantic type is directly mapped to the 'Age group' category. In this case, we categorized all the search queries with semantic type 'AGGP' into the 'Age group' category.
2) For a few health categories (e.g., 'Test and Diagnosis') we utilized multiple semantic types ('DIAP', 'LBPR', 'LBTR'). In this case, we categorized all the queries with at least one semantic type ('DIAP', 'LBPR', 'LBTR') into the 'Test & Diagnosis' category.
3) For a few health categories (e.g., 'Food and Diet'), there are certain concepts that are closely associated with the health category are not mapped to the selected semantic type. In such cases, we utilized both semantic types and well as semantic concepts for the categorization. For example, 'FOOD' semantic type does not include concepts such as 'meal', 'menu', 'diet', 'recipe' and 'lunch' as they are not actually food items. We categorized all the search queries that have 'FOOD' semantic type or at least one concept from (meal', 'menu', 'diet', etc.) into the 'Food and Diet' category.
4) For a few health categories (e.g., 'Cause') there is no directly associated semantic type. In such cases, we utilized semantic concepts for the categorization. For example, we categorized all the search queries which have either 'Cause' or 'Reason' semantic concepts into the 'Cause' category.
5) For a few health categories (e.g., 'Living with'), apart from semantic concepts, we also considered the presence of keywords ('Living with') within the search query as 'Living with' is not a concept in the UMLS.
6) Few semantic types include some undesired concepts (in the context of our customized categorization, not in the terms of UMLS concept hierarchy). For example, semantic types 'ORCH\| PHSU' and 'PHSU' are associated with the 'Drugs and Medication' category. These semantic types include some concepts that are not considered as drugs to a consumer/lay population: caffeine, fruit, prevent, etc. In such cases, we do not categorize the search

queries with semantic types 'ORCH/PHSU' or 'PHSU' and with semantic concepts caffeine, fruit, prevent, etc. into the 'Drugs and Medication' category.

7) We also considered lexical variants, as well as partial matches, of some concepts for example: diagnose, diagnosis, test, testing, etc. A search query can be categorized into zero, one or more than one health category depending on the mapping of the query to UMLS concepts and semantic types. We empirically evaluated queries in each category and performed several iterations to evaluate the semantic type/concepts for each category then defined the categorization scheme **(Table 2)**.

2.4. Categorization Evaluation: We evaluated the performance of the categorization approach as following:

1) Gold standard dataset creation: We randomly selected 1000 search queries from the analysis dataset. Two domain experts manually annotated 1000 search queries by labeling one search query with zero, one, or more than one health category. The annotators first discussed and agreed upon the annotation scheme. To reduce the probability of human errors and subjectivity, the two annotators discussed together and annotated each query and created a gold standard dataset with 1000 search queries.
2) Precision-Recall calculation: We categorized 1000 search queries from the gold standard dataset using the categorization approach as discussed in Section 2.3 and evaluated the categorization approach with respect to the gold standard dataset. Since we categorized search queries into 14 health categories, we also calculated Micro average Precision and Recall. Based on the evaluation, our categorization approach has very good Precision: 0.8842, Recall: 0.8642 and F-Score: 0.8723.
3) We also performed Precision and Recall analysis for each health category independently (Table 3) to check the performance of the categorization approach for individual health categories. The categorization approach works well for most of the categories while for a few categories the approach shows above average Precision/Recall. One observed reason that affected the Precision/Recall is multiple interpretations of the concepts that sometime may not be contextually correct, e.g. for the search query 'nuts good for your heart', MetaMap annotated 'nuts' as 'FOOD' as well as 'MEDD' (Nut - Medical Device Component or Accessory).

Table 3. List of health categories and their respective Precision and Recall

No	Categories	Precision	Recall	F1 Score	No.	Categories	Precision	Recall	F1 Score
1	Symptoms	0.9274	0.8042	0.8614	8	Living with	0.8659	0.9342	0.8988
2	Causes	0.8861	0.9859	0.9333	9	Prevention	0.8333	1.0000	0.9091
3	Risks and Complication	1.0000	1.0000	1.0000	10	Side effects	1.0000	1.0000	1.0000
4	Drugs and Medications	0.8582	0.9350	0.8950	11	Medical devices	0.8077	0.7500	0.7778
5	Treatments	0.7083	0.9444	0.8095	12	Diseases	0.9291	0.7751	0.8451
6	Tests and Diagnosis	0.6389	1.0	0.7797	13	Age-group References	1.0000	0.8889	0.9412
7	Food & Diet	0.9391	0.9558	0.9474	14	Vital signs	0.8872	0.8669	0.8769
Overall Micro Average **Precision (0.8842), Recall (0.8607)** and **F1 Score (0.8723)**									

3. Health query length: We calculated search query length by computing the number of words (separated by white space) and the number of characters (excluding white space) in the search queries.

4. Usage of query operators and special characters: In search queries, query operators ('and', 'or', 'not', etc.) are used to formulate complex queries. In this study, we have considered following operators: AND, OR, +, &, other (NOT, AND NOT, OR NOT, & NOT). Special characters are the characters apart from letters (a-z) and digits (0-9). The significance of special characters in health search query depends upon the usage of special characters in the medical domain. For example, OHISs may mention values in different formats, e.g., 2.3 ml, 40%, 17-19, 125/90 (for blood pressure) or $200 (for the cost of a drug or procedure). We analyzed the usage of search query operators and special characters in the CVD search queries based on their usage frequency in the search queries.

5. Misspellings in health queries: OHISs occasionally make spelling mistakes while searching for health information. To analyze the frequency of such errors, we used a dictionary-based approach. We first generated a dictionary of words using the Zyzzyva wordlist[23], the Hunspell dictionary[24], and its medical version (OpenMedSpell[25]), comprising a total of 275,270 unique words. We used this dictionary to check misspellings in the CVD search queries.

6. Type of Search queries: OHISs express their health information need by formulating search queries on Web search engines. OHISs can express their information need either by formulating search queries using keywords or asking questions (Wh-questions and Yes/No questions). For this analysis, we have considered the following Wh-questions (lexicon): 'What', 'How', '?', 'When', 'Why' and others ('Who' 'Where', 'Which'). Note that, although '?' does not come under Wh-questions category, we have included it for the simplicity. Yes/No questions are usually used to check some factual information, for example, whether coffee is bad for the heart. In this analysis, we have considered Yes/No questions that start with 'Can', 'Is', 'Does', 'Do', 'Are', and others ('Could' 'Should', 'Will', 'Would'). Using the lexicon for Wh-questions and Yes/No questions, we performed text analysis on the search queries to count the number of queries with Wh-questions and Yes/No Questions. Search queries that do not contain any questions (Wh- or Yes/No) are classified as Keyword-based. Additionally, for different Wh-questions and Yes/No questions, we computed their usage frequency in the search queries.

7. Linguistic analysis of health queries: Linguistic structure of the search queries has implications on information retrieval using Web search engines[8]. Thus we analyzed basic linguistic characteristics of the CVD search queries. We performed part-of-speech analysis on the search queries using Stanford's POS tagger[26]. For this analysis, we considered nouns, verbs, adjectives and adverbs. We mapped all the subtypes in part-of-speech (e.g. proper nouns, common nouns, compound nouns) to the main part-of-speech types (e.g. nouns). We analyzed usage of different part-of-speech types in the CVD queries based on their usage frequency in the search queries.

RESULTS

1. Top health queries: Most of the top search queries are related to major CVD diseases and conditions. At the same time, questions about blood pressure (high/low) and heart rate were also searched frequently (Table 4).

Table 4. Top search queries related to CVD

Top 1-5 Queries	Top 6-10 Queries	Top 11-15 Queries	Top 16-20 Queries
heart attack symptom	congestive heart failure	cardiomyopathy	echocardiogram
blood pressure chart	low blood pressure	heart palpitations	heart disease
how to lower blood pressure	stroke symptoms	blood pressure medication	orthostatic hypotension
heart rate	normal blood pressure	symptoms of stroke	heart healthy recipes
broken heart syndrome	high blood pressure symptoms	heat stroke	heart arrhythmia

2. Health categories: Based on **Table 5**, the most popular health categories while searching for CVD information are 'Diseases and Conditions' and 'Vital signs'. One in every two searches is related to either 'Diseases and Conditions' or 'Vital signs'. Due to close association of vital sings (such as blood pressure and heart rate) with CVD, OHIS might be searching it frequently. Other popular health categories that users search for includes 'Symptoms', 'Living with', 'Treatments', 'Food and Diet' and 'Causes'. Mostly due to the chronic nature of the CVD and as the patients are in charge of managing the disease with day-to-day care, many CVD patients might be searching for 'Living with' related search queries. As diet has a significant impact on the CVD, we observed large search traffic for 'Food and Diet' category. Many OHISs are also interested in learning about CVD 'Treatments', 'Medical Devices' (e.g. pacemaker), 'Drugs and Medication', and 'Cause'. Although CVD can be prevented with some lifestyle and diet changes, interestingly very few OHISs search for CVD 'Prevention'.

Table 5. Categorization of CVD search queries into 14 health categories

No	Health categories	Total Queries	Percentage Distribution	No	Health categories	Total Queries	Percentage Distribution
1	Diseases	4,232,398	28.66	8	Drugs and Medications	603,905	4.09
2	Vital signs	3,455,809	23.40	9	Causes	599,895	4.06
3	Symptoms	1,422,826	9.64	10	Tests & Diagnosis	344,747	2.33
4	Living with	1,178,756	7.98	11	Risks and Complication	277,294	1.88
5	Treatments	955,701	6.47	12	Prevention	136,428	0.92
6	Food and Diet	779,949	5.28	13	Age-group References	87,929	0.60
7	Med Devices	665,484	4.51	14	Side effects	25,655	0.17
					Total	14,766,776	100

Using our categorization approach, we categorized 92% of the 10 million CVD related queries into at least one health category (**Table 6**). Most of the queries (around 88%) are categorized into either one or two categories (Table 6). Very few CVD queries (4.28%) are categorized into 3 or more categories. Our approach did not categorize 8.13% of the queries into any health categories. After studying the uncategorized search queries, we found that there are few queries that do not fit into any of the selected 14 categories such as cardiac surgeon, cardiology mayo, video on cardiovascular, pediatric cardiology, and orthostatic.

Table 6. A search query can be categorized into zero, one, or more health categories. The table shows the distribution of search queries by number of health categories in which they are categorized.

Number of health Categories	Number of search queries	Percentage Distribution
0	845,744	8.13%
1	4,967,337	47.72%
2	4,149,803	39.87%
3	420,622	4.04%
4 and 5	25,415	0.24%
Total	10,408,921	100.00%

3. Health query length: Average search query length (**Figure 2**) for CVD is 3.88 words and 22.22 characters. Around 80% of the CVD search queries have 3 or more words. The analysis implies that, CVD search queries are longer than previously reported non-medical, as well as medical queries, as the average length for both of them is around 2.35 words[15,27]. This potentially indicates that OHISs describe their CVD information needs in more detail by adding relevant health context to the search query. Longer search queries also denote OHIS' interest in more specific information about the disease; subsequently OHISs use more words to narrow down to a particular health topic. Another possible reason for longer CVD search queries might be that even simple CVD related search queries have multiple words (Table 1, Table 4).

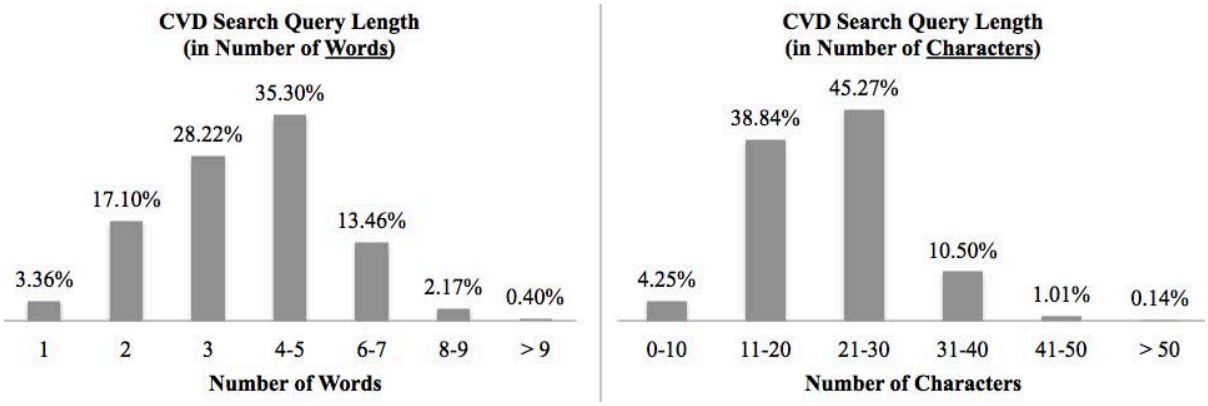


Figure 2. Distribution of length of search queries by number of words and number of characters

Table 7. Search query operators usage, special characters usage, and misspellings in the CVD search queries

Structural Analysis	Usage Frequency	Total Queries	Percentage Distribution
Number of Query Operators	0	10,011,257	96.18%
	>0	397,664	3.82%
Query Operators Usage	AND	366,117	92.07%
	+	7,115	1.79%
	OR	16,063	4.04%
	&	4,159	1.05%
	Other	4,210	1.06%
Special Characters	0	10,288,916	98.85%
	>0	120,005	1.15%
Spelling Mistake	0	10,075,665	96.80%
	>0	333,256	3.20%

4. Query operators usage, special characters usage, and misspellings in the CVD search queries (Table 7): Around 4% of CVD search queries use at least one query operator. 'AND' is the most popular operator (92%), followed by 'OR' (4%) and '+' (1.7%). Overall variations of 'and' (AND, &, +) operators comprise around 95% of operator usage in the search queries. OHISs formulate very few (1%) CVD search queries with special characters. In CVD search queries, 3.2% of the queries have at least one spelling mistake. Web search engine's "auto-completion" as well as "spelling correction" functionalities might be one reason to lower misspellings in the search queries.

5. Type of Search queries: As indicated by the analysis in **Figure 3**, OHISs predominantly formulate search queries using keywords (80%), though queries with Wh-Questions are also significant. Few queries (2.5%) are formulated as Yes/No type questions. In Wh-questions, OHISs mostly use "How" and "What" in the search queries and both of them generally signify that more descriptive information is needed. In Yes/No Questions, OHISs more often start the search queries with "does" "can" and "is".

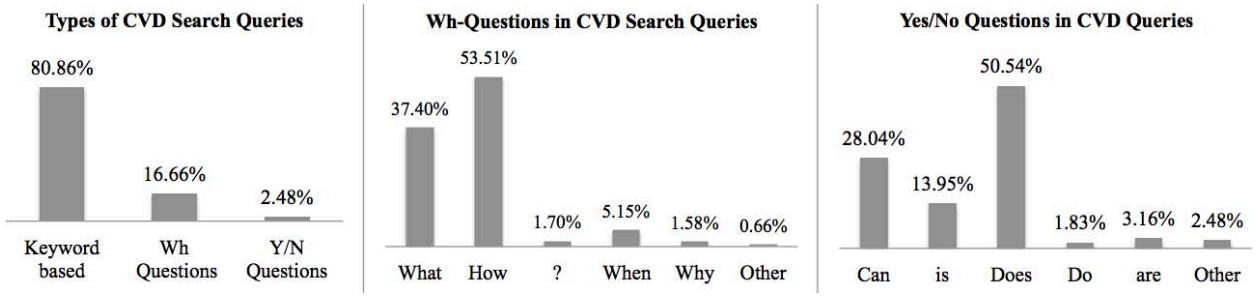


Figure 3. Types of search queries (how health information need is expressed) and distribution of Wh-Questions and Yes/No Questions based on frequency of their usage in the CVD search queries.

6. Linguistic analysis (Table 8): In health search queries, nouns typically denote entities like disease names, health categories, etc. Almost all the CVD search queries have at least one noun and most of the search queries (81.5%) have 2 or 3 nouns. A verb conveys an action or an occurrence, for example "how to control (verb) hypertension (noun)". Approximately 26% of the search queries have at least one verb. Adverbs are the words that modify a verb, an adjective and another adverb, while an adjective is a 'describing' word, giving more information about the object signified, for example "extremely (adverb) bad (adjective) heart (noun) pain (noun)." Very few search queries have at least one adverb and 47% of the queries have at least one adjective.

Table 8. Linguistic analysis (part-of-speech) on CVD search queries

Part-of-Speech	Usage Frequency	Total Queries	Percentage Distribution
Nouns	**0**	6,279	0.06%
	1	1,075,467	10.33%
	2	4,964,074	47.69%
	3	3,523,204	33.85%
	>3	839,897	8.07%
Verb	**0**	7,722,752	74.19%
	>0	2,686,169	25.81%
Adverb	**0**	10,151,641	97.53%
	>0	257,280	2.47%
Adjective	**0**	5,519,489	53.03%
	>0	4,889,432	46.97%

DISCUSSION

In this study, we analyzed a significantly large dataset of 10 million CVD related search queries in order to understand online health information searching for CVD. We implemented a rule based categorization approach (with Precision: 0.8842, Recall: 0.8607and F1 Score: 0.8723) using UMLS concepts/semantic types and categorized 92% of the 10 million CVD related search queries into 14 "consumer oriented" health categories. As per our analysis, the top searched health categories ("information needs") for CVD are 'Diseases and Conditions', 'Vital Sings', 'Symptoms', and 'Living with'. Other frequently searched CVD health categories are 'Treatments', 'Food

and Diet', and 'Causes'. Most of the queries (around 88%) are categorized into either one or two health categories. Even though CVD can be prevented with some lifestyle and diet changes, very few OHIS search for preventive health information. We found that use of MetaMap and UMLS concepts/semantic type to be a very good approach for categorization of the health related search queries as UMLS incorporates variety of medical vocabularies and concepts, and mapping of each concept to semantic types. However for customized categorization, we have to carefully select/eliminate UMLS semantic types and concepts considering the alignment of their scope with desired categories.

Our study reveals some interesting insights about structural and syntactic properties of CVD search queries. The average length of CVD search queries is longer than that of previously reported general search queries as well as medical search queries[15,27]. Our study implies OHISs may be interested in more specific information. Only 3.2% of the search queries contain at least one spelling mistake. The search engine's "auto-completion" feature, and spelling correction/suggestion functionality might be contributing to reduce misspelled words in search queries. Very few OHISs use search query operators and variations of 'and' (AND, &, +) operators comprise around 95% of operator usage in the search queries. OHISs formulate search queries primarily using keywords (around 80%), followed by Wh-Questions, and Yes/No Questions. In Wh-questions, OHISs mostly use 'What' and 'How' in the search queries, and both of them generally signify a need for more descriptive information, while search queries in the form of Yes/No questions indicate interest in factual information. Almost all the search queries have one noun. OHISs also use adjectives and verbs frequently in the search queries to add context to the topic of interest.

Following are some of the limitations of this study. The results of this study are derived from the analysis limited to CVD search queries from Web search engines that led users to MayoClinic.com. Even though Mayo Clinic web pages often ranked high in Web search engines, not all health information seekers visit MayoClinic.com. The focus of this study is limited to analysis of the search query log and we have not analyzed associated socioeconomic factors due to the anonymized nature of the data. To the best of our knowledge, there is not much research on understanding online health information searching for chronic diseases and especially for CVD. This study addresses this knowledge gap and extends our knowledge about online health information search behavior. The study provides interesting and valuable insights that can further be leveraged in multiple ways, such as:

1. Web search engines: to understand details about structural and linguistic characteristics of health search queries, search query complexity, and popular health categories in order to improve health information retrieval systems;
2. Websites that provide health information: to better understand an OHIS's health information need, and do a better organization of health information content;
3. Healthcare providers: to better understand their patients and their health information interest;
4. Healthcare-centric application developers: to better understand what and how OHISs search for and to build applications around consumer health information needs and priorities;
5. OHISs: we anticipate that this work will help empower OHISs in their quest for health information, and facilitate their health information search efforts by enabling the development of smarter and more sophisticated consumer health information delivery mechanisms.

In the future, we plan to leverage insights from this work to facilitate a better health search experience by developing more advanced next-generation knowledge and content delivery systems. Also, we plan to perform comparative analysis of major diseases to learn similarities and differences between them in online health information searching.

Conclusion

We presented a comprehensive analysis on CVD related search queries in order to understand what users search for ("information need") and how they formulate search queries ("expression of information need"). We found that using MetaMap and UMLS concepts/semantic type is a very good approach for categorization of health related search queries into health categories. The categorization approach can be reused for different set of health categories by defining new association rules**.** Distribution of the CVD queries by health categories indicates that, OHISs have most information needs for CVD related 'Diseases and Conditions', 'Vital Signs' 'Symptoms' and post CVD information ('Living with', 'Diet', 'Treatments, Drugs'). OHISs predominantly formulate search queries using keywords followed by Wh-Questions and Yes/No Questions. Almost all CVD search queries have at least one noun. A greater understanding of OHIS's needs may help us to accomplish the changes that will lead to improvement in online health information searching and a more balanced approach for health information intervention. This study extends our knowledge about online health information searching, and provides useful insights for Web search

engines, health-centric websites and application developers. Finally, we anticipate that this work will help empower OHISs in their quest for health information, and facilitate their health information search efforts by enabling the development of more advanced next-generation knowledge and content delivery systems.

Acknowledgement: We thank the Mayo Clinic Web Analytics team for their valuable contribution in data provision. This work is supported by the Mayo Clinic NIH Relief Fund Award (FP00068008). We acknowledge Rashmi Dusane for her help in this work.

References

1. Fox S, Duggan M. Health online 2013. Pew internet & American Life Project 2013.
2. Higgins O SJ, Barry MM, Domegan C. A literature review on health information seeking behaviour on the web: a health consumer and health professional perspective. In: ECDC, ed. Stockholm2011.
3. National Center for Health Statistics. Health, United States, 2010: With special feature on death and dying Hyattsville, MD. 2011.
4. Ayers SL, Kronenfeld JJ. Chronic illness and health-seeking information on the Internet. Health:. 2007;11(3):327-347.
5. Fox S, Duggan M. Who Lives with Chronic Conditions Pew internet & American Life Project. 2013.
6. Mayo Clinic's consumer health information website. http://www.mayoclinic.com/ Accessed March 9, 2014
7. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Paper presented at: Proceedings of the AMIA Symposium2001.
8. Croft WB et.al. Search engines: Information retrieval in practice. Addison-Wesley Reading; 2010.
9. Drentea P, Goldner M, Cotten S, Hale T. The association among gender, computer use and online health searching, and mental health. Information, Communication & Society. 2008;11(4):509-525.
10. Weaver III JB, Mays D, Weaver SS, Hopkins GL, Eroglu D, Bernhardt JM. Health information-seeking behaviors, health indicators, and health risks. American journal of public health. 2010;100(8):1520-1525.
11. Atkinson NL, Saperstein SL, Pleis J. Using the internet for health-related activities: findings from a national probability sample. Journal of Medical Internet Research. 2009;11(1).
12. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. 2009;457\(7232\):1012\-1014\.
13. Herskovic JR, Tanaka LY, Hersh W, Bernstam EV. A Day in the Life of PubMed: Analysis of a Typical Day's Query Log. Journal of the American Medical Informatics Association. 3// 2007;14(2):212-220.
14. Cartright M-A, White RW, Horvitz E. Intentions and attention in exploratory health search. SIGIR2011.
15. Spink A, Yang Y, Jansen J, et al. A study of medical and health queries to web search engines. Health Information & Libraries Journal. 2004;21(1):44-51.
16. Yang CC, Winston F, Zarro MA, Kassam-Adams N. A study of user queries leading to a health information website: AfterTheInjury. org. Proceedings of the 2011 iConference: ACM; 2011:267-272.
17. White RW, Horvitz E. From health search to healthcare: explorations of intention and utilization via query logs and user surveys. Journal of the American Medical Informatics Association. 2013.
18. Jadhav A, Andrews D, et al. Comparative Analysis of Online Health Queries Originating From Personal Computers and Smart Devices on a Consumer Health Information Portal J Med Internet Res 2014;16(7):e160
19. White RW, Horvitz E. Cyberchondria: studies of the escalation of medical concerns in web search. ACM Transactions on Information Systems (TOIS). 2009;27(4):23.
20. Winkleby MA et al. Socioeconomic status and health: how education, income, and occupation contribute to risk factors for cardiovascular disease. American journal of public health. 1992;82(6):816-820.
21. Celler BG, Lovell NH, Basilakis J. Using information technology to improve the management of chronic disease. Medical Journal of Australia. 2003;179(5):242-246.
22. Lorig KR, Ritter PL, Laurent DD, Plant K. Internet-based chronic disease self-management: a randomized trial. Medical care. 2006;44(11):964-971.
23. Zyzzyva: The Last Word in Word Study http://www.zyzzyva.net/wordlists.shtml Accessed March 9, 2014.
24. Hunspell dictonary. http://hunspell.sourceforge.net/. Accessed March 9, 2014.
25. OpenMedSpel for Hunspell http://www.e-medtools.com/Hunspel_openmedspel.html Accessed March 9, 2014.
26. Toutanova K, Klein D, Manning CD, Singer Y. Feature-rich part-of-speech tagging with a cyclic dependency network. Paper presented at: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 12003.
27. Spink A, Wolfram D, Jansen MB, Saracevic T. Searching the web: The public and their queries. Journal of the American society for information science and technology. 2001;52(3):226-234.

Characterization of a Handoff Documentation Tool Through Usage Log Data

Silis Y. Jiang[1], Alexandrea Murphy[1] MPH, David Vawdrey[1] PhD, R. Stanley Hum[2] MD, Lena Mamykina[1] PhD
[1]Department of Biomedical Informatics, Columbia University
[2]Department of Pediatrics, Columbia University

Abstract

Handoffs are a critical component of coordinated patient care; however, poor handoffs have been associated with near misses and adverse events. To address this, national agencies have recommended standardizing handoffs, for example through the use of handoff documentation tools. Recent research suggests that handoff tools, typically designed for physicians, are often used by non-physician providers as information sources. In this study, we investigated patterns of edits of an electronic handoff tool in a large teaching hospital through examination of its usage log data. Qualitative interviews with clinicians were used to triangulate log data findings. The analysis showed that despite its primary focus on facilitating transitions of care, information in the handoff documentation tool was updated throughout the day. Interviews with residents confirmed that they purposefully updated information to make it available for other members of their patient care teams. This further reiterates the view of electronic handoff tools as facilitators of team communication and coordination. However, the study also showed considerable variability in the frequency of updates between different units and across different patients. Further research is required to understand what factors drive such diversity in the use of electronic handoff tool and whether this diversity can be used to make inferences about patients' conditions.

Introduction

Handoffs are a critical aspect of providing continuous care for patients in inpatient services[1]. Due to recent changes towards more restrictive resident work hours, handoffs have become commonplace in the hospital[2]. However, poor handoffs have been associated with near misses and adverse events[3,4]. Institutions such as the Joint Commission and the Accreditation Counsel of Graduate Medical Education (ACGME) have made recommendations towards improving handoffs[5] and require hospitals to ensure resident competency in handoffs[6]. These identified limitations of the current handoff practices suggest the need for new tools for facilitating handoff. With few nationally published or commercially available handoff tools, many healthcare institutions developed and optimized handoff tools to their own needs and workflows[7]. While many of these tools are paper based, there is a growing trend towards developing electronic and often EHR-integrated handoff documentation systems[8-10]. There is emerging evidence that interventions that include handoff documentation tools reduce errors in patient care and improve the quality and structure of handoffs[11].

Previous research on handoff tools has primarily focused on the impact of these tools on handoff standardization or patient care[12]. Furthermore, these previous studies usually adopted a singular user perspective, typically focusing on physicians[12,13]. At the same time, an emerging theme from recent handoff tool research suggests that electronic resident handoff documentation tools are adopted beyond their original user group. For example, Vawdrey et al. showed that non-physician clinicians (nurses, pharmacists, social workers, etc.) accounted for 60% of active users in a resident handoff documentation tool[7]. Similarly, Schuster et al. reported that the physician handoff tool was integrated into the daily workflow of non-physicians[14]. However, these studies have focused on handoff tools from a user viewing or information retrieval perspective. The process of creating or updating handoff documents is much less understood.

This study focuses on investigating patterns of use of a resident oriented handoff documentation tool, with a particular attention to creation of the handoff document, including edits and updates. The study was conducted at Columbia University Medical Center, where clinicians use a handoff documentation tool known as Handoff Tab. The tool's interface includes a series of labeled free-text boxes that can be directly edited by authorized users. Additionally, clinicians have the option to print the user-contributed data along with recent structured data, such as lab values and medication list. The tool is primarily edited by residents but available to all clinicians to view. This study sought to identify: 1) whether there are any systematic temporal patterns in the edits to the different fields of

the Handoff Tab; 2) whether there are any differences in the frequency and patterns of updates between different clinical units; and 3) whether there were any differences in how frequently Handoff Tab was updates across different patients.

Materials and Methods

Handoff Tab

Handoff Tab is a custom-designed module included in the commercial EHR system (Sunrise Eclypsis). It has 9 free-text boxes: Active Issues, Consult Notes, Contact Info, Coverage To Do List, Discharge Planning, Hospital Course, Notes and Comments, Patient Summary, and Primary To Do List. The tool also provided the functionality to produce a printable report that included structured data, such as labs and medication. Furthermore, users could choose to print a cover sheet that summarized select patients.

Dataset

To analyze patterns in Handoff Tab documentation, we used the document-editing event logs for all of October 2013 at the Columbia University Medical Center campuses of New York Presbyterian Hospital. The date and time of each edit event, which free-text box was edited, patient medical record number, patient bed location, clinician user ID, and care provider role were extracted from the audit log. All analyses on the unit level were normalized based on the number of unique patients in that unit. It is important to note, however, that not all patients in a unit may have Handoff Tab documents associated with them. Therefore, there may actually be more patients in the unit at a given time than can be estimated by the dataset. Analyses on the patient level were normalized based on the number of handoff days. The number of handoff days was calculated based on the first and last day a clinician edited the Handoff Tab document of each patient.

Additionally, four interviews were conducted with the medical staff in the pediatric intensive care unit at Columbia University Medical Center (CUMC). Three residents and one attending were interviewed. The interviewees were selected by convenience. The three residents were in their final year of residency, while the attending physician had previously used a handoff documentation tool during the final year of critical care fellowship training.

Quantitative Analysis

To identify editing patterns throughout the workday, all edit events were aggregated into a single 24-hour period and visualized, regardless of the day of the month.

To understand the influence of unit practices on editing patterns, the dataset was divided by units and then clustered. First, the total number of edits and the number of edits per Handoff Tab Field was calculated for each unit. These values were then normalized based on the patient count for individual units. Euclidean distance was used to measure the difference of each unit based on the edit frequency of individual fields. Hierarchical clustering based on the complete method was then used to group the units. A *k*-value was selected based on the outcome. This *k*-value was then applied to a *k*-means clustering algorithm (Hartigan and Wong method) of each analysis dataset.

To understand the influence of patients on editing patterns, the dataset was split by unique patients and then clustered. A similar method similar to the unit cluster analysis was used. Briefly, edits per Handoff Tab field was calculated on a patient basis and then normalized based on the handoff days. A distance matrix was used to calculate the Euclidean distance between each patient in the dataset. Hierarchical clustering based on the complete method was used to group the patients based on that matrix. A *k*-value was extracted from the hierarchical clustering analysis and applied to the k-means clustering algorithm (Hartigan and Wong method).

The statistical tool R was used for cluster data analysis and visualization. All clustering related methods were performed using the *stats* package, and all visualizations were created using the *ggplot2* package. All statistics are presented as means along with the standard error of the mean.

Qualitative Analysis

To supplement the quantitative data, interviews with physicians were conducted to provide insight and reasoning for unit differences in documentation behaviors. Four interviews were conducted with the medical staff in the pediatric intensive care unit at Columbia University Medical Center (CUMC). Three residents and one attending were interviewed. The interviewees were selected by convenience. The clinicians were consented prior to the start of the interview. The interviews were recorded and transcribed for analysis. Transcripts were reviewed for common themes.

This study was approved by the Columbia University Institutional Review Board.

Results

Temporal Distribution of Handoff Tab Updates

On average, each unit updated the Handoff Tab 13.67 times per patient during the study period with a standard error of the mean (SEM) score of 0.21 edits. The patients in this dataset had an average stay of 5.12 days (SEM = 0.007 days) and an average of 22.42 edits to their handoff document during their stay (SEM = 0.003 edits).

Table 1. Dataset Characteristics

Characteristic	**Sample size**
Total Number of Edit Entries During Study Period	145,872
Total Number of Units	132
Total Number of Patients	6515

After aggregating the dataset into a single 24-hour time frame, the data showed that Handoff Tab was updated throughout the day (Figure 1). However, not all fields were utilized equally. In general, the field “Consult Notes” were updated the least and the “Primary To Do List” field was updated the most. Updates were most frequently made during two timeframes corresponding to the periods right before handoff. However, the data also shows that Handoff Tab was updated throughout the workday.

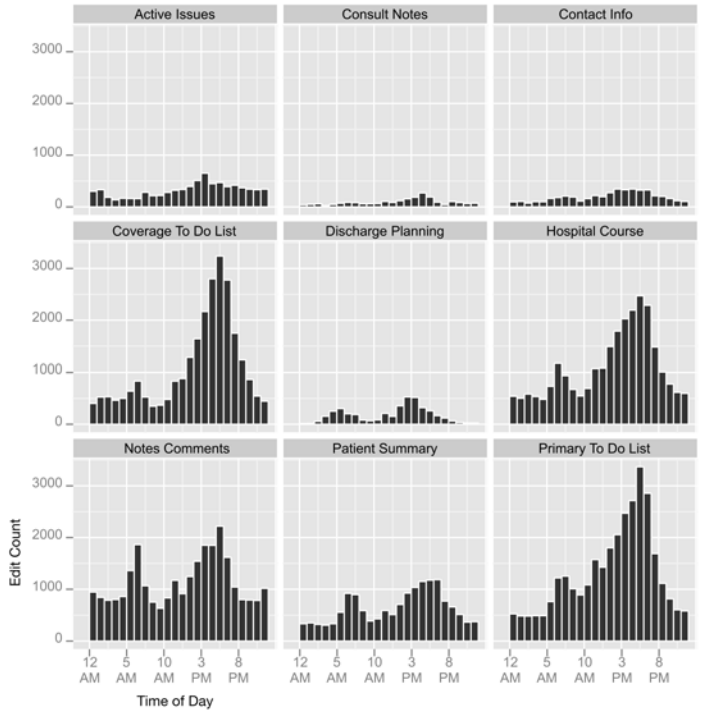


Figure 1. Hourly distributions of edit frequencies per Handoff Tab field

During interviews clinicians expressed awareness of other clinicians utilizing the tool and their notes as information sources. For this reason, residents reported updating the Handoff Tab up to three or four times per shift to make sure the information remain up to date.

> *"If we have down time and I have time to update during the day, I do that as well just so it is up to date for any one who refers to it and sometimes I know our nurses refer to it and the attendings refer to it." – Resident 1*
>
> *"Yeah. I'm usually – I usually keep track of things on my paper printout but then probably like three or four times a day I would do that [updating Handoff Tab throughout the day]." – Resident 3*

In these examples, the residents indicate a conscious effort to regularly maintain the information found in Handoff Tab. The desire to maintain an up-to-date handoff document provides one explanation for continuous updates throughout the day.

Unit Cluster Analysis

To understand the update pattern of the different units, a cluster analysis was performed based on individual unit update frequencies. Hierarchical clustering was used to identify possible cluster numbers for *k*-means clustering. In this case, a *k*-value of 5 was selected based on the hierarchical clustering (Figure 2A and 2B). The units in each cluster are differentiated by two factors: update frequency magnitude and field update distribution. Cluster 1 is characterized by having well above median update frequency magnitude. Cluster 1 has a unique distribution in its frequency peak for the "Notes and Comments" field, but the distribution also shows the "Patient Summary" field being edited more frequently than the "Hospital Course" field. Cluster 2 has well below median update frequency magnitude. Units in this cluster made less than 0.5 updates per patient in all of the fields. Furthermore, the distribution of updates per patient is relatively flat. Cluster 3 has the median frequency magnitude. For this cluster, the distribution of edits is appears to be roughly equal between the "Hospital Course", "Notes and Comments", and "Primary To Do List" fields. Cluster 4 has below median magnitude and roughly equal distributions between the "Coverage To Do List", "Hospital Course", and "Primary To Do List" fields. Cluster 5 has above median update frequency magnitude. The update distribution for this cluster is primarily focused on the "Notes and Comments" section and higher update frequency for ""Hospital Course" compared to "Patient Summary" (the opposite trend from cluster 1).

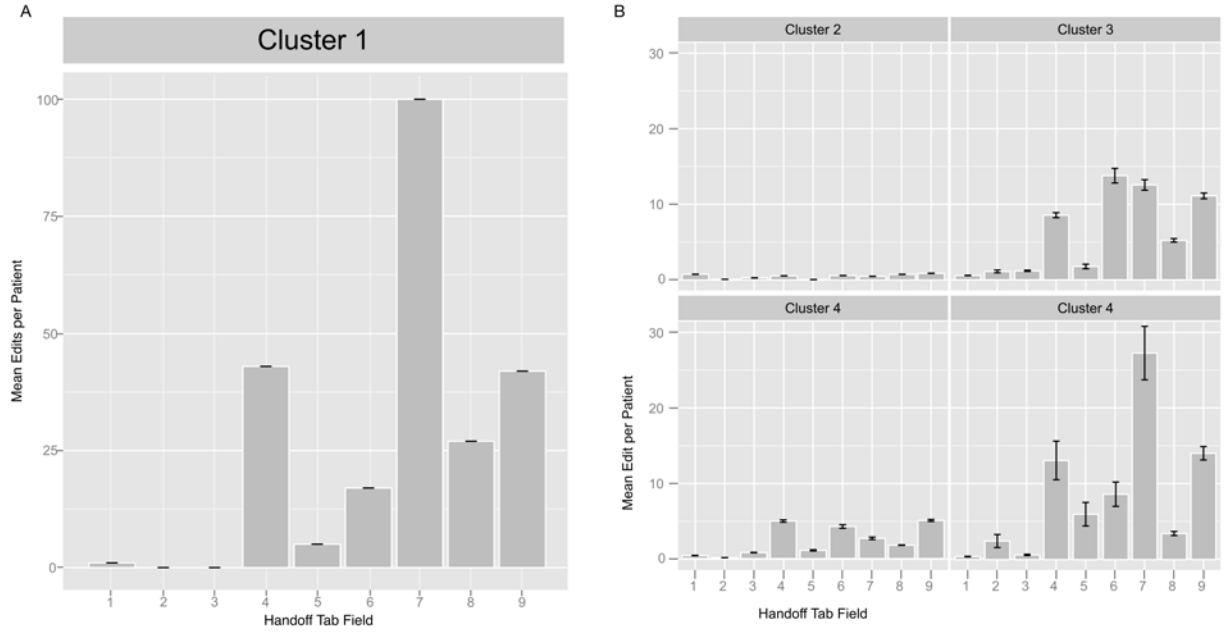


Figure 2A and 2B. Mean and standard error of the mean (SEM) of field frequencies per unit clustering. 1 = Active Issues, 2 = Consult Notes, 3 = Contact Info, 4 = Coverage To Do List, 5 = Discharge Planning, 6 = Hospital Course, 7 = Notes and Comments, 8 = Patient Summary, 9 = Primary To Do List

To test whether clusters could characterize unit properties, the patient count of each cluster was compared. Analysis comparing patient counts in each cluster showed no difference between clusters. Some units that had many patients during the study period update the handoff tab very frequently for each patient, while others with similar patient counts did not. Conversely, some units that did not have very many patients updated the Handoff Tab very frequently. Therefore, considering only the patient count was not predictive of that unit's cluster membership.

Patient Cluster Analysis

To identify whether different patients had differences in the update pattern of their associated handoff document, clustering was used to group patients. A hierarchical cluster analysis showed that a *k*-value of 8 was optimal for a *k*-means cluster analysis. Based on Figure 3, each cluster of patients has a different distribution of update frequencies per Handoff Tab field.

Unlike the unit cluster analysis, each cluster of the patient cluster analysis is differentiated by the distribution of updates between the Handoff Tab fields. For the majority of the clusters (clusters 1 to 6), the magnitude of update frequencies is roughly equal. The distinctive feature of patients in cluster 1 is the low update frequency in the "Consult Notes" field compared to other fields. Cluster 2 patients have relatively more "Consult Notes" and "Discharge Planning" edits compared to the other fields. Cluster 3 patients have a roughly equal number of edits in each field. While cluster 4 patients are similar to those in cluster 3, the edit frequency to "Patient Summary" and "Primary To Do List" appear to be slightly more relative to other fields in cluster 4. Cluster 5 patients have a distinctly high edit frequency in the "Active Issues" field. Cluster 6 patients have relative edit frequency troughs in three fields. Patients in cluster 7 have very distinct peaks in edit frequency in the "Active Issues" and "Contact Information" fields. The final cluster is characterized by having a very low edit frequency in every field, with the exception of the "Patient Summary" field.

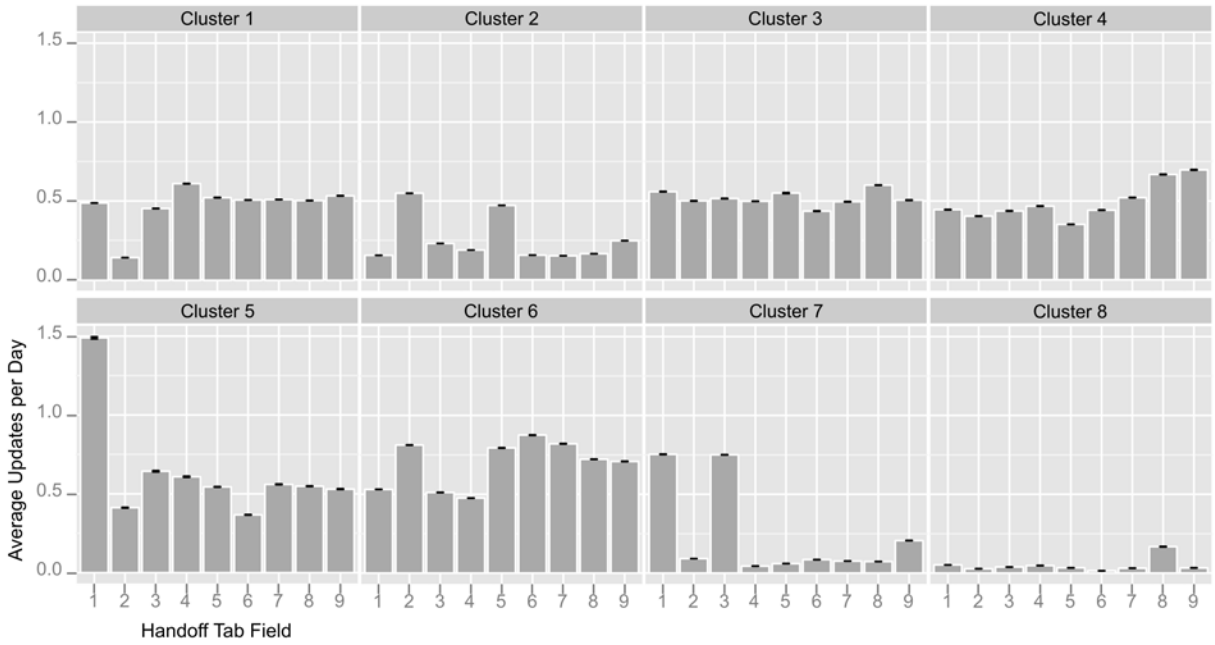


Figure 3. Mean and standard error of the mean (SEM) of field frequencies per patient cluster. 1 = Active Issues, 2 = Consult Notes, 3 = Contact Info, 4 = Coverage To Do List, 5 = Discharge Planning, 6 = Hospital Course, 7 = Notes and Comments, 8 = Patient Summary, 9 = Primary To Do List

Following cluster analysis, a comparison of the handoff days for each cluster was made. Based on Table 2, the mean handoff days for each cluster is similar. This suggests that handoff days is not a good indicator for predicting cluster membership. However, it may be possible for more complex patient characteristics, such as patient severity score, to be indicative of being a member of a certain cluster.

Table 2. Means and deviations of patient handoff days (Days) by patient clusters

Cluster	Patient Count	Mean Stays	Stay Standard Deviation	Stay Standard Error of the Mean
1	821	4.790499	6.513398	0.007933
2	2025	5.568395	6.99678	0.003455
3	589	5.198642	6.517084	0.011065
4	697	4.731707	6.043999	0.008671
5	292	6.130137	8.733759	0.02991
6	927	4.738943	6.235724	0.006727
7	886	4.715576	6.430535	0.007258
8	278	5.309353	6.192868	0.022277

Discussion

There is a considerable body of research suggesting that communication is a key component of teamwork and team coordination[15-18]. Particularly in inpatient settings, where care is provided by patient care teams, communication, both verbal and non-verbal, is a crucial part of a team's ability to coordinate their efforts[15]. Due to common delays in written documentation, clinicians continue to rely on verbal communication to stay up to date[19]. At the same time, there is a growing awareness that verbal communication can be disruptive for work, requires both parties to be available simultaneously, and leaves no record[20]. Computer-mediated communication among members of patient care team can provide an alternative or complementary channel, less disruptive to the flow of clinical work. Yet, the patterns of communication through EHRs are not well understood[21].

Recently, handoff documentation tools were shown to be useful as a source of information for many user groups, beyond physicians[14]. The results of our study suggest that residents, who are primarily responsible for editing information in the Handoff Tab, purposefully use it as an improvised team communication tool, rather than strictly resident handoff tool. The analysis of the temporal patterns of Handoff Tool updates show that residents enter new information throughout the day, not only in preparation for transitions of care. Our interviews with residents further confirmed that often they updated the information to make it available to other members of their teams. Moreover, the residents indicated that Handoff Tab can be used as a collaborative writing platform, with other members of their teams suggesting updates and changes, even if they do not edit the fields directly.

At the same time, the results of our cluster analysis suggested that Handoff Tool is used differently by different departments, and even for different patients. Previous evaluations of handoff tools that included multiple departments were generally taken from a homogenous perspective[12,13]. In contrast to these studies, our analysis found substantial differences in the frequency of updates to Handoff Tool between different departments. Some of these differences are easy to account for. For example, departments with more critically ill patients, such as Intensive Care Units, had a higher frequency of updates than many other hospital units. Yet other differences are not as straightforward. For example, we found that neither the number of patients in the unit nor the average lengths of stay were indicative of the frequency of Handoff Tool updates in that unit. Moreover, interviews with residents suggested that different units developed varying practices in regards to what information is captured in different text boxes. Further research is needed to further explicate local practices in handoff tool use and the different factors that lead to the diversity in its adoption.

Finally, the analysis of differences in frequencies between different patients suggests that the use of Handoff Tab may reflect differences in patients' conditions and care. For example, patients with complex conditions who require ongoing care from different clinical specialties may have more updates in their Consult Notes and Contact Info fields. In contrast, patients with rapidly changing conditions might have more updates in the Active Issues list. Finally, patients who are nearing their discharge from the hospital may have more updates in the Discharge Planning area. While our dataset did not allow us to examine these relationships in greater detail, our interviews with the residents confirmed these intuitions. Recently, nursing documentation patterns were shown to be predictive of patients who experience a cardiac arrest or death[22]. Our study further suggests that there is great utility in understanding the documentation patterns of individual units and patients. For instance deviations in established handoff update frequency patterns for patients in a cardiology ward could prove to be indicative of future complications or deteriorations, such as cardiac arrest.

Limitations

A key limitation of the dataset is that it only provides the fields clinicians edited but not the information that was input. Since Handoff Tab consists of free-text fields, clinicians are free to write information unrelated to that field in the text box. This limitation partially hinders the ability to draw conclusions about what the cluster membership means for either units or patients. In other words, while cluster 1 in the unit clustering analysis suggests that the "Notes and Comments" field is edited frequently, it is impossible to determine what content the edits were focused on. Another limitation regards the correlation between handoff days and length of stay for patients in this dataset. Patients may have been admitted to the hospital before a Handoff Tab document was created or discharged to a unit that did not incorporate Handoff Tab into the unit workflow. Therefore the length of stay may be longer than estimated by handoff days. Furthermore approximately 1.5% of the patient population had length of stays that equaled or exceeded the study period. Additionally, this dataset provided little information characterizing either the unit or the patient. It was impossible to determine the service (cardiology, nephrology, surgical intensive care, etc.) that patient was being treated by. The dataset contained little information about patient severity or any patient outcomes. This limited the abilities to draw conclusions about the patients in each cluster. Lastly, it should be recognized that this dataset represents a small time slice within one academic medical center. With handoff workflows constantly changing, it must be acknowledged these findings may not capture all variations currently in clinical practice.

Conclusion

Handoff Tab was frequently edited throughout the day. This suggests that handoff documentation tools may be a valuable source of up to date information. Furthermore, clusters analysis indicates that different update patterns exist based on units and patients.

Acknowledgements

The research described was supported by the **T15LM007079** grant from the National Library of Medicine.

References

1. Ahmed J, Mehmood S, Rehman S, Ilyas C, Khan LUR. Impact of a structured template and staff training on compliance and quality of clinical handover. Int J Surg. 2012;10(9):571–4.

2. Choma NN, Vasilevskis EE, Sponsler KC, Hathaway J, Kripalani S. Effect of the ACGME 16-Hour Rule on Efficiency and Quality of Care: Duty Hours 2.0. JAMA Intern Med. 2013 Apr 1;:1–2.

3. Horwitz LI, Meredith T, Schuur JD, Shah NR, Kulkarni RG, Jenq GY. Dropping the baton: a qualitative analysis of failures during the transition from emergency department to inpatient care. Annals of Emergency Medicine. 2009 Jun;53(6):701–4.

4. Cohen MD, Hilligoss PB. The published literature on handoffs in hospitals: deficiencies identified in an extensive review. Qual Saf Health Care. 2010;19(6):493–7.

5. Agency for Healthcare Research and Quality (AHRQ). AHRQ Patient Safety Network - Handoffs and Signouts [Internet]. psnet.ahrq.gov. [cited 2014 Mar 13]. Available from: http://psnet.ahrq.gov/primer.aspx?primerID=9

6. Education ACFGM. Graduate Medical Education > Duty Hours [Internet]. www.acgme.org. [cited 2014 Mar 13]. Available from: https://www.acgme.org/acgmeweb/tabid/271/GraduateMedicalEducation/DutyHours.aspx

7. Vawdrey DK, Stein DM, Fred MR, Bostwick SB, Stetson PD. Implementation of a computerized patient handoff application. AMIA Annu Symp Proc. 2013;2013:1395–400.

8. Van Eaton EG, Horvath KD, Lober WB, Rossini AJ, Pellegrini CA. A randomized, controlled trial evaluating the impact of a computerized rounding and sign-out system on continuity of care and resident work hours. Journal of the American College of Surgeons. 2005 Apr;200(4):538–45.

9. Bernstein JA, Imler DL, Sharek P, Longhurst CA. Improved physician work flow after integrating sign-out notes into the electronic medical record. Jt Comm J Qual Patient Saf. 2010 Feb 1;36(2):72–8.

10. Abraham J, Kannampallil TG, Patel VL. Bridging gaps in handoffs: A continuity of care based approach. Journal of Biomedical Informatics. Elsevier Inc; 2012 Apr 1;45(2):240–54.

11. Horwitz LI. Does improving handoffs reduce medical error rates? JAMA. 2013 Dec 4;310(21):2255–6.

12. Riesenberg LA, Leitzsch J, Massucci JL, Jaeger J, Rosenfeld JC, Patow C, et al. Residents“ and attending physicians” handoffs: a systematic review of the literature. Acad Med. 2009 Dec;84(12):1775–87.

13. Riesenberg LA, Leitzsch J, Cunningham JM. Nursing handoffs: a systematic review of the literature. Am J Nurs. 2010 Apr;110(4):24–34–quiz35–6.

14. Schuster KM, Jenq GY, Thung SF, Hersh DC, Nunes J, Silverman DG, et al. Electronic handoff instruments: a truly multidisciplinary tool? Journal of the American Medical Informatics Association. 2014 Feb 19.

15. Sarcevic A, Marsic I, Burd RS. Teamwork Errors in Trauma Resuscitation. ACM Trans Comput-Hum Interact. 19(2):13:1–13:30.

16. Manser T. Teamwork and patient safety in dynamic domains of healthcare: a review of the literature. Acta Anaesthesiol Scand. 53(2):143–51.

17. Leonard M. The human factor: the critical importance of effective teamwork and communication in providing safe care. Qual Saf Health Care. 2004 Oct 1;13(suppl_1):i85–i90.

18. Thomas EJ, Sexton JB, Lasky RE, Helmreich RL, Crandell DS, Tyson J. Teamwork and quality during neonatal care in the delivery room. J Perinatol. 2006;26(3):163–9.

19. Collins SA, Bakken S, Vawdrey DK, Coiera E, Currie L. Clinician preferences for verbal communication compared to EHR documentation in the ICU. Appl Clin Inform. 2011;2(2):190–201.

20. Coiera E. The science of interruption. BMJ Quality & Safety. 2012 May;21(5):357–60.

21. Walsh C, Siegler EL, Cheston E, O'Donnell H, Collins S, Stein D, et al. Provider-to-provider electronic communication in the era of meaningful use: A review of the evidence. Journal of Hospital Medicine. 2013;8(10):589–97.

22. Collins SA, Cato K, Albers D, Scott K, Stetson PD, Bakken S, et al. Relationship Between Nursing Documentation and Patients' Mortality. Am J Crit Care. 2013 Jul 1;22(4):306–13.

An Automated Approach for Ranking Journals to Help in Clinician Decision Support

Siddhartha R. Jonnalagadda, PhD[1,2], Soheil Moosavinasab, BS[2], Chinmoy Nath, PhD[1], Dingcheng Li, PhD[2], Christopher G. Chute, MD, DrPH[2], Hongfang Liu, PhD[2]
[1]Division of Health and Biomedical Informatics, Northwestern University Feinberg School of Medicine, Chicago, IL
[2]Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN

Abstract

Point of care access to knowledge from full text journal articles supports decision-making and decreases medical errors. However, it is an overwhelming task to search through full text journal articles and find quality information needed by clinicians. We developed a method to rate journals for a given clinical topic, Congestive Heart Failure (CHF). Our method enables filtering of journals and ranking of journal articles based on source journal in relation to CHF. We also obtained a journal priority score, which automatically rates any journal based on its importance to CHF. Comparing our ranking with data gathered by surveying 169 cardiologists, who publish on CHF, our best Multiple Linear Regression model showed a correlation of 0.880, based on five-fold cross validation. Our ranking system can be extended to other clinical topics.

Introduction

Since medical errors have become a leading cause of death[1,2] and the healthcare costs and complexity of the diseases have increased, Clinical Decision Support (CDS) shifted its focus from an auxiliary technology to a practical necessity.[3] Due to the sheer amount of information available on clinical topics and the limited time of the user, it is important to deliver context-specific natural language answers. Manual approaches to CDS is cumbersome for content maintainers. Therefore, an automated approach, like text mining, would be desirable.

Our previous research has demonstrated the feasibility of automatically extracting relevant sentences from MEDLINE® abstracts[4] to aid clinicians in finding relevant information quickly. But, much of the relevant information is in the full text article, not the abstract. Taking two randomly selected articles from UpToDate®[5] (a clinical knowledge system) titled "Overview of the Therapy of Heart Failure Due to Systolic Dysfunction" and "Clinically Isolated Syndromes Suggestive of Multiple Sclerosis", we can illustrate the shortcomings of abstract only processing. By counting the number of citations in the abstract and the full-text, we found that only 49.55% (56 out of 113) of the references in the Heart Failure related article and 68.11% (47 out of 69) of the citations in the Multiple Sclerosis related article refer to information in abstract text. Combining the two, only 56.59% (103 out of 182) of references in clinical articles refer to abstracts, suggesting the importance of processing full-text for providing clinical decision support.

Today, there are over 2500 journals that have at least one article indexed in Medline with the Major MeSH® Term – "Congestive Heart Failure". Therefore, it is important to prioritize the journals not only to assign a manageable task for text mining, but also to give confidence to clinicians by providing content from the journals they trust. In addition, journal ranking can be used as a feature in article prioritization task where systems rank[6] or classify[7] articles. There has been research focused on prioritizing individual articles based on quality, specifically for clinical decision support[8], but few focused on prioritizing journals first. Prioritizing journals first aids the clinical knowledge system's decision on the number of full-text articles to process. In addition, directly measuring an article's relevance to a topic is a manageable task using standard text categorization and machine learning techniques[6,7]; however, obtaining the full-texts of the articles for all the 2500 journals and processing each article is not.

Abridged Index Medicus (AIM)[9], also known as core clinical journal titles, is an existing resource that ranks top journals, which has been available since 1970.[10] However, this list is not suitable for identifying the journals from which articles about a specific topic or even a branch of medicine could be extracted. For example, the list misses cardiology journals such as *Circulation: Heart Failure, JACC: Heart Failure* and *European Heart Journal* that are perceived to be important by cardiologists. The McMaster Plus project[11] collaborates with the BMJ Group to provide citations, rated for quality by hand, from over 120 top journals. However, this list of top journals also

misses journals important for a specific topic such as heart failure (example: *Circulation: Heart Failure, JACC: Heart Failure* and *Journal of Heart Failure*).

In this study, we identified the top journals for "Congestive Heart Failure" and the factors that favored their positive perception among cardiologists. We then developed a regression model to prioritize automatically journals for any topic.

Methods

A. Survey Procedure

All the Medline abstracts indexed with the MeSH Major Topic – "Congestive Heart Failure" were downloaded in XML format using the PubMed® interface.[12] The email addresses of the corresponding authors and the countries of their affiliation were extracted using a set of rules and regular expressions.[13] To decrease the variability of the survey participants, we limited them to US organizations.

To reduce outliers and selection bias, we took into consideration that cardiologists are busy and an individual follows only a few journals. Therefore, we first ranked the journals by the number of their articles indexed in Medline by the MeSH Major Topic – "Congestive Heart Failure". The top-100 journals presented in random order were rated by eight cardiologists locally on a 1-5 scale based on their value for clinical decision-making (same criterion as actual survey). Sixty journals that are above a threshold agreed after looking at the distribution of the ratings were considered for the final survey. Based on our experience with the Mayo Clinic cardiologists, we asked the survey participants to rank at least 20 journals among these. The survey was sent to all the participants on March 8, 2013. They were reminded to participate on March 15, 2013 and finally again on March 22, 2013.

Informed consent was obtained for each participant by asking at the beginning:

We want to extract automatically sentences from journals that are relevant to clinician information needs. For the sake of building a prototype, we are asking cardiologists to provide a subjective opinion by rating few journals (presented in random order) in relation to congestive heart failure. By clicking the next button, please confirm that you are consenting to take part in the survey and that you are very familiar with cardiology and congestive heart failure in particular.

If they provided informed consent, we asked them the mandatory questions listed below:

- For how many years have you been practicing medicine?
- How many publications do you have in cardiology?
- How many journals related to congestive heart failure do you follow regularly?
- Choose all your current roles.
- Please rate how valuable information in each journal is to your clinical decision making.

Once they finished the mandatory questions, we asked them two optional questions:

- When choosing a rating for the journal, what are the factors important to you?
- Any comments about the survey?

B. Metric Analysis

With the help of a librarian, we searched existing medical literature to know what factors of a journal such as impact factors, aims & scope and download counts are important in determining whether clinicians follow that journal or use the information at point of care. The metrics mentioned in the articles[14-28] were used as choices in the optional questions.

In addition, the journal-related numeric metrics were compared to the rating average of the corresponding journal using Multiple Linear Regression[29] algorithm:

1. Impact Factor (IF)
2. H-Index
3. SCImago Journal Ranking (SJR)
4. Number of articles (Total Docs)
5. Number of references in the articles (Total Refs)

6. Number of articles for 3 years (3yr Docs)
7. Number of citations to articles for 3 years (3yr Cites)
8. Number of cited-articles for 3 years (3yr Cited)
9. Number of references per article (Ref/Doc)
10. Number of CHF-indexed Medline abstracts (CHF count)
11. Broad Journal Heading – cardiology (0/1) (BJH)
12. Core clinical journal (0/1) (AIM)

The metrics 1-9 are freely available every year from the online resource for journal ranking provided by SCImago Journal & Country Rank[30] from Scopus. The metrics 10-12 are also available freely from the NLM.

The impact factor (IF) of a journal is the number of times that articles published in the past two years are cited by indexed journals during current year, divided by the total number of articles published by that journal in those two years.[31] H-Index of a journal, similar to that of a scientist, is defined to be the maximum possible h such that h number of articles in the journal have at least h citations.[32] SJR is an adaptation of the PageRank metric[33] used by Google©, where the nodes are the individual journals and the edges have weights corresponding to the amount to the citations between the journals representing the nodes. Metrics 4-9 are self-explanatory.

The "Number of CHF-indexed Medline abstracts" is obtained from PubMed interface by limiting the search to the name of the journal AND the MeSH Major Topic – "Congestive Heart Failure". The "Broad Journal Heading" is extracted from the journal descriptions[34] provided by NLM®. For example, "American journal of cardiovascular drugs: drugs, devices and other interventions" has the following broad journal headings: Cardiology, Drug Therapy and Vascular Diseases. When one of the headings is Cardiology, we assign a score of '1'; otherwise, the score is '0'.

Cardiologists' survey rating averages were used as the response variable (Y) and the 12 journal metrics were used as random independent variables (X) for Multiple Linear Regression. We tried all the combinations of the independent variables ($\sum_{k=1}^{12}\binom{12}{k} = 4095$). Variables 11 and 12, which are ordinal, were also compared with the rating averages using the non-parametric (rating averages unlikely to be normal in distribution and the journals in at least one group is anticipated to be less than 25) Wilcoxon's Rank Sum Test.[35]

Results

A. Survey Output

Among all the Medline abstracts with the MeSH Major Topic – "Congestive Heart Failure", there were 32,702 abstracts with an affiliation sentence. Parsing them revealed 14,525 affiliation sentences with at least one email address and these are from multiple countries. Among them, 5,584 sentences correspond to the USA. In these affiliation sentences, we found 3,443 unique email addresses. Among these, 1,088 addresses expired or the recipient is out of office. Of the successful recipients, 169 accepted the eligibility criteria with informed consent and 161 finished the survey. Of these, 19 participants had 0-5 years of experience. For comparing the rating average of journals with the journal metrics, we considered the 142 participants with six or more years of experience.

Table 1 describes the characteristics of the participants. The majority of them have been in practice for over 15 years and published more than 20 cardiology articles.

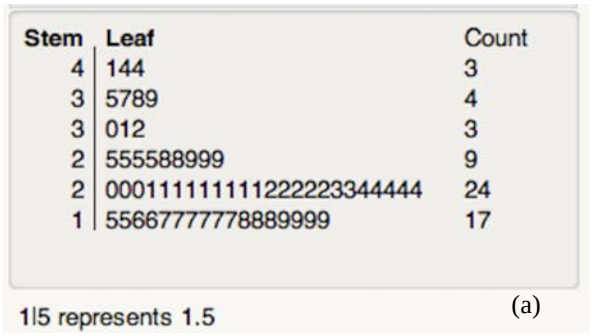
We presented to them 60 randomly ordered journals preselected by the eight Mayo Clinic cardiologists, with an average score of at least 1.5. They were asked to rate these journals by their value in information about CHF in a 1-5 scale (1=least value, 5=most value). The rating averages of the 142 participants with at least 6 years of experience are right-skewed with a median of 2.17 (25% quartile =1.87 and 75% quartile = 2.54). Figure 1(a) shows the Stem and Leaf plot for the distribution of rating averages. It illustrates how a majority of journals got an average rating of 1.5-2.5. Figure 1(b) shows the Stem and Leaf plot for the distribution of response counts. It illustrates that each journal is rated by at least 96 participants.

Table 1. Characteristics of survey participants.

Participant characteristic (N=161)	No. (%) of respondents
Experience (years)	

0-5 6-10 11-15 16-25 >25	19 (11.8%) 20(17.4%) 22(13.7%) 45(28.0%) 47(29.2%)
Cardiology publications 0-5 6-20 21-50 51-100 >100	 24(14.9%) 41(25.5%) 31(19.3%) 23(14.3%) 42(26.1%)
CHF related journals following 0-5 6-10 11-15 16-25 >25	 118(73.3%) 35(21.7%) 7(4.3%) 0(0.0%) 1(0.6%)
All current roles Clinician Researcher Instructor	 130(80.7%) 132(82.0%) 98(60.9%)

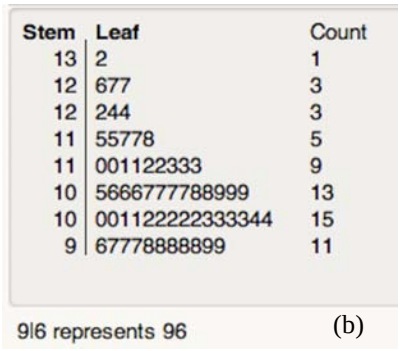
The top-10 journals with their rating averages are in Table 2. The highest rating average is 4.35 (The New England Journal of Medicine) and the lowest rating average is 1.58 (Basic research in cardiology). Table 2 lists the rating averages with the number of responses for 10 journals with highest rating averages. The journal titles in **bold** are those that are NOT indexed in AIM (core-clinical journal list).



Stem	Leaf	Count
4	144	3
3	5789	4
3	012	3
2	555588999	9
2	000111111111222223344444	24
1	55667777778889999	17

1|5 represents 1.5

(a)



Stem	Leaf	Count
13	2	1
12	677	3
12	244	3
11	55778	5
11	001122333	9
10	5666777788999	13
10	001122222333344	15
9	67778888899	11

9|6 represents 96

(b)

Figure 1. Stem and Leaf plots for the ratings from 142 experienced participants: (a) individual rating averages for each of the 60 journals (b) individual responses for each of the 60 journals.

Table 2. Top journal rating averages for 142 participants (at least six years of experience).

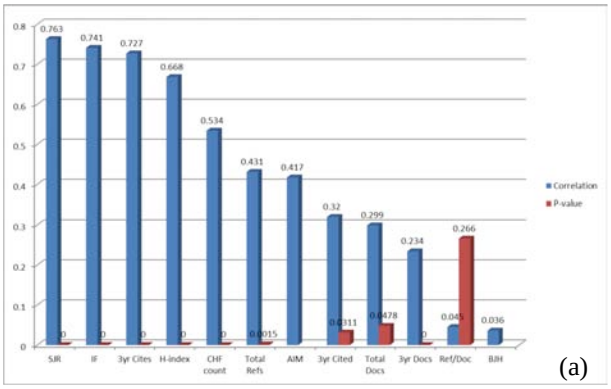
Journal names	Rating Average	Response Count
The New England journal of medicine	4.35	132
Circulation	4.35	127
Journal of the American College of Cardiology (JACC)	4.13	127
JAMA : the journal of the American Medical Association	3.86	124

Circulation. Heart failure	3.79	124
Lancet	3.74	126
JACC. Heart Failure	3.52	104
European heart journal	3.21	112
Annals of internal medicine	3.14	118
Journal of cardiac failure	3.04	117

Figure 2(a) shows the correlation coefficients of the first 10 variables, which are continuous along with the respective p-values. The rows with p-values in bold indicate that the correlation of the variable with ranking average is statistically significant (alpha=0.05). It should be noted from Table 3 and Figure 2(a) that although the topic-specific metrics (CHF count, BJH and AIM) are itself less predictive than other features, they provide complementary information that improves the overall predictive power significantly in combination with top-metrics such as SJR.

Table 3. Top-10 accurate multiple linear regression combinations (numbers rounded to 3 decimal places).

IF	H-Index	SJR	Total Docs	Total Refs	3yr Docs	3yr Cites	3yr Cited	Ref/ Doc	CHF count	BJH	AIM	Correlation
		0.826	-0.003		0.002	0	-0.001	-0.018	0.002	0.628	-0.327	**0.880**
		1.065	-0.002	-0.000	0.001			0.019	0.003		-0.913	**0.879**
		0.699	-0.004		0.002		-0.002		0.003			**0.879**
		1.005	-0.003	-0.000	0.002				0.002	0.406		**0.878**
		0.661	-0.003		0.002		-0.002	-0.015	0.003			**0.877**
0.293			-0.002		0.001		-0.001	-0.009	0.003			**0.876**
		0.677	-0.002		0.001	0.000	-0.002	-0.007	0.003			**0.875**
0.332			-0.002		0.001		-0.001		0.003	0.328	-0.108	**0.874**
		1.055		-0.000	0.000	0.000			0.003	-0.210		**0.874**
		0.729	-0.004		0.003		-0.002		0.002	0.445	-0.529	**0.874**



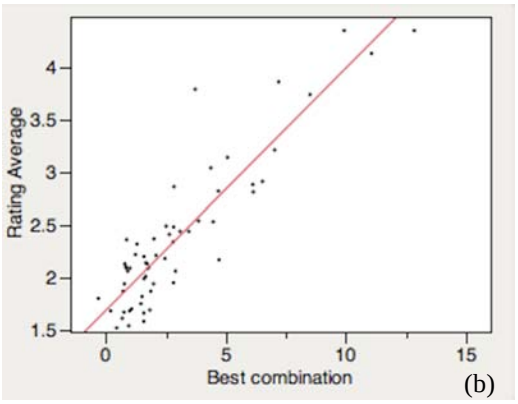


Figure 2. (a) Correlation of the continuous variables with rating averages (b) Linear bivariate fit of rating average by weighted combination of best model

B. Metric Analysis

The 60 rated journals by 142 cardiologists and twelve variables were used to create a Multiple Linear Regression model for rating average. Table 3 lists the top-10 combinations (among the 4095 combinations tested) of metrics and their coefficients. The best model has a correlation of 0.880. The correlation was calculated from the 5-fold cross validation and the final coefficients were obtained from the sixth run combining all the 60 journals. The only variable that is present in all the top-10 models is CHF count (Number of CHF-indexed Medline abstracts). In fact, the best 1,388 models employ the CHF count variable. Without employing this variable, the best correlation obtained is .806 (8.4% lesser). The other variables derived from NLM's sources – BJH (Broad Journal Heading) and AIM (core clinical journals) respectively take into account whether the journal is related to cardiology or is a top medical journal. Using none of these three metrics, the best correlation is only 0.792 (10.0% lesser).

We obtained the below formula to automatically rate the journals.

Formula 1: Journal rating metrics and coefficients

Journal priority score =

(0.82640 * SJR) -
(0.00377 * Total Docs) +
(0.00258 * 3yr Docs) -
(0.00190 * 3yr Cited) -
(0.01846 * Ref/Doc) +
(0.00295 * CHF count) +
(0.62864 * BJH) -
(0.32753 * AIM))

The two groups – core clinical journals (AIM) and the rest have statistically significant difference in means of rating averages ($p=0.0003<0.05$, Rank sums test). The respective means are, however, close – 2.9 vs. 2.1. The groups – cardiology journals and non-cardiology journals (based on BJH) do not have statistically significant difference in means of rating averages ($p=0.1888>0.05$, Rank sums test). Surprisingly, the non-cardiology journals have a slightly higher rating average mean (2.5) than cardiology journals (2.3). This is because the non-cardiology journals with higher impact factor, h-index or SJR such as NEJM and JAMA have a high acceptance even among cardiologists.

In addition, the participants (159 participants answered this question) rated the factors mentioned in literature and suggested by local cardiologists as shown below:

1. Clinical relevance: 108 (76.6%)
2. Impact factor: 93 (66.0%)
3. Type of articles – Human Vs. Animal Research: 66 (46.8%)
4. Electronic Access: 61 (43.3%)
5. Study design of the articles: 61 (43.3%)
6. Quantity of topic-related articles: 53 (37.6%)
7. Aims and scope: 50 (35.5%)
8. Proportion of topic-related articles: 49 (34.8%)
9. Open-access: 28 (19.9%)
10. Readership composition: 26 (18.4%)
11. Editorial board composition: 26 (18.4%)

Discussion

A. Selection Bias

Such a survey conducted at a single institution might be influenced by selection bias (sample being not representative of population). In our case, we invited participants from across the United States. This study is part of a 4-year project funded by the NLM (USA) intending to help cardiologists in the USA in areas such as CHF. The cardiologists invited are affiliated with organizations from all the 50 states of USA and Puerto Rico and over 500 different cities of USA.

Table 4. Distribution of citations to journal names

Heart Failure related article				
Journal names	*Rank based on Formula 1*	*Rank based on survey*	*AIM journal?*	*# of references in article*
The New England journal of medicine	1	1	Yes	23
Circulation	3	1	Yes	31
Journal of the American College of Cardiology	2	3	Yes	20
Lancet	4	6	Yes	8
JAMA : the journal of the American	5	4	Yes	3

Medical Association				
European heart journal	6	7	No	9
The American journal of cardiology	7	10	Yes	2
American heart journal	9	11	Yes	3
Annals of internal medicine	10	8	Yes	2
Journal of cardiac failure	14	9	No	6
BMJ (Clinical research ed.)	17	19	Yes	1
Cochrane database of systematic reviews (Online)	44		No	1
American journal of hypertension	45		No	1
The Canadian journal of cardiology	48	52	No	2
Blood pressure	60		No	1
Multiple Sclerosis related article				
The New England journal of medicine	3		Yes	4
Neurology	4		Yes	22
Lancet	5		Yes	11
Multiple sclerosis (Houndmills, Basingstoke, England)	7		No	3
Lancet neurology	8		No	9
Annals of neurology	11		No	6
Archives of neurology	14		Yes	6
Brain : a journal of neurology	15		Yes	4
Cochrane database of systematic reviews (Online)	38		No	2
Archives of ophthalmology	57		Yes	2

B. Number of Participants vs. Number of Journals (Nonparticipation Bias)

More participants will minimize a journal getting a significantly higher or lower rank than its true mean (central limit theorem). A higher number of journals allows for more statistical power in the calculations of association between the metrics and rating averages. However, we observed a tradeoff between the number of participants and number of journals they rate while testing the survey at Mayo Clinic. Several cardiologists recommended reducing the journals significantly. Therefore, we chose 60 journals instead of 100 and imposed a constraint in the survey so that the participants rated at least 20 (33%) journals. Despite these, we observed statistically significant differences for all metrics except Ref/Doc (Number of references per article) and BJH (Broad Journal Heading). On average, each participant rated 45.0 journals (out of 60 maximum) and each journal received 120.75 ratings (out of 161 maximum).

To further verify this, a replication survey was conducted at Mayo clinic in which 18 cardiology fellows participated. This step is also useful for comparing results of the previous survey where the participants are established clinicians or researchers and with this survey with a different population. The demographic characteristics of the participants are presented in Table 5. In Table 6, rating averages with the number of responses for the 10 journals with highest rating averages are presented. The highest rating average is 4.71 (The New England Journal of Medicine) and the lowest rating average is 1.50. The ranking of the journals show a large overlap.

Table 5. Characteristics of survey participants of 18 respondents

Participant characteristic (N=18)	No. (%) of respondents
Experience (years)	
0-5	7 (38.9%)
6-10	8(44.4%)
>11	3(16.7%)
Cardiology publications	
0-5	13(72.2%)
6-20	5(27.8%)
>21	0(0.0%)
CHF related journals following	
0-5	16(88.9%)
6-10	2(11.1%)
>11	0(0.0%)
All current roles	
Clinician (7 serve also as researcher, 1 as researcher & instructor)	14 (77.8%)
Researcher	4(22.8%)
Instructor	0(0.0%)

Table 6. Top journal rating averages for 18 participants (rank in the original survey in brackets)

Journal names	Rating Average	Response Count
The New England journal of medicine (1)	4.71	14
Circulation (2)	4.57	14
Journal of the American College of Cardiology (JACC) (3)	4.46	13
JAMA : the journal of the American Medical Association (4)	3.93	14
European heart journal (8)	3.75	16
Lancet (6)	3.75	16
Circulation. Heart failure (5)	3.58	12
Heart (British Cardiac Society) (21)	3.50	14
The American journal of cardiology (11)	3.38	13
JACC. Cardiovascular imaging (19)	3.36	11

C. Generalizability

The best Multiple Linear Regression model has a correlation as high as 0.880. Figure 2(b) shows the regression fit with the weighted combination in the X-axis. We will also be able to update the list with new journals related to CHF or automatically update the scores for existing journals with time. The focus of our project was to help cardiologists with the topics – CHF and atrial fibrillation (afib). The survey we conducted locally solicited cardiologists to rate the journals for both CHF and afib. The correlation between the ratings is 89.16%. Therefore, it is likely that the list might be repurposed for other cardiology topics with expert editing.

To check the generalizability of Formula 1 for non-cardiology topics, we used the same two articles from UpToDate to assess how the rank of a journal obtained using the formula is related to the number of references from the journal. Table 4 shows all the journal names in the respective articles sorted by the journal priority score calculated using Formula 1.

It should be noted that the ranks of the journals obtained using the formula are similar to the rankings obtained from survey ratings. The top 15 journals as per the journal priority score calculated by Formula 1 for Heart Failure domain cover 94.69% (107 out of 113) of references. Using the same formula for the MS related article (with mesh counts calculated using “multiple sclerosis”), almost the same coverage is obtained (94.20% (65 out of 69)). In comparison, if the 119 AIM journals are used, 82.30% (93 out of 113) and 71.01% (49 out of 69) of references are covered in Heart Failure and Multiple Sclerosis topics respectively – a 14% less coverage on average. For building a topic-specific clinical decision support system using knowledge in literature, we might achieve a higher coverage even using lesser number of journals (about 15) than using all the AIM journals. When building a generic decision support system, one might expand the AIM journals for the key topics using the journal priority score formula.

Although this simple analysis suggests the formula might be generalizable on non-CHF related articles (94.20% coverage on MS topic), increasing the size of the corpus by conducting multiple surveys and training the system with more than one topic will increase the accuracy (measured by correlation) of the formula.

D. Choice of Algorithm

This was determined by the fact that this is not a classification problem (where algorithms such as Naïve Bayes are used) or a clustering problem (where algorithms such as K-means are used), but a regression problem. Since the individual metrics are linearly correlated, Linear Regression is more appropriate than other regression algorithms such as Logistic Regression. Here we chose Multiple Linear Regression algorithm since it accounts for all possible combinations of the metrics.

Conclusion

With the help of cardiologists across the United States, we rated and ranked journals related to CHF. A replication survey with cardiology fellows was done to verify that the ratings are generalizable to other demographics. While these ratings will help us move forward in CHF, we have also obtained a formula to automate the process for other topics such as Multiple Sclerosis. The Multiple Linear Regression model achieves a correlation of 0.880 (five-fold cross-validation). We demonstrated that general journal metrics such as impact factor, h-index and number of articles per year provide better results when used in combination with topic-specific metrics such as number of abstracts indexed with a MeSH term. Our participants ranked these and other factors in this order: clinical relevance (77%), impact factor (66%), human vs. animal research (47%), etc. All these factors need to be taken in consideration for further improving and generalizing this work.

Acknowledgement

We thank Mayo Clinic's cardiologists for helpful comments and discussion. We are also thankful to UpToDate© for allowing us to use their data for research purposes. This paper would not have been possible without the anonymous participation of cardiologists across USA. This publication was made possible by the NLM K99/R00 grant LM011389.

References

1. Kohn LT, Corrigan J, Donaldson MS, America CoQoHCi. To err is human: building a safer health system. Washington, D.C.: National Academic Press; 1999.
2. Leape LL, Berwick DM. Five years after to err is human. Journal of American Medical Association 2005;293:2384.
3. Greenes RA. Clinical decision support: the road ahead. Burlington, MA: Academic Press; 2007.
4. Jonnalagadda SR, Del Fiol G, Medlin R, et al. Automatically extracting sentences from Medline citations to support clinicians' information needs. Journal of the American Medical Informatics Association : JAMIA 2012.
5. UpToDate Inc. 2011. at http://www.uptodate.com.)
6. Jonnalagadda S, Petitti D. A new iterative method to reduce workload in systematic review process. Int J Comput Biol Drug Des 2013;6:5-17.
7. Aphinyanaphongs Y, Tsamardinos I, Statnikov A, Hardin D, Aliferis CF. Text categorization models for high-quality article retrieval in internal medicine. Journal of the American Medical Informatics Association: JAMIA 2005;12:207-16.
8. Wang D, Kaufman DR, Mendonca EA, Seol YH, Johnson SB, Cimino JJ. The cognitive demands of an innovative query user interface. Proc AMIA Symp 2002:850-4.
9. Abridged Index Medicus (AIM). 2012. at http://www.nlm.nih.gov/bsd/aim.html.)
10. Slagle AD. Abridged Index Medicus. Archives of Pediatrics & Adolescent Medicine 1970;119:193.
11. Hemens BJ, Haynes RB. McMaster Premium LiteratUre Service (PLUS) performed well for identifying new studies for updated Cochrane reviews. J Clin Epidemiol 2012;65:62-72 e1.
12. PubMed interface. 2013. at http://www.ncbi.nlm.nih.gov/pubmed.)
13. Jonnalagadda SR, Topham P. NEMO: Extraction and normalization of organization names from PubMed affiliations. Journal of Biomedical Discovery and Collaboration 2010;5:50-75.
14. Jones TH, Hanney S, Buxton MJ. The role of the national general medical journal: surveys of which journals UK clinicians read to inform their clinical practice. Med Clin (Barc) 2008;131 Suppl 5:30-5.
15. Pendlebury DA. The use and misuse of journal metrics and other citation indicators. Arch Immunol Ther Ex 2009;57:1-11.
16. Fassoulaki A, Karabinis G, Paraskeva A. How readers perceive the quality of six anesthesia journals, their editors and reviewers: a European survey. Acta Anaesthesiologica Belgica 2010;61:195-201.
17. Brown T. Journal quality metrics: options to consider other than impact factors. American Journal of Occupational Therapy 2011;65:346-50.
18. Gibson JC. Impact factor in general practice. Quality in Primary Care 2011;19:3-4.
19. Eysenbach G. Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. Journal of Medical Internet Research 2011;13:e123.

20. Lokker C, Haynes RB, Chu R, McKibbon KA, Wilczynski NL, Walter SD. How well are journal and clinical article characteristics associated with the journal impact factor? a retrospective cohort study. Journal of the Medical Library Association 2012;100:28-33.
21. Frederickson RM, Brenner MK. Assessing Journal Influence: Impacted Wisdom. Mol Ther 2012;20:1481-2.
22. AbdullGaffar B. Impact factor in cytopathology journals: what does it reflect and how much does it matter? Cytopathology 2012;23:320-4.
23. Barraclough K. Why doctors don't read research papers. Brit Med J 2004;329:1411-.
24. O'Donnell M. Why doctors don't read research papers - Scientific papers are not written to disseminate information. Brit Med J 2005;330:256-.
25. Ide CW. Why doctors don't read research papers - Editors' behaviour might have something to do with it. Brit Med J 2005;330:256-.
26. Barraclough K. Why doctors don't read research papers - Reply. Brit Med J 2005;330:256-.
27. Khaliq MF, Noorani MM, Siddiqui UA, Anwar M. Physicians reading and writing practices: a cross-sectional study from Civil Hospital, Karachi, Pakistan. BMC Medical Informatics and Decision Making 2012;12.
28. Hudson C, Cecere E, Yalamanchili R, et al. Utilization and attitudes on technological advances in medical publications. Curr Med Res Opin 2012;28:S16-S.
29. Aiken LS, West SG, Pitts SC. Multiple linear regression. Handbook of psychology 2003.
30. SCImago Journal & Country Rank. 2007. at http://www.scimagojr.com/.)
31. Garfield E. The history and meaning of the journal impact factor. JAMA: the journal of the American Medical Association 2006;295:90-3.
32. Braun T, Glänzel W, Schubert A. A Hirsch-type index for journals. Scientometrics 2006;69:169-73.
33. Page L, Brin S, Motwani R, Winograd T. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report: Stanford InfoLab; 1999. Report No.: SIDL-WP-1999-0120.
34. Journal description metadata. 2013. at ftp://ftp.nlm.nih.gov/online/journals/)
35. Wolfe DA, Hollander M. Nonparametric statistical methods. Nonparametric statistical methods 1973.

De-identification of Address, Date, and Alphanumeric Identifiers in Narrative Clinical Reports

Mehmet Kayaalp, MD, PhD, Allen C. Browne, MS,
Zeyno A. Dodd, PhD, Pamela Sagan, RN, Clement J. McDonald, MD
Lister Hill National Center for Biomedical Communications,
U.S. National Library of Medicine, National Institutes of Health, Bethesda, MD

Abstract

Introduction: *The Privacy Rule of Health Insurance Portability and Accountability Act requires that clinical documents be stripped of personally identifying information before they can be released to researchers and others. We have been developing a software application, NLM Scrubber, to de-identify narrative clinical reports.* ***Methods****: We compared NLM Scrubber with MIT's and MITRE's de-identification systems on 3,093 clinical reports about 1,636 patients. The performance of each system was analyzed on address, date, and alphanumeric identifier recognition separately. Their overall performance on de-identification and on conservation of the remaining clinical text was analyzed as well.* ***Results****: NLM Scrubber's sensitivity on de-identifying these identifiers was 99%. It's specificity on conserving the text with no personal identifiers was 99% as well.* ***Conclusion****: The current version of the system recognizes and redacts patient names, alphanumeric identifiers, addresses and dates. We plan to make the system available prior to the AMIA Annual Symposium in 2014.*

1. Introduction

Electronic health records are treasure troves for clinical scientists because with the availability of high volumes of electronic reports, clinicians are no longer limited to a cohort of their patients and can easily test their hypotheses on much larger samples. Access to those records, however, is not easy and involves overcoming a number of institutional barriers. These barriers have been raised purposefully to ensure that only the right person could access the private information of the patient. While these barriers had been the primary means to protect patient privacy, the requirements were so difficult to attain that they also became a barrier to scientific progress. Having seen both sides of the issue, the U.S. Congress enacted Health Insurance Portability and Accountability Act (HIPAA) in 1991, which tasked the U.S. Department of Health and Human Services (HHS) with regulating access to health records while protecting the health information of individuals.

The Privacy Rule of HIPAA requires that clinical documents be stripped of personally identifying information before they can be released to researchers and others. There have been several attempts to de-identify clinical text data automatically via software with an upward trend of performance, yet the clinical research community still considers human verification and validation necessary prior to the release of any automatically de-identified clinical data.

We have been developing a software system to automatically de-identify clinical records, which we have named NLM Scrubber. The current version of the system recognizes and redacts patient names, alphanumeric identifiers, addresses, and dates. The primary goal in clinical text de-identification is to raise the sensitivity performance of the system to an acceptable level so that de-identified data can be used with no need of human verification. We designed NLM Scrubber with that goal in mind and this study is a new step forward to reach that goal.

2. Background

As defined by HHS, Protected Health Information (PHI) comprises a subset of the health information of an individual who is the subject of the health record *and* personally identifiable information[*] (PII), including demographic information, collected from the same individual to be used by the health care provider, health plan, employer or health clearinghouse. PII is any information that distinguishes or traces an individual's identity such as

[*] The text of CFR 45 § 164.514 uses the term *individually* identifiable information instead of *personally* identifiable information. One possible reason is that the meaning of the legal term *person* also includes entities other than natural person (human) such as trust, estate, partnership, corporation, and professional association among others. Since personally identifiable information and its acronym PII are more widely known terms, we used them here instead.

name, social security number, date of birth or biometric records and any other information such as medical, financial and employment information that is linkable to an individual.[2 3]

Table 1. Per HIPAA Privacy Rule, the following identifiers must be deleted from PHI to fully de-identify health information. (*) As of 2010, there were 18 sets of zip codes with distinct initial three digits whose corresponding population sizes were less than or equal to 20,000.[1]

1. Names
2. All geographic subdivisions smaller than a state, except the first two digits of the zip code of the postal address. The third digit of the zip code can also be left intact, only if the size of the population in the area of the censored two digits is greater than 20,000 according to the most recent census data.(*)
3. All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older.
4. Telephone numbers.
5. Fax numbers.
6. Electronic mail addresses.
7. Social security numbers.
8. Medical record numbers.
9. Health plan beneficiary numbers.
10. Account numbers.
11. Certificate/license numbers.
12. Vehicle identifiers and serial numbers, including license plate numbers.
13. Device identifiers and serial numbers.
14. Web universal resource locators (URLs).
15. Internet Protocol (IP) address numbers.
16. Biometric identifiers, including fingerprints and voiceprints.
17. Full-face photographic images and any comparable images.
18. Any other unique identifying number, characteristic, or code, except the ones that may be generated by the covered entity for re-identification.

HHS developed the Privacy Rule, defining certain identifiers as part of PHI, which should be de-identified before health records are accessed for research purposes (see Table 1). Note that the health information dissociated from those identifiers of the individual is not considered PHI. According to the Privacy Rule the identifiers in Table 1 that belong to the individual or relatives, employers or household members of the individual, should not be present in any de-identified health records.[4]

2.1 Current Clinical Text De-identification Systems

De-identification of structured data is a fairly straightforward process, where the content of fields defined as containing PHI (e.g., name and birth date fields) be removed and made inaccessible to researchers. De-identification of unstructured data or free text, on the other hand, is a challenge. Because of the idiosyncrasies of any natural language, including English, the utterances of information are not always predictable and we have to devise intelligent tools to recognize those words and phrases containing PHI.

A thorough review of 18 clinical text de-identification systems has been published recently.[5] Since then only four additional systems have appeared in major journals.[6-9] These 22 systems can be categorized in two groups based on their target documents: general purpose vs. niche (specialized) de-identifiers. They can also be classified in terms of their underlying methodologies, which use either symbolic or machine learning approaches. Symbolic approaches mainly rely on rules, regular expressions, and lookup tables (also referred to as dictionaries or gazetteers). The availability of a de-identification system is another important characteristic; some are freely available, some are commercial products, and others have not been made available.

Currently, there are only five freely available systems, three of which were specialized to de-identify surgical pathology reports only.[10-12] The other two systems are general purpose de-identification systems developed by researchers at MIT and MITRE. MIT's system took a symbolic approach; whereas, MITRE's is a machine learning system using conditional random fields.

The name of MIT's system was not mentioned in their publication[13] but the filename of the code was *deid.pl*. Since there is another (commercial) system with the same name, De-ID, to prevent any confusion, we here call MIT's system "MITdeid". MITdeid provides various features that are closely tuned to clinical setting, such as accepting a list of provider names of the institute and the full name of the patient per report.

MITRE's system, MIST, was developed to demonstrate how an existing conditional random field program designed for generic use could be repurposed quickly as a successful clinical text de-identification system.[14] MIST has proven itself as one of the most successful systems in the i2b2 competition in 2006.[15] As a machine learning system, MIST requires a training dataset. The current version of the system does not store the constructed model and has to be re-trained before each testing session.

As part of Clinical Text De-identification Project at NLM, we studied personal name recognition in great depth. Since we previously reported the results of the name scrubbing performance of our system elsewhere, we do not repeat them here.[16]

2.2 Contribution of NLM Scrubber

There have been several attempts to de-identify clinical text data automatically via software, but all those systems that are freely available had failed to scrub some identifiers; showing the need for improvement. As part of HHS, NLM started the clinical text de-identification project to respond to this need and promote scientific progress by enabling the research community to access large amount of health information that does not contain personal identifiers. NLM Scrubber will be freely available at the time of this publication. The goal of NLM Scrubber is to facilitate the production of de-identified clinical text data, and minimize (if not eliminate) the burden of manual de-identification for the clinical research community.

3. Methods

In this section, we present (1) how we select and process the clinical text data from a large corpus of clinical reports in order to reliably study and develop robust de-identification methods; (2) the methods and components of our de-identification tool; and (3) our evaluation methods.

3.1 Data Selection

A hospital information system may retain many copies of a given report (e.g., the initial report, the report with addenda or corrections) and HL7 messages logs that we used do the same. A sample with duplicate reports may inflate the magnitude of events observed in the study. Duplicate reports in a de-identification test set would distort the analysis of tests and the success of a de-identifying program.

To ensure the reliability of the study results, we devised a random sampling method to exclude any redundant reports. For each randomly selected patient, we collected all reports generated during a particular visit of the patient, grouped them by report types, sorted each group by report filing time, and took only the most recent, presumably the most developed report in that group.

Our sampling method relies on the assumption that each visit is associated with a unique visit number and reports of the same type in two different visits are sufficiently dissimilar. Note that this assumption may not always hold. After performing the sampling, we sorted reports of each patient by word counts. The manual comparison of the reports that were similar in size helped us discover partially duplicate records from distinct visits of the same patients. We eliminated the earlier reports from these sets. This approach may inadvertently eliminate some non-duplicate reports, but in the final analysis, it yields an unbiased, large spectrum of reports per visit with distinct report types.

In this study, we used 3093 distinct clinical reports about 1636 distinct patients of the Clinical Center at NIH. The maximum number of reports per patient was 20.

We developed NLM Scrubber using a training set of 1140 clinical reports from the same institution. Unlike the study test data, retrieval of the training data was done in several iterations over a long period of time in an ad hoc and not truly randomized manner. The patient cohorts in the training and test data were mutually exclusive.

3.2 PII Recognition Methods

We collapse 18 types of identifiers defined by HIPAA into four distinct categories: 1) names, 2) addresses, 3) dates (incl. ages), and 4) alphanumeric identifiers (e.g., medical record numbers, report identifiers, telephone numbers). We reported our results with the first category, names, previously [16]. Here we report our results with the other three categories.

3.3 Alphanumeric identifier recognition

We define an alphanumeric string as a string of characters containing at least one or more digits. It may or may not contain other characters. Alphanumeric Identifier Recognizer (AIR) has a two-prong approach: It detects patterns that correspond to alphanumeric strings such as phone and social security numbers that need to be labeled as alphanumeric-Id, but it also detects patterns of known clinical entities such as lab values that need to be preserved. AIR also attempts to distinguish alphanumeric strings from date-like patterns so that dates would not be mislabeled.

If a token of alphanumeric string contains only a single digit, AIR ignores it completely; otherwise, it analyzes the content of token t and its context. If t is preceded by a token containing certain strings such as number, protocol, or # sign, it labels t as an alphanumeric identifier. If an alphanumeric token containing a sequence of two or more upper case letters is followed by certain tokens such as "protocol", it is labeled as alphanumeric identifier. A 9–10 digit number pattern with or without delimiters in between is also defined as alphanumeric identifiers (i.e., phone or social security numbers).

AIR also checks numerous conditions (e.g., a number followed by a unit of measure) that may indicate that token t be a valid piece of clinical data to be conserved and marks most other alphanumeric strings as alphanumeric identifiers.

3.4 Date and age recognition

Algorithms for identifying dates and ages are based on a set of regular expressions to detect the corresponding patterns. Some date patterns are listed in Table 2. For example, string 07-08-2012 would be identified using the pattern DD*MM*YYYY, where "*" is a delimiter and D, M, Y are digits such that YYYY should be greater than 1900 and less than the current year, $1 \leq \text{DD} \leq 31$ and $1 \leq \text{MM} \leq 12$.

Table 2. Date Patterns: D, M, Y, h, and m are date, month, year, hour and minute digits; * is a delimiter; MONTH, HOLIDAY are literal values of month and holiday incl. their abbreviations; X? denotes that X is optional; | indicates distinct disjunctive patterns

Pattern	Example	Pattern	Example
YYYY*MM*DD	2012-08-07	**DD*MM?**	07-08, 07-8
DD*MM*YYYY	07-08-2012	**YYYYMMDD**	20120708
MM*DD*YYYY	08-07-2012	**YYYYMMDDhhmm**	201207081215
MM*DD*YY	08-07-12	**YYYY**	2012
M*DD*YY	8-07-12	**DD?*?MONTH**	7-August, 7August, 7 Aug
YYYY*YYYY	2011-2012	**MONTH*YY(YY)?**	August.2012, August'12
DD*MM*YY	07-08-12	**(early|mid|late)*YYYY**	Mid-2012
M*D*YY	8-7-12	**YYYY*?MONTH**	2012/August, 2012Aug
M*DD*YY	8-07-12	**'YY*?MONTH**	'12-August, '12Aug
MM*YYYY	08-2012	**DD?MONTH*YY**	7August'12
DD*MM*DD*MM	07-08/08-08	**MONTH*DD?**	Aug7, August 7
MM?*DD	08-07, 8-07	**MONTH**	Aug, August
DD?*MM	07-08, 7-08	**HOLIDAY**	Christmas, Easter
MM*DD?	08-07, 08-8		

Unlike date patterns, age patterns are more involved. For example, age patterns may be required to catch phrases like "on his **ninety-third** birthday" or "in his late **90s**".We classified alphanumeric age expressions and labeled them with specific names (see Table 3). The corresponding patterns are recognized through regular expressions.

Whenever a date (age) regular expression is matched with the tokens in the text, those tokens are labeled as date (age).

Table 3. Alphanumeric Age Expression Classes

Expression Classes	Examples
AGE-WITH-SUCCEEDING-MARKER	he was [93 years-old]
AGE-WITH-PRECEDING-MARKER	at the [age of 93],
AGE-WITH-APPENDED-UOM	his father, [93yo], has
AGE-FRACTION-EXPR	he is [5-years and 3-months] old
AGE-FROM-PHRASE-CONTEXT	she [was nearly 93].
AGE-BIRTHDAY-CONTEXT	on his [ninety-third birthday]
AGE-DECADE-CONTEXT	in his late [90s]
AGE-SIMPLE-CATCH-ALL	(as 93)
AGE-COMPOUND-CATCH-ALL	(93 and 90)

3.5 Address identifier recognition

Addresses are recognized mostly via the "shapes" component of dTagger, a specialized part-of-speech tagger extended with limited pattern tagging abilities for entities, such as addresses.[17] The dTagger searches address terms in various lexicons, which contain city and states names as well as street types and their abbreviations (e.g., Avenue, Alley, Blvd, and Circle). In its current format, the recognizer is difficult to maintain and will be revised before the release of the software package; therefore, we do not provide any further specifications of the soon-outgoing recognizer in this report.

3.6 Redaction

The redactor removes the PHI content in a post-processing step, where it replaces the removed text with a corresponding standard PHI label. For example, John would be replaced with [NAME]. If two distinct recognizers (e.g., both date and alphanumeric-Id recognizers) tag the same token as PHI, the redactor labels the content as [PHI] instead of choosing one tag over another.

3.7 Evaluation Methods

We evaluated the NLM Scrubber on a test set of 3,093 dictated narrative reports generated at the NIH Clinical Center. The set was annotated by two experts, a linguist and a registered nurse, producing the gold standard for the test data. Following NLM Scrubber's run on the study data, we compared the resulting tags against the gold standard and evaluated them in terms of sensitivity, specificity, precision, and F_2 measures. We also evaluated the privacy risks due to the revealed PHI tokens.

We tested the scrubbing performance of two of the most prominent and freely available de-identification systems, MIST and MITdeid against the same data and used the same evaluation approach for all of them.

Since MIST is a machine learning system, it requires training before testing. We used our held-out set of 1,140 annotated reports as training data for MIST. After testing MIST extensively on the training dataset using various parameterizations and based on consultations with its developers, we decided to run it with a bias of –4, which greatly favors sensitivity over precision but not to the extent that the results become unreliable. We appreciate the generous assistance we received from many members of the MIST developer team at different phases of our study.

3.8 Evaluation of differently tokenized results

Most de-identification systems come with their own tokenizers producing different sets of incompatible results. In order to compare the results and to report token misalignment errors, evaluators devised terminology such as colliding tokens, boundary detection failure and partially tagged tokens. For example, Deleger et al. reported that partially tagged PHIs due to boundary detection errors were 13% of all tagging errors.[7] Some researchers in the NLP community also use complex alignment schemas to remedy the problem.[18] When tokens produced by different systems do not match, the evaluation gets complicated and the differences between results become obscured. The situation gets more complicated as the number of systems to be compared increases. In the literature, we have not seen any proposed solution to the problem for robust evaluation of de-identification systems.

In this study, we align all outputs to be compared to the tokens of the gold standard. This method simplifies the evaluation without introducing any bias favoring one system over another: (1) We re-tokenize all outputs using the

same tokenization scheme that the gold standard annotation has adopted. (2) When a token t in a system output does not correspond one-to-one to a gold standard token t_G, one of the following three scenarios is observed: (a) t may be a proper substring of t_G; (b) t_G may be a proper substring of t; or (c) t and t_G may overlap partially. After re-tokenization, the string of characters in t is distributed into a sequence of one or more tokens. We tag the resulting sequence of tokens with the original tag of t. (3) If t was tagged with a set of multiple tags originally, we apply them simultaneously to all tokens in the resulting sequence.

3.9 Evaluation of system's labeling of single tokens

Accurately distinguishing patient identifiers (e.g., telephone numbers and addresses) from provider identifiers is too difficult to attain for any text de-identification system. Recall that provider identifiers are not PHI. By design, NLM Scrubber (as well as most other clinical text de-identification systems) attempts to de-identify all personal identifiers regardless of whom they belong to. Throwing such a wide net does not degrade the quality of the de-identified text, since provider identifiers usually have no information value for clinical scientists.

Although this approach simplifies our task of catching *all* patient identifiers, it also complicates the evaluation process significantly. How should we evaluate labels of tokens related to providers? For example, if the system labels the physician's phone number as an alphanumeric identifier, should we count this label as true positive (TP)? Conversely, if the system misses a provider's phone number, should we count the instance as false negative (FN)?

When the aim is to protect patient privacy, the performance of a de-identification system should indicate the level of protection it provides. Inflating the TP count with the inclusion of de-identified non-PHI tokens would distort the actual performance. Note that in clinical reports, physician phone numbers are mentioned very frequently but patient phone numbers are extremely rare. Consider the case where a system de-identifies *all* physician phone numbers but misses a few, rarely-occurring patient phone numbers. Had we counted physician phone numbers as TPs, they would totally wash out the system's failure vis-à-vis the missed patient phone numbers. The resulting statistics would give a false diminutive impression on the system's failure on protecting patient privacy.

When we create the gold standard, should we then avoid labeling provider identifiers as PII? In that scenario, whenever a de-identification system recognizes a provider identifier as PII, we would have to count it as false positive (FP) and when it misses the provider identifier, we would have to count it as true negative (TN). Note, the design of our system requires de-identifying all personal identifiers. It would be incongruous to reward the system with a TN count when it misses a provider identifier or to penalize it with a FP count when it finds a provider identifier by following its design requirement faithfully. De-identifying a provider token (or keeping it intact in the text) has no effect on the actual performance of privacy protection or clinical information preservation; thus, we excluded provider tokens from performance evaluation.

The main purpose of our false positive rate analysis is to obtain an indirect measure about the level of preservation of scientific information present in clinical reports from de-identification. Most clinical information of any scientific value is represented in tokens that are not labeled in our gold standard annotation. The only exception to this rule is tokens representing non-PHI age (i.e., age < 90). The rate of preservation of non-labeled tokens and non-PHI age tokens is an indicator for the rate of preservation of clinical information.

If a system assigns a PII label to a token that actually is a patient identifier, we consider the decision as TP. If the system fails to label a patient identifier as PII, we call it as FN. A PII label would be FP if the token has no label in the gold standard (i.e., it is not PII) or if it denotes a non-PHI patient's age, because consequently the token would be redacted and the information would not be available to clinical scientists. If a non-labeled token or a non-PHI age token is labeled as non-PII, we consider it as TN.

3.10 Nonparametric analytic methods

Confidence intervals are staples of biostatics where samples usually come from a well-known parametric distribution and observations are random variables distributed independently and identically, which are not applicable to words in our dataset.

To estimate confidence intervals (CIs) in this study, we adopted a nonparametric bootstrap method,[19] bias-corrected, accelerated (BC_a) percentile intervals as implemented in package *boot* in R.[20] Through a bootstrap resampling strategy, we could truly simulate our initial sampling method. For each bootstrap sample, we randomly selected a patient and then included all reports (hence all associated token sequences) of the patient into the sample. We repeated this process until reaching the same number of patients in our original test data.

We computed statistical significance for the scores where CIs were overlapping, based on Wilcoxon paired signed test with Pratt's adjustments,[21] using the package *coin* in R.[22] This method is more suitable than bootstrap based *p* value estimation because it can successfully take into account that two sets of compared results are paired datasets.

These methods can be used in a wide-range of computational linguistic studies and provide a strong analytic footing for comparisons of different study results. We previously used them to compare and analyze information extraction performances of various systems.[23]

4. Results

In Table 4 on Alphanumeric-Id, Address, and Date rows, we reported each identifier recognition performance separately. For example, on the first row, the alphanumeric identifier recognition performance of NLM-S was isolated from the effects of date and address recognizers of NLM-S on the TP and FN counts. On the PHI rows, however, we reported the total effect of all three recognizers of each system.

Note that for any given system, neither TP nor FN counts on the PHI row are direct sum of the other three rows. For example, the total FN count of NLM-S (excluding PHI row) was 2 + 48 + 311 = 361, but the total missed PHI token count was 201, because some of the missed tokens by one recognizer (e.g., by Date recognizer) were caught by others (e.g., by Alphanumeric identifier recognizer). While we analyze each recognizer in isolation (before evaluating overall de-identification performance), it is important not to lose sight from the overall picture.

4.1 Alphanumeric Identifiers

In alphanumeric identifiers, NLM Scrubber (NLM-S) performance was clearly superior to others. It missed only two tokens, one of which was "406," a 3-digit area code of a telephone number, which should be considered non-PII since the area it covers is the entire state of Montana.[24] The other missed token was a protocol number, which is considered a low risk to privacy as the necessary information to re-identify the patient is not publicly available and such information is usually given to the patient's health care providers only.[25] MITdeid did not produce a viable alphanumeric de-identification on this dataset.

4.2 Addresses and Dates

In both addresses and dates, MIST results yield the best sensitivity scores, but on addresses, the sensitivity score difference between NLM-S and MIST was not statistically significant. After reviewing the false negative cases of NLM-S, we observed that most of the NLM-S's "missed address tokens" were actually non-PHI tokens such as geographical direction (e.g., Northern), state name abbreviation (e.g., VA), large city names in other countries (e.g., Beijing) and country names (e.g., England). None of the missed address tokens revealed a street address, but three of the revealed address tokens may cause some privacy concerns. They were Falls Church (Falls Church, VA: pop. 12,751) and Takoma (Takoma Park, MD: pop. 17,021).

Table 4. De-identification Sensitivity Results of NLM Scrubber (NLM-S), MIT's de-identification system (MITdeid), and MIST: Bold fonts denote the best results among the three, which are also statistically significant if their confidence intervals are written in bold fonts.

Identifier	Gold	System	TP	FN	Sensitivity
Alpha-Numeric-Id	4165	**NLM-S**	4163	2	**1.000 (0.998,1.000)**
		MITdeid	1444	2721	0.347 (0.334,0.359)
		MIST	4091	74	0.982 (0.977,0.986)
Address	292	NLM-S	244	48	0.836 (0.769,0.888)
		MITdeid	129	163	0.442 (0.371,0.510)
		MIST	250	42	**0.856** (0.791,0.905)
Date	29134	NLM-S	28823	311	0.989 (0.984,0.992)
		MITdeid	27595	1539	0.947 (0.942,0.951)
		MIST	28906	228	**0.992 (0.988,0.994)**
PHI	33591	**NLM-S**	33390	201	**0.994 (0.992,0.995)**
		MITdeid	29347	4244	0.874 (0.868,0.879)
		MIST	33310	281	0.992 (0.988,0.994)

In dates, MIST missed fewer date tokens than the other two systems and the differences were statistically significant. The sensitivity performance difference between MIST and NLM-S was 0.003, which however was statistically significant. NLM-S requires further sensitivity improvement on dates. None of the dates by NLM-S was tagged as PHI-Age (i.e., age > 89) in the gold standard.

Although trailing behind the other two systems, MITdeid showed strong sensitivity performance on dates (0.947), but not on addresses (0.442).

4.3 Overall Performance

It is not uncommon that a system tags a PHI token (e.g., a date) with a wrong label (e.g., an alphanumeric identifier). In such cases, there is neither a leakage of PHI nor a loss of clinical information. The PHI row in Table 4 indicates that there were a total of 33,591 PHI alphanumeric, address, and date tokens, of which NLM-S missed only 201, MIST 281 (40% more than NLM-S), and MITdeid 4,244 (21 times as many as NLM-S). NLM-S was clearly superior in overall sensitivity.

The decomposition of the revealed PHI tokens by PII types is displayed in Table 5. Note that the superiority of MIST on date and address de-identification that we observed in Table 4 is not present when we focus only on missed PHI instead of the accuracy of the classification into specific categories.

Table 5. Decompositions of Revealed PHI tokens by System

Tag	PII Type	Gold	NLM-S	MIST	MITdeid
AlphaNumericId	**AlphaNumeric-Id**	3502	0	24	1885
	Protocol-Id	660	1	3	659
	Telecom	3	1	0	0
Address	**Address**	292	48	40	163
Date	**Date**	29124	151	207	1532
	Age 90+	10	0	7	5
All		**33591**	**201**	**281**	**4244**

Although we do not have direct information about the inadvertent loss of clinical information due to over-de-identification, measures such as specificity, precision and F_2 may be used instead as indirect indicators (see Table 6).

Table 6. False Positive (FP), specificity, precision, and F_2 measures of the de-identification systems

System	FP	Specificity	Precision	F_2
NLM-S	5370	0.9950 (0.9947,0.9953)	0.861 (0.853,0.869)	0.964 (0.962,0.967)
MITdeid	2143	**0.9980 (0.9978,0.9982)**	**0.932 (0.926,0.938)**	0.885 (0.880,0.890)
MIST	3143	0.9971 (0.9967,0.9974)	0.914 (0.903,0.922)	**0.975 (0.971,0.978)**

MITdeid was superior in specificity and precision. Due to its low sensitivity score, however, its F_2 measure was not on par with others. In terms of F_2 measure, MIST did better than the other systems and NLM-S was a close second with a 0.011 difference.

5. Discussion

NLM-S revealed much fewer personal identifier tokens than the other two systems (see Table 5) and none of those revealed tokens were informative enough to disclose the identity of any patient. Our primary goal and our main criterion for success are to eliminate all patient related PII tokens when possible. As seen in results, NLM-S incurred a substantial number of false positives in order to catch the maximum number of identifiers, but the specificity penalty it incurred as a result was not higher than 0.005, which means that out of 1000 words only 5 of them were inadvertently deleted. This level of specificity does not have any significant adverse effect on the preservation of

clinical information and on the readability of the resulting text. For keeping the trust of the U.S. Public to the research community, we have to continue working on improving the sensitivity of NLM-S especially on dates and addresses even if it costs us more false positives to achieve that. On the other hand, we are also cognizant of the needs of the research community and have to pay great attention to false positive rates and to the effective conservation of clinical information in the upcoming versions of NLM-S.

In our study data, NLM-S has recognized more PHI tokens than MIST and MITdeid have, which are the only freely available, general-purpose clinical text de-identification systems at the time of this study. Our risk analysis indicates that the revealed tokens would not cause any substantial risk to the patient privacy. Only three instances of address identifiers revealed the home city of three distinct patients, where the population sizes were less than 20,000 but greater than 12,500. Population size 20,000 was devised by the Privacy Rule as a threshold for further censoring zip codes (see Table 1).

MIST was clearly the second best performing system of this study. Due to their underlying methodological power, probabilistic machine learning systems do very well in this domain. Given that we devised our system based on the characteristics of the clinical corpus in our hand, we should not be surprised if MIST outperforms NLM-S in another clinical dataset with different characteristics.

As we indicated in one of our earlier studies,[23] probabilistic machine learning and symbolic linguistic methods are not an either-or proposition, a good NLP system should incorporate methods of both paradigms and reap the benefits of both worlds. We plan to develop a robust machine learning component to our scrubber so that it could perform well on a variety of reports from different origins.

Acknowledgements

We are grateful to the Scientific Counselors of LHNCBC and Dr. Olivier Bodenreider, Chief of Cognitive Science Branch at LHNCBC for their generous inputs and contributions to the project and to an earlier, extended version of this text. We are grateful to Dr. Jon McKeeby, CIO, Clinical Center at NIH and his staff for their help in obtaining and interpreting the clinical data. We thank Guy Divita, Dr. Yanna Kang, Selcuk Ozturk, and Shuang Cai for their contributions to the project. We also thank Drs. Lynette Hirschman, Samuel Bayer and Ben Wellner as well as John Aberdeen of MITRE, for their generous help and offers to test MIST on our study data.

Funding

This work was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

Competing Interests

The first author receives royalties from University of Pittsburgh for his contribution to a de-identification project. NLM's Ethics Office reviewed and approved his appointment.

References

1. U.S. Census Bureau. ZIP code tabulation areas, 2010.
2. Department of Health and Human Services. Public Welfare; Administrative Data Standards and Related Requirements; General Administrative Requirements; General Provisions; Definitions. 45 CFR § 160.103.
3. McCallister E, Grance T, Scarfone K. Guide to protecting the confidentiality of personally identifiable information (PII). Recommendations of the National Institute of Standards and Technology. U.S. Department of Commerce, NIST, 2010.
4. Department of Health and Human Services. Public Welfare; Administrative Data Standards and Related Requirements; Security and Privacy; Privacy of Individually Identifiable Health Information; Other Requirements Relating to Uses and Disclosures of Protected Health Information. 45 CFR § 164.514.
5. Meystre S, Friedlin F, South B, Shen S, Samore M. Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Medical Research Methodology 2010;10(1):70.
6. Benton A, Hill S, Ungar L, Chung A, Leonard C, Freeman C, et al. A system for de-identifying medical message board text. *BMC bioinformatics* 2011;12 Suppl 3:S2.
7. Deleger L, Molnar K, Savova G, Xia F, Lingren T, Li Q, et al. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. J Am Med Inform Assoc 2013;20(1):84-94.

8. Fernandes AC, Cloete D, Broadbent MT, Hayes RD, Chang CK, Jackson RG, et al. Development and evaluation of a de-identification procedure for a case register sourced from mental health electronic records. *BMC Med Inform Decis Mak* 2013;13:71.
9. Ferrández O, South BR, Shen S, Friedlin FJ, Samore MH, Meystre SM. BoB, a best-of-breed automated text de-identification system for VHA clinical documents. J Am Med Inform Assn 2013;20(1):77-83.
10. Beckwith BA, Mahaadevan R, Balis UJ, Kuo F. Development and evaluation of an open source software tool for deidentification of pathology reports. BMC Med Inform Decis Mak 2006;6:12.
11. Berman JJ. Concept-match medical data scrubbing. How pathology text can be used in research. Arch Pathol Lab Med 2003;127(6):680-686.
12. Gardner J, Xiong L. HIDE: An integrated system for health information de-identification. Proceedings of the 2008 21st IEEE International Symposium on Computer-Based Medical Systems 2008:254-259.
13. Neamatullah I. Automated de-identification of free-text medical records. BMC Med Inform Decis Mak 2008;8:32.
14. Wellner B, Huyck M, Mardis S, Aberdeen J, Morgan A, Peshkin L, et al. Rapidly retargetable approaches to de-identification in medical records. J Am Med Inform Assn 2007;14(5):564–573.
15. Uzuner Ö, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. J Am Med Inform Assn 2007;14(5):550-563.
16. Kayaalp M, Browne AC, Callaghan FM, Dodd ZA, Divita G, Ozturk S, et al. The pattern of name tokens in narrative clinical text and a comparison of five systems for redacting them. J Am Med Inform Assn 2013.
17. Divita G, Browne AC, Loane R. dTagger: a POS tagger. *AMIA Annu Symp Proc* 2006:200-3.
18. NTT system description for the WMT 2006 shared task. Workshop on Statistical Machine Translation; 2006; New York, NY. Association for Computational Linguistics.
19. Efron B. Better bootstrap confidence interval. Journal of the American Statistical Association 1987;82(397):171-185.
20. Davison AC, Hinkley DV. Bootstrap methods and their application: Cambridge University Press, 1997.
21. Pratt JW. Remarks on zero and ties in the Wilcoxon signed rank procedures. Journal of the American Statistical Association 1959;54(287):655-667.
22. Hothorn T, Hornik K, van de Wiel MA, Zeileis A. Implementing a class of permutation test: the coin package. Journal of Statistical Software 2008;28(8):1–23.
23. Kang YS, Kayaalp M. Extracting laboratory test information from biomedical text. Journal of Pathology Informatics 2013;4(1):23-35. URL: http://www.jpathinformatics.org/ text.asp?2013/4/1/23/117450. Accessed in 9/3/2013.
24. Wikipedia. Area code 406: Wikipedia, 2013. URL: http://en.wikipedia.org/wiki/Area_ code_406. Accessed in 8/20/2013.
25. Office of Civil Rights. Guidance regarding methods for de-identification of protected health information in accordance with health insurance portability and accountability act (HIPAA) privacy rule, 2012.

Engineering for reliability in at-home chronic disease management

Logan Kendall[1], Jordan Eschler[1], Paula Lozano, MD, MPH[2], Jennifer B. McClure, PhD[2], Lisa M. Vizer, PhD[1], James D. Ralston, MD, MPH[2], Wanda Pratt, PhD[1]

[1]University of Washington, Seattle, WA; [2]Group Health Research Institute, Seattle, WA

Abstract

Individuals with chronic conditions face challenges with maintaining lifelong adherence to self-management activities. Although reminders can help support the cognitive demands of managing daily and future health tasks, we understand little of how they fit into people's daily lives. Utilizing a maximum variation sampling method, we interviewed and compared the experiences of 20 older adults with diabetes and 19 mothers of children with asthma to understand reminder use for at-home chronic disease management. Based on our participants' experiences, we contend that many self-management failures should be viewed as systems failures, rather than individual failures and non-compliance. Furthermore, we identify key principles from reliability engineering that both explain current behavior and suggest strategies to improve patient reminder systems.

Introduction

People with chronic conditions, or those who care for others with chronic conditions, are expected to manage complex medical regimens. For many, the demands are monthly, weekly, or even daily. For example, patients with well-controlled hypertension, diabetes, and hyperlipidemia must perform more than 3,000 health-management activities a year to be adherent to recommended self-care guidelines[1]. These activities include making changes to their diet, requesting medication refills, adhering to a medication regimen, monitoring critical health indicators (e.g. blood pressure), getting lab tests done, attending appointments, getting annual screenings and immunizations, and managing symptoms[2]. As a result of these challenges, many patients are not able to meet their goals for daily self-care activities successfully. In fact, half of individuals diagnosed with a chronic condition—such as asthma, hypertension, and diabetes—do not adequately adhere to their prescribed medication regimens[3] and miss as many as 21-34% of their scheduled appointments[4]. These failures in chronic disease management can lead to adverse patient outcomes, increased care costs, and create challenges to the patient-provider relationship[5–7].

To support patients managing a chronic disease, health care systems increasingly send reminders for appointments and chronic and preventive care activities. Successful reminder systems can alert people to scheduled medical visits and screenings, improve adherence, and enhance communication between patients and their provider team[8,9]. Moreover, many patients already utilize personal reminder systems to remember everyday tasks. The following scenarios help illustrate how people incorporate explicit and implicit reminders into their daily routines:

> *Bob is a 60-year-old diabetic. His typical morning routine is to wake up, walk into his bathroom to take his medication that he keeps out on the counter, then have breakfast. As he walks into the kitchen, he sees his blood glucose meter sitting on the kitchen table and remembers that he needs to check and record his blood sugar. He then starts up his coffee machine, and glances at his wall calendar to see what is happening that week. While eating breakfast, he receives a call from his clinic reminding him about an appointment the next day. Later on, his wife, who is out of town, calls him to check-in. Realizing he forgot to check his blood sugar while making coffee, he walks back to the kitchen to get his glucose meter.*

> *Cindy is a 34-year-old mother of a 9-year old child with persistent asthma. When she wakes up, she always glances at her smartphone's calendar to see her agenda for the day. She also takes a minute to enter in a to-do list. Among other things, her son needs his allergy shot at the clinic and she needs to send off her sister's birthday package. Before heading out, Cindy notices and grabs her son's inhaler by the door to make sure it gets into his backpack for school that day. In the rush to get her son to school from the appointment, she forgets to stop by the post office. But on her way home later that day, Cindy passes a mailbox that reminds her to send off the package and she makes a turn to the nearest post office.*

These vignettes, based on activities and experiences described by our study participants, demonstrate how people rely on a variety of tools and subtle triggers to help remember to perform future actions. In the first scenario, Bob relies mostly on environmental cues that are part of his morning routine. However, his wife will frequently check-in with him, which serves as a backup in case he does forget to do something. In contrast, Cindy makes heavy use of her

mobile phone to track what she needs to do for the day. She also makes deliberate use of visual cues like placing the inhaler by the door to make sure her son takes his inhaler. Yet, in both situations, the two people still experienced minor failures in achieving their intended tasks.

In this paper, we examine how individuals responsible for managing their own or others' chronic conditions integrate reminders and notification systems into their daily routines. By understanding these diverse individual experiences, we hope to gain insight into the optimal design characteristics for future patient reminder and support systems. We further contend that many self-care management failures may be accurately viewed as system failures, as opposed to failures of individuals[10]. Our work highlights the complex ecosystem of interactions, tasks, and reminders between the clinic and home environment for a person managing a chronic illness. Finally, we apply key principles from reliability engineering to help explain participants' self-management behavior and offer suggestions for strategies to further improve patient reminder systems in the future.

Background and Related Work

This work draws on several divergent literatures: prospective memory as a basis for task planning and recall, clinical and personal reminder systems, and systems reliability engineering to frame patient self-care tasks and failures.

Prospective Memory as a Basis for Task Planning and Recall

Remembering to perform all the tasks expected for proper self-management requires effective recall of what has already happened and a continuous scan of what needs to happen in the near future. The process of remembering is frequently framed as either of two types: (1) retrospective memory that is concerned with the retrieval of past memories of people, events, and words, or (2) prospective memory that is concerned with remembering to perform a planned action or intention in the future[11]. The latter process includes short-term intentions—such as daily intake of a medication—as well as delayed actions—such as going to an annual checkup appointment—that could occur weeks or months in the future. Outlined in Figure 1, the process for realizing a delayed intention begins with encoding the future action, retaining the intention, and then retrieving the intention at the appropriate time to complete the action. This can occur through either an explicit reminder system or through spontaneous retrieval. Actions such as remembering to take medication at breakfast often rely on spontaneous retrieval of the intention that is triggered through environmental and physiological cues linked to daily routines. However, intermittent actions further out in time often involve a more explicit signaling cue—such as creating an alarm on a phone—to retrieve and execute the action at the right time[12]. In the case of an individual managing a chronic condition, the capacity to reliably shape and direct future behavior is critical to successfully managing the disease. The role of both explicit and implicit reminder systems within this memory process is the focus of this paper.

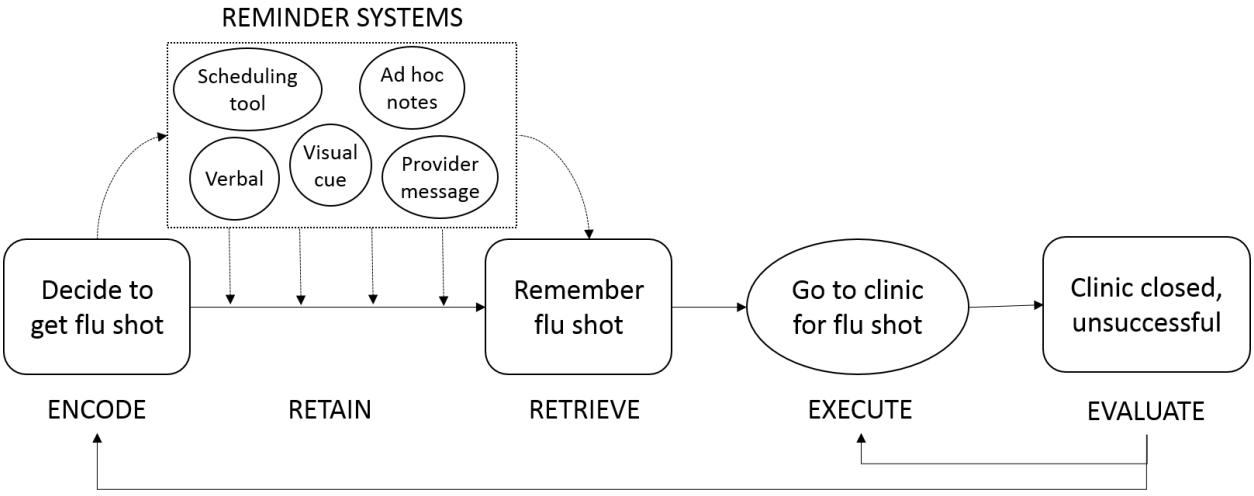


Figure 1. Model for establishing and realizing delayed intentions. Adapted from Ellis, 1996 [13]. When a person decides to get an annual flu shot (encoding), they need to maintain that intention over a delayed period (retention). At some point the person receives a signal to remember the flu shot (retrieval) and then goes to the clinic (execution). If successful (evaluation), the task is complete. Otherwise, the person attempts the activity again or re-plans the intention. In this model reminder systems provide a way for individuals to externalize the retention process as well as establish a way to cue retrieval of the delayed intention.

Clinic and Personal Reminder Systems to Support Recall

Reminders can be useful mechanisms to support the execution of delayed intentions described by prospective memory. James Reason describes the value of reminders in mitigating errors of omission that can lead to failures in planning and intention formation (encoding), intention storage (retention), action execution, and monitoring[14]. Research into reminders for health has typically evaluated clinic use of reminder systems. For example, health care organizations use postal mail or telephone calls to help prevent missed appointments and thereby reduce costs associated with disruptions to clinic workflow[15]. More recent articles have started to explore mobile phone-based strategies, such as text messaging and other types of electronic reminders[8,16]. Vervloet's review of electronic reminder systems initiated by providers showed short-term effectiveness of these systems in supporting medication adherence but suggested more investigation of reminder content and timing[8]. In a cross-sectional survey of patient preferences, Finkelstein et al. found that personalizing the delivery mechanism improved patients' responsiveness to reminders[4]. The research into clinic-initiated reminder systems indicates that they support improvements in patient adherence, are effective across diverse technologies, and are more likely to be successful through the personalization of message content based on patient preferences.

Outside of a clinic setting, individuals leverage a variety of tools for personal reminder systems. Grimes and Brush interviewed working parents and identified a number of challenges that they experienced in coordinating and interacting across their personal and professional schedules[17]. In a separate study, Brush highlighted the value of triggers in the workplace that are created from a mixture of explicit self-reminder systems, tacit "breadcrumbs" from recent activities, or based solely on memory to recall specific tasks[18]. However, there has been limited research into how these types of personal systems support patients' multiple chronic and preventive health care needs. For example, a review by Kapur et al. focused on the use of memory aids for neurological rehabilitation of individuals with severe memory impairments due to injury or a disorder[19]. While reminder systems are prevalent in a variety of contexts, there is an opportunity to explore how personal systems are used for chronic disease management, and importantly, how these tools integrate with clinic reminder activities and the home environment.

Systems Reliability Engineering to Understand Errors and Failures

Within this paper, we frame chronic disease management as a complex system of technical, organizational, economic, environmental, and human components that interact for a common purpose—the maintenance and coordination of an individual's health. This type of system highlights a growing trend in health care of utilizing human factors and ergonomics (HFE) concepts for designing patient-centered care. The National Research Council conducted a HFE evaluation of home health care that stressed systems engineering for designing technology interventions that facilitate interactions between the complex elements and tasks in the home environment[20]. Furthermore, Holden, Carayon, and colleagues introduced a popular framework for HFE thinking and reliability in health care with the Systems Engineering Initiative for Patient Safety (SEIPS) model[21]. The latest version of this model discusses the importance of a person-centered approach to work systems that include organizations, tools & technology, people, tasks, and the internal and external environments where the work takes place. In the context of reminders, the activity of remembering and executing self-management tasks at home involves the interplay between the patient, clinic, home environment, and devices and tools used to coordinate and execute care activities.

Building on HFE concepts, we are concerned with the reliability of the system to support individuals' ability to use reminders to perform their care coordination tasks successfully. The principles of reliability science and engineering help to describe design strategies for mitigating and responding to failures in complex systems. Reliability refers to the probability that a system will meet its minimum performance requirements, without failure, for a given period of time[22]. Three activities in particular—engineered redundancy, diverse and independent design, and regular monitoring—can enhance the reliability of a system[23]. **Redundancy** describes systems where duplicate processes or components are used to increase availability in case of a single point of failure. For example, many cars are designed with twin hydraulic brake circuits to ensure that the vehicle can still stop if one circuit fails. However, redundant systems do not increase reliability in situations where they are vulnerable to the same defect—referred to as common cause failure. To mitigate this vulnerability, engineers incorporate **diversity** into their design so that a system uses two or more different, independent techniques or processes for the same functional purpose[24]. For example, when backing up a car, a driver can visually assess their environment with the car mirrors as well as listen to the beeping sounds produced by proximity sensors built into many modern vehicles. Finally, reliable systems can use **monitoring** to actively audit the system and mitigate the impact of a breakdown through early detection of failures. Modern automotive anti-lock systems actively monitor wheel deceleration and distance from other objects to adjust the brake speed and prevent uncontrolled skidding. Although researchers have used reliability engineering to frame inpatient safety[25], none have used this framework to examine patient adherence to chronic disease management activities.

Study Overview

We conducted a series of semi-structured, in-depth interviews with patients diagnosed with Type 2 diabetes (n = 20) and mothers of children receiving treatment for asthma (n = 19). The choice of the two populations maximized the variation in the perspectives and experiences of individuals managing a chronic disease. Our sampling of parents of children with asthma focused on mothers, rather than fathers, since women in this age group also have a large number of prevention and maintenance activities and are typically the primary health information managers within the household[26]. The semi-structured interviews focused on reminder tools and systems that patients used for maintenance and care coordination activities such as appointment scheduling, medication adherence, and communication with their providers outside of the clinic. In addition to the interviews, the authors (LK, JE) toured the participants' homes in order to directly observe the systems and tools used in this context. This study was reviewed and approved by our institution's human subjects review board.

Sampling

We used purposeful sampling to identify participants that were representative of the general population in the Northwest United States based on gender, ethnicity, technology use (with recorded use of a patient portal as a proxy) and education. All participants were enrollees of an integrated healthcare delivery system that covers more than 300,000 members. Based on the sampling criteria, we contacted 586 individuals identified as either persons with diabetes or parents of a child with asthma. Of these, 402 could not be reached, refused to participate, or were lost to follow-up, and 118 were deemed ineligible based on follow-up screening. Of the remaining 66, we enrolled 39 participants, aged 27 to 88 (median=61). The diabetes cohort was older on average than the mothers of children with asthma and were less likely to use technology such as a patient portal for their health. Other details on our participant population are in Table 1. Each participant received a gratuity of $50 for completing the interview.

Table 1. Description of study participants.

	Diabetes		Asthma		Combined	
	N	**%**	**N**	**%**	**N**	**%**
Participants	20	100	19	100	39	100
Female	10	50	19	100	29	74
Race/Ethnicity						
White	10	50	9	47	19	49
Black	6	30	6	32	12	31
Asian	2	10	1	5	3	8
Hispanic	1	5	1	5	2	5
Mixed/Other	1	5	2	11	3	8
Technology Use						
Patient Portal User	13	65	18	95	31	79
Education						
High School or Less	12	60	5	26	17	44
More than High School	8	40	15	79	23	59

Analysis

The interviews lasted from 40 to 90 minutes. The audio recordings from each interview were transcribed and reviewed using Atlas.ti 7 and identified as "A##" and "D##" for the asthma and diabetes cohorts, respectively. To maintain confidentiality, we removed all information that could identify the patient from the transcripts. During the data collection process, the authors periodically reviewed transcripts to identify emerging themes and to assess topic saturation. Using an open coding technique and an affinity diagramming process[27], the authors clustered related terms and organized preliminary themes into higher-level categories. A subset of the authors then iterated on the codes through several rounds of transcript review to solidify the coding schema used for this article. One author used this schema to code each interview using Atlas.ti.

Results

Through this content analysis, we focused on two areas. First, we highlighted the reasons for failures in self-management routines as identified by participants during the interviews. Second, we characterized three strategies used by participants to improve the reliability of their reminder systems and overall management of their health. These strategies are summarized in Table 2.

Sources of Reminder Errors and Task Failures

Throughout the study, the participants provided examples of breakdowns in their self-management routines that included failures in memory, missed environmental signals, and failures in capturing tasks within a reminder system.

Memory failures were a common reason for failure as well as a rationale for using explicit reminder systems. Based on the prospective memory process in Figure 1, these types of errors can be described as retention and retrieval failures. For example, participants described how easy it was to forget because of their hectic schedules. A07 expresses the challenges representative of many busy mothers. *"If I don't have anything immediately reminding me of it, it's out of my head because I have so much going on. We have kids to pick up, drop off. We have cleaning house, I've got selling things, I've got to meet people." [A07]*. D02 and many elderly participants with diabetes described concerns with growing memory deficits as they aged. They expressed interest in updating their reminder systems as their memory started to become less reliable. Despite this interest, participants often relied on informal, tacit signals in the environment. For example, it was only after scheduling a separate appointment for herself that A08 remembered her child's annual well-child visit was overdue. Similarly, other participants remembered to go in for screenings and tests only after scheduling a visit for a separate health need. Reliance on environmental cues was particularly common for medication management. One participant described managing her child's inhaler by paying attention to when "*the propellant in it doesn't expel quite as well and so he doesn't feel like he's getting or receiving the medication as well and so he'll tell us that he needs a new one" [A11]*. Without having an explicit visual cue or timely notification from her child, she relies on trying to remember how much time has passed since the last refill. Tacit signals are dependent on the environmental cue occurring at the right time and place. If the person misses this signal, they can experience a retrieval failure.

In addition to missing cues from informal reminders, participants also described breakdowns with their reminder tools. In some situations the reminder was never captured and therefore never signaled the appropriate behavior. This type of encoding failure often occurs because individuals get distracted from capturing the task or do not have access to their reminder tool when they need it. One participant described how, "*I'll make an appointment and then start doing something else so I didn't write it down and just hope that I'll remember it." [A10]*. In this situation, A10 missed an appointment for her baby because the event did not get recorded into her phone and she did not get a reminder call. A14 used a system where she printed out calendars and captured appointments, bills, and other items. However, when she misplaced it, *"I was freaking out, because I had actually taken the time to write everything down—they had a trove of information on it. I couldn't find it and then I was like how the heck am I going to do this if I don't have a backup?" [A14]*. Her situation mirrored many of the participants where, because of a busy schedule, being away from home, or not having access to their normal reminder tool, they failed to set up or retrieve the cue from their system to perform the action in the future. Despite having systems that they often described as being fairly reliable, participants still experienced situations in which some aspect of their reminder process fell apart.

Table 2. Summary of reliability system design strategies used by participants to enhance self-care reminders.

Reliability Strategy	Benefits	Drawbacks
Redundancy: *Repeated reminders, back-up systems*	• Back-ups ensure availability of reminder for retrieval at the right time. • Repeated alerts enhance retention.	• Too many back-up systems can create unwanted noise and increase chance of being ignored.
Diversity: *Independent systems, different communication modes*	• Separate clinic and personal systems reduces reliance on single operator. • Multiple modalities improves availability of reminder for retrieval.	• Additional systems increase complexity and can create new opportunities for failure.
Monitoring: *Validating reminder capture, double-checking for errors*	• Improves retention through repeated exposure. • Increases likelihood of catching errors in the reminder system.	• Additional work required of system operators • Challenges with maintaining this behavior routinely.

Strategies for Redundancy in Reminder Systems

To compensate for these types of memory and organizational errors, participants employed a number of strategies to improve the reliability of remembering certain tasks for their care management. In particular, to mitigate the chance that one reminder system might fail, participants integrated duplicative or redundant reminder systems in their homes. In many cases, having redundancy was simply a backup. D04 used the calendar on her computer to manage most of her reminders. However, she purchased a smartphone because she recently got a virus on her computer and was concerned about its reliability in the future. Other individuals in the study described how having their spouse or children aware of their health management tasks served as a useful backup. They would regularly discuss upcoming events with the family member or ask them to check-in to make sure the participant took their medication. Even though it was rarely necessary, *"they're there if we do need them" [D08]*.

Participants expressed that having multiple, redundant systems helped to ensure that reminders get encoded. D02 maintained three calendars so that the intended task was captured regardless of where she was in the home. An added value of having multiple systems was that it helped to retain what she need to do. *"It's like repetition. If you say something to me and say the same thing several times, it rings a bell and I remember it. I think that's what the calendar does." [D02]*. This system served as both a way to deal with memory deficits as well as ensure that upcoming events are always captured. A09 was worried that her husband would forget to pick up the kids and so "*I told him the night before, I told him the morning of, I sent him a text from work, I called him to make sure - don't forget." [A09]*. While that many reminders were likely unnecessary, it reassured the participant and increased her spouse's awareness.

Although redundant systems have advantages, they can be a burden as well. Providers are typically very proactive about letting a patient know of an upcoming appointment to reduce the impact of a missed appointment on their clinic workflow. Some participants expressed an interest in streamlining their different reminder systems—from both their providers as well as their personal reminder tools—to remove duplicative work and unnecessary reminders.

> *"My son, his dentist office annoys the heck out of me because they send reminders two months in advance and then send another one a month in advance and at that time they want you to click confirm on the appointment and then after you confirm it, they send another one a week in advance and then call you. It's overkill." [A09]*

While repeated reminders increases the chance that the patient is aware of the appointment or task, it also creates more noise. The participants described many redundant reminders as unhelpful in situations where they already were aware of the appointment or task or it was not applicable to their needs. If they felt that it was an unnecessary reminder, they would simply ignore it. This situation is analogous to the alert fatigue that health care providers face with poorly designed EMR and other practice management systems.

Strategies for Diversity in Reminder Systems

The individuals in this study not only performed redundant work as a protection against failures, but also used a diverse set of reminder systems (Figure 2). Diversity provides greater protection against system failures that can affect even redundant systems. For example, a clinic that sends multiple appointment reminder notices in the mail will fail if the patient's address is incorrect. A phone that runs out of power will not be able to receive a text message about taking a medication. Throughout the interviews, the study participants described an array of modalities used for communication and reminders about upcoming tasks.

Figure 2. Sample of personal reminder systems used by study participants. From left to right: a whiteboard-style calendar; a sticky note wrapped around the strap of a handbag; pill bottles placed upside down to indicate they have been taken for the day.

All of the study participants emphasized the value of traditional reminders initiated by the provider such as paper mail, phone calls, and secure messaging within a patient portal. However, participants still used a number of additional, personal reminder systems that ranged from technology-centric approaches—such as a mobile phone—to informal mechanisms such as the placement of a pillbox on the kitchen table. A06 would get multiple text messages and phone calls from her dentist leading up to an appointment, but would still enter it into her phone right away. The major tenet of the diversity principle is to ensure independence in redundant systems and subsystems. Thus, by incorporating reminders from clinics as well as their own systems, patients add a layer of reliability—they are able to rely on the provider messages if their personal system fails or rely on their own system if they do not receive a reminder from their clinic. D01 described an incident where she had a scheduled phone consult with her doctor. She recorded the appointment in her personal planner. However, the clinic did not record the phone call information correctly, and the physician didn't call. D01 ended up having to call and remind the clinic about her phone appointment.

During the interviews, participants also explained how the mode of communication can have a meaningful impact. *"The phone call I think it's an accountability thing for me, if I actually physically speak to a person, it's easier for me to go okay, I need to write it down and remember it." [A03]*. Synchronous communication modalities, particularly where the patient is interacting with a health professional, make it difficult to ignore the reminder and ensure that the reminder reaches its intended audience. Conversely, other participants discussed how it was easy to overlook or disregard asynchronous, electronic messages like emails because they get overloaded with too many messages in that format. Paper notifications were often valued because they are more tangible and visible around the home. *"Well again, because I don't get a lot of paper notifications, it feels serious and this should be done." [A01]*

Diversity can enhance reliability, but it can also create a chaotic and complex experience as the patient tries to integrate multiple systems. Several of the participants, for example, owned multiple computing devices that each had a different, incompatible operating system. This creates barriers in syncing activities across the platforms and in some situations encouraged participants to rely on a paper method instead. Additionally, participants that received care from multiple physicians described challenges with getting phone calls from some clinics and paper mail or electronic notifications from others. The diversity of communication approaches made it challenging to reliably integrate the reminders into their personal systems.

Strategies for Monitoring Reminder Systems

Study participants incorporated a habit of active monitoring of their reminder systems as an additional method for addressing reminder failures. Monitoring provided a way to identify errors, validate their reminder systems, and involve their friends and family members in supporting self-management. First, routine monitoring helps to identify when failures in self-management occurred.

> *"I have to be more conscious about did I take my pills…Did I do that? Sometimes I'll actually get up and look to see whether I took my pills. It's kind of like turning your headlights off on your car, yeah, I took my pills but when I stop and think about it, I got to go check because I don't remember doing it." [D20]*

Even if a participant forgets to do something like take a medication, the process of monitoring can lead the person to recognize the error and then be in a position to make changes for the future. Although a clinic can notify a patient when they miss an appointment or test, no feedback process informs patients when they forget activities at home. Having a system in place for monitoring can be helpful to evaluate one's behavior.

Second, active monitoring serves as a way of double-checking the reminder system and ensuring the right content is captured. The added benefit is that this helped participants retain what they needed to do in the future. Participants described deliberate efforts to consistently review and scan for future activities. *"I try to check the next couple days ahead, like I'm thinking today plus two or something so I'm aware what's coming." [A09]*. This participant also started involving her son this process by setting up a calendar in his room so he could cross off activities as they occurred. *"If I get an appointment, then I write it down...I go in [the patient website], because it shows upcoming appointments so then I'll print off the deal and go in there and check my calendar, so it's a backup type, so I got a system, my checks and balances." [D05]*. Redundant, diverse reminder systems are helpful in supporting this kind of auditing behavior by helping participants compare the content of a reminder from multiple sources.

Third, participants often shared the responsibility of monitoring activities across family members to help reduce the workload. This demonstrated a distributed process for auditing reminder systems and ensuring completion of self-care behaviors. Informal dialogue between spouses or between parent and child appeared to be a subtle but valuable mechanism for reviewing and validating upcoming appointments, medications, and tasks on a regular basis. A01 spoke about how she had transitioned from administering her child's inhaler directly to watching him do it himself. While

this created other challenges around knowing if he took the inhaler correctly, A01 experience exemplified a transition in self-care responsibility common to many of the mothers of children with asthma.

Active monitoring does require additional work by participants and may be onerous when considering the activities already required for chronic disease management. Moreover, participants mentioned difficulties with making sure this type of monitoring is routine. A wall calendar with appointments listed is only useful as a reminder of upcoming events if the target audience makes the effort to review and validate the content. One participant assumed she had the correct information for an appointment on her wall calendar, but realized she had written the appointment on the wrong week after her husband walked by the calendar and pointed it out [D08]. Wall calendars contrast with systems like email and text messaging that will push information to the target based on a predefined event. However, sometimes these push systems actually discourage any active monitoring of a reminder system. *"I guess the hope is that it'll perform how I've asked it to, or that I've remembered to ask it to perform how I want it to." [A08]*. Another participant described how he messed up an appointment twice because he did not look back at his email about the appointment and his did not have his smartphone set properly to alert him [D16]. Becoming too dependent on the reminder system's capacity to create notifications can create a new avenue for failure.

Discussion

Forgetfulness, confusion, external distractors, fatigue and even a person's health condition can all lead to failures in self-management and perception of non-adherence. These types of errors, referred to as slips in human factors research, describe instances where an individual's intentions get waylaid en route[28]. Although slips are often small failures, they are particularly relevant to breakdowns in the automatic, routine behavior that is at the heart of chronic disease management. Solely relying on memory or informal environmental cues may be adequate for a period of time, but if an individual does not retrieve the prospective action accurately or at the right time, no backup will ensure the activity happens. We observed participants adopting a variety of strategies—such as redundancy, diversity, and active monitoring behaviors—to improve the reliability of managing self-care responsibilities. Incorporating different systems, people, and modes of communication ensured multiple, repeated communication paths for reminders in the event of one system failing. Moreover, participants did not just rely on different tools and modalities, but continually audited these systems to ensure that they had the correct information and that activities were completed successfully.

The challenge is that incorporating redundancy and diversity across multiple, separate systems involves tradeoffs between reliability and complexity. Diversity inherently increases complexity, and the need for synchronization among multiple systems could create more opportunities for failure. Furthermore, as participants in this study described, there is a risk of noise fatigue when dealing with multiple reminders from multiple sources. Health care organizations need to be aware of the additional workload placed on patients at home and reduce complexity through more tailored communication and easier integration of clinic-based reminders with patient reminder systems. By evaluating chronic disease self-management through a systems lens, we argue that trying to fix breakdowns in self-management should focus on designing system-level changes that focus on the experience of patients.

Designing for Human Error

Importantly for systems-thinking, our study highlights the variety of ways that failures can occur in remembering to perform self-care management activities. Therefore, the design of reminder systems to support self-management should account for errors by making it easier to detect, evaluate, and respond to failures when they do occur. The concepts of redundancy, diversity, and monitoring represent system design concepts that, integrated with the prospective memory process (Figure 1), can inform approaches to the future design of reminder systems. In addition, well-established systems engineering methodologies and tools can support this design and evaluation process. For example, concurrent engineering and quality functional deployment (QFD) use methods similar to participatory design to explicitly capture all stakeholder needs in a complex system[29], including less visible needs—such as the personal reminder work of patients with chronic diseases. The SEIPS system model further reinforces the importance of design that incorporates the needs of patients and caregivers involved in chronic disease management work[21].

James Reason suggests that all reminders should meet certain universal criteria so that they are conspicuous at the right time, contiguous or available in time and space, provide the necessary context and content for the intended actions or tasks that need to be done[14]. While the patient reminder systems in this study incorporated many of these qualities and avoided errors in encoding, retention, and retrieval through redundant and diverse design, they often lacked a feedback loop to track if an activity was performed or to support evaluation of errors that may have occurred. An important aspect of high reliability systems is the practice of assessing failures in order to actively identify, correct, or mitigate the sources of failure in the future. Assessing variation in self-care management and if performance is

within acceptable boundaries requires better tools and processes for capturing the metrics and data to calculate the variation. Information technology has the ability to support more robust monitoring and learning through different notification processes, passive data collection on behaviors/activity, and enhanced summary reports and real-time feedback. Participants across both cohorts leveraged mobile phones and other computing devices to support their self-management efforts. Increasingly, mobile phones link with ubiquitous sensing tools to support automated, detailed tracking of health metrics and performance around daily activities. These sensing tools can reduce patient work, integrate diverse metrics, provide feedback on progress towards achieving health goals, and identify deviations that are the result of errors or other failures. However, systems thinking and systems engineering principles should be considered in the design and the use of these tools so that they integrate in the wider context of strategies for more reliable self-care management.

Study Limitations

As with any qualitative study, our findings might not be fully representative of the populations under consideration. Despite making considerable effort to sample for a representative patient population, it is possible that our participants differ from other patients in terms of self-efficacy or their organization with managing their care. Because the diabetic population in this study encompassed a narrow demographic of mostly elderly, retired individuals, it is possible that we did not adequately capture a wide enough array of experiences and strategies for managing diabetes. However, we were able to contrast their experiences with those of young mothers taking care of their children with asthma. These mothers were often working and were more likely to use technology. Finally, the way that the participants used reminder systems in this study could reflect disease-specific needs of our two cohorts.

Conclusion and Future Work

In summary, the experiences of our participants managing chronic disease at home highlights the diverse strategies they employ to manage their schedules and tasks. Moreover, even though many clinics and health systems have implemented reminder systems—such as follow-up phone calls—our study showed that patients must still do extensive work at home to integrate these reminders into their daily lives. These separate reminders enhance reliability through increased diversity, but also add to the overall complexity of the system. Participants often felt confident in relying on their memory for most routine needs, yet still valued redundant reminders as a backup to deal with any memory slips. We also note that in addition to formal systems—such as calendars—many participants in this study relied on subtle systems similar to what Donald Norman refers to as "knowledge in the world"[28]. These cues that are visible in the environment and trigger prospective memories are less visible but important to be aware of with evaluating reminders and support systems for chronic care management. Technologies that recognize and integrate with these tacit signals have the potential to provide more context-sensitive reminders.

Our findings also support evaluating failures in self-management from a systems perspective, rather than simply attributing a failure to individual's lack of responsibility. While non-adherence is a legitimate issue, our participants' experiences make it clear that some self-care failures are unintentional and can best be characterized as breakdowns in the interaction between people, machines, and environments within a system. To mitigate failures, many patients incorporated key characteristics from reliability science into their personal reminder systems—characteristics such as redundancy, diversity, and monitoring behaviors. We are not aware of other studies that have examined these strategies from a system reliability perspective. As future work seeks to support patients' ability to manage chronic conditions outside of the clinic, it will be important to design self-care and reminder tools that also capitalize on these reliability principles. Healthcare providers, systems, and designers should consider the use of engineering design, evaluation, and control methods to explore this subsystem of patient reminder work that is prevalent among individuals with chronic conditions. Greater understanding of how this patient work integrates with clinic workflows and programs can lead to more reliable care management and ideally improved outcomes.

Acknowledgements

This project was supported by grant # R01HS021590 from the Agency for Healthcare Research and Quality (AHRQ).

References

1. Steiner JF. Rethinking adherence. Ann Intern Med. 2012 Oct 16;157(8):580–5.
2. Barlow J, Wright C, Sheasby J, Turner A, Hainsworth J. Self-management approaches for people with chronic conditions: a review. Patient Educ Couns. 2002;48(2):177–87.
3. Sabaté E. Adherence to long-term therapies: evidence for action. Geneva, Switzerland: World Health Organization; 2003.

4. Finkelstein SR, Liu N, Jani B, Rosenthal D, Poghosyan L. Appointment reminder systems and patient preferences: Patient technology usage and familiarity with other service providers as predictive variables. Health Informatics J. 2013 Jun;19(2):79–90.
5. Junod Perron N, Dao MD, Righini NC, Humair J-P, Broers B, Narring F, et al. Text-messaging versus telephone reminders to reduce missed appointments in an academic primary care clinic: a randomized controlled trial. BMC Health Serv Res. 2013 Jan;13:125.
6. Schectman JM, Schorling JB, Voss JD. Appointment adherence and disparities in outcomes among patients with diabetes. J Gen Intern Med. 2008 Oct;23(10):1685–7.
7. Hussain-Gambles M, Neal RD, Dempsey O, Lawlor D a, Hodgson J. Missed appointments in primary care: questionnaire and focus group study of health professionals. Br J Gen Pract. 2004 Feb;54(499):108–13.
8. Vervloet M, Linn AJ, van Weert JCM, de Bakker DH, Bouvy ML, van Dijk L. The effectiveness of interventions using electronic reminders to improve adherence to chronic medication: a systematic review of the literature. J Am Med Inform Assoc. 2012;19(5):696–704.
9. Szilagyi PG, Bordley C, Vann JC, Chelminski A, Kraus RM, Margolis P a, et al. Effect of patient reminder/recall interventions on immunization rates: A review. JAMA. 2000 Oct 11;284(14):1820–7.
10. Steiner JF. Rethinking adherence. Ann Intern Med. 2012 Oct;157(8):580–5.
11. Kliegel M, McDaniel MA, Einstein GO, editors. Prospective Memory: Cognitive, Neuroscience, Developmental, and Applied Perspectives. Taylor & Francis Group; 2008.
12. Daniel MAMC, Einstein GO. Strategic and Automatic Processes in Prospective Memory Retrieval : A Multiprocess Framework. 2000;144(September):127–44.
13. Ellis J. Retrieval Cue Specificity and the Realization of Delayed Intentions. Q J Exp Psychol Sect A. 1996 Nov;49(4):862–87.
14. Reason J. Combating omission errors through task analysis and good reminders. Qual Saf Health Care. 2002 Mar;11(1):40–4.
15. Hashim MJ, Franks P, Fiscella K. Effectiveness of telephone reminders in improving rate of appointments kept at an outpatient clinic: a randomized controlled trial. J Am Board Fam Pract. 2001;14(3):193–6.
16. Gurol-Urganci I, de Jongh T, Vodopivec-Jamsek V, Atun R, Car J. Mobile phone messaging reminders for attendance at healthcare appointments. Cochrane database Syst Rev. 2013 Jan;12(12):CD007458.
17. Grimes A, Brush AJ. Life scheduling to support multiple social roles. Proceeding of the ACM CHI conference 2008. Florence, Italy: ACM Press; 2008. p. 821.
18. Brush AJB, Meyers BR, Tan DS, Czerwinski M. Understanding memory triggers for task tracking. Proceedings of the AMC CHI conference 2007. San Jose, CA: ACM Press; 2007. p. 947.
19. Kapur N, Glisky EL, Wilson B a. Technological memory aids for people with memory deficits. Neuropsychol Rehabil. 2004 Mar;14(1-2):41–60.
20. National Research Council. Health Care Comes Home: The Human Factors. Washington, DC: The National Academies Press; 2011.
21. Holden RJ, Carayon P, Gurses AP, Hoonakker P, Hundt AS, Ozok AA, et al. SEIPS 2.0: a human factors framework for studying and improving the work of healthcare professionals and patients. Ergonomics. 2013 Jan;56(11):1669–86.
22. Zio E. Reliability engineering: Old problems and new challenges. Reliab Eng Syst Saf. 2009 Feb;94(2):125–41.
23. International Atomic Energy Agency. Protecting against Common Cause Failures in Digital I & C Systems of Nuclear Power Plants. Vienna: International Atomic Energy Agency; 2009.
24. Littlewood B, Strigini L. Redundancy and Diversity in Security. 9th European Symposium on Research in Computer Security. Sophia Antipolis, France; 2004. p. 423–38.
25. Luria JW, Muething SE, Schoettker PJ, Kotagal UR. Reliability science and patient safety. Pediatr Clin North Am. 2006 Dec;53(6):1121–33.
26. Moen A, Brennan PF. Health@Home: the work of health information management in the household (HIMH): implications for consumer health informatics (CHI) innovations. J Am Med Inform Assoc. 12(6):648–56.
27. Martin B, Hanington BM. Universal methods of design: 100 ways to research complex problems, develop innovative ideas, and design effective solutions. Rockport Publishers; 2012.
28. Norman DA. The Design of Everyday Things. New York: Basic books; 2002.
29. Reid PP, Compton WD, Grossman JH, Fanijang G, editors. Building a Better Delivery System: A New Engineering/Health Care Partnership. Washington, DC: The National Academies Press; 2005.

Automatic Extraction of Drug Indications from FDA Drug Labels

Ritu Khare[1], PhD, Chih-Hsuan Wei[1], PhD, Zhiyong Lu[1], PhD
[1]National Center for Biotechnology Information (NCBI), NIH, Bethesda, MD 20894

Abstract

Extracting computable indications, i.e. drug-disease treatment relationships, from narrative drug resources is the key for building a gold standard drug indication repository. The two steps to the extraction problem are disease named-entity recognition (NER) to identify disease mentions from a free-text description and disease classification to distinguish indications from other disease mentions in the description. While there exist many tools for disease NER, disease classification is mostly achieved through human annotations. For example, we recently resorted to human annotations to prepare a corpus, LabeledIn, capturing structured indications from the drug labels submitted to FDA by pharmaceutical companies. In this study, we present an automatic end-to-end framework to extract structured and normalized indications from FDA drug labels. In addition to automatic disease NER, a key component of our framework is a machine learning method that is trained on the LabeledIn corpus to classify the NER-computed disease mentions as "indication vs. non-indication." Through experiments with 500 drug labels, our end-to-end system delivered 86.3% F1-measure in drug indication extraction, with 17% improvement over baseline. Further analysis shows that the indication classifier delivers a performance comparable to human experts and that the remaining errors are mostly due to disease NER (more than 50%). Given its performance, we conclude that our end-to-end approach has the potential to significantly reduce human annotation costs.

Introduction

Drug-disease treatment relationships, i.e. drugs and their indications, are among the top searched topics in PubMed[1, 2]. The primary application of this information is to inform healthcare professionals and patients for questions[3] like "what are the indications of fluoxetine." Such information is also valuable for developing computational methods for predicting and validating results of novel drug indications[4-6] and drug side effects[7], controlling data-entry and medication errors in the electronic medical records[8-10], feeding the Google Knowledge Graph, and assisting PubMed Health® (http://www.ncbi.nlm.nih.gov/pubmedhealth/) editors to cross-link drug and disease monographs[11]. Given the variety of applications, it is important to build a structured and normalized drug indication repository, or a computable "gold standard" of drug-disease relationships, from credible drug resources. Most existing high-quality drug resources, such as DailyMed[12], DrugBank[13], MedlinePlus[14], and MedicineNet[15] are described in free text. Figure 1 presents excerpts from DailyMed descriptions, aka *FDA drug labels*, for two drugs.

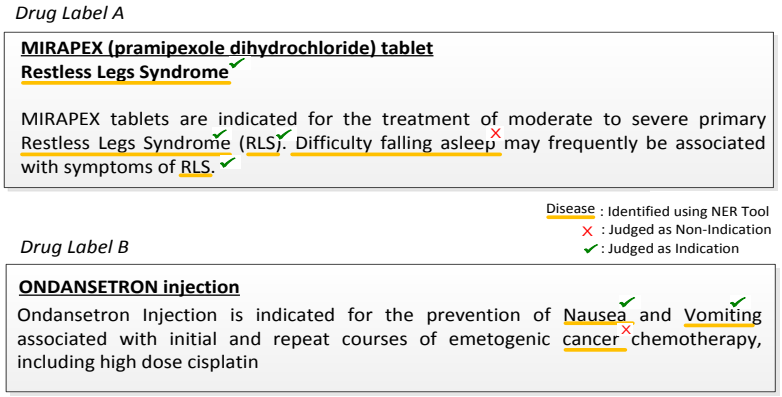


Figure 1. Illustration of Disease Named Entity Recognition (NER) and Classification Problems (Source: DailyMed Drug Labels)

Figure 1 also illustrates the two key steps in creating a drug indication repository from existing free-text resources: (i) *disease named entity recognition (NER)* to identify all the disease mentions (underlined) from drug narratives, and (ii) *disease classification* to distinguish indications from other disease mentions, demonstrated using checks for yes and 'X' for no. Disease NER, in general, is not a new problem; there exist high-performing and state-of-the-art tools

such as MetaMap[16], DNorm[17, 18], and KMCI[19] that automatically identify UMLS disease concepts from biomedical or clinical narratives. Also, many text-mining methods[20-25] have been proposed to extract disease mentions from drug resources such as DailyMed, MedlinePlus, and even Wikipedia.

The disease classification problem has not received much attention, however. Through detailed analysis of 100 FDA drug labels, we learned that even in the designated "INDICATIONS AND USAGE" section, about half of the disease mentions are not indications and include false positives such as characteristics of indications, contraindications, side effects, usages of another drug, unrelated diseases, etc[26]. There have been a few earlier attempts to address the disease classification problem in context of building a drug indication repository, e.g. SIDER-2 contains structured indications from FDA drug labels by filtering the disease mentions that overlap with its side effects repository[23], and Wei et al.[21] classify a given disease mention as an indication based on its frequency of occurrence across multiple structured and unstructured resources. However, the performances of these hand-crafted rule-based methods are limited either in terms of precision[23] or recall[21]. More recently, we resorted to manual expert annotations to identify indications from pre-recognized diseases in FDA drug labels, and curated a source-linked resource called *LabeledIn*[24, 26]. Upon systematic comparison with an existing automatically-curated resource[27], we concluded that human judgment is a reliable solution to the classification problem. However, it is time-consuming, and hence, expensive to be scalable with respect to building a repository with wide coverage of drugs.

In this study, we present an automatic end-to-end indication extraction framework that given an FDA drug label, (1) extracts the relevant section, (2) recognizes and normalizes all the disease mentions, (3) more importantly, classifies the disease mentions as "indications" vs. "non-indications," and (4) outputs the indication mentions and corresponding UMLS disease concepts that could be directly used to populate the target drug indication repository. The main contribution of this work is the use of supervised machine learning to address the under-explored disease classification problem in (3) and achieve a performance that is comparable to human experts.

Methods

The overall framework to extract drug indications from FDA drug labels is shown in Figure 2. As our data source, we use the FDA drug labels from the DailyMed website[12], which contains the most up-to-date drug labels submitted to FDA by drug manufacturers. The output of the framework is a set of structured and precisely normalized drug-disease relationships that can be used to automatically populate a computable drug indication repository. In the following subsections, we describe our disease NER and classification methods in further detail.

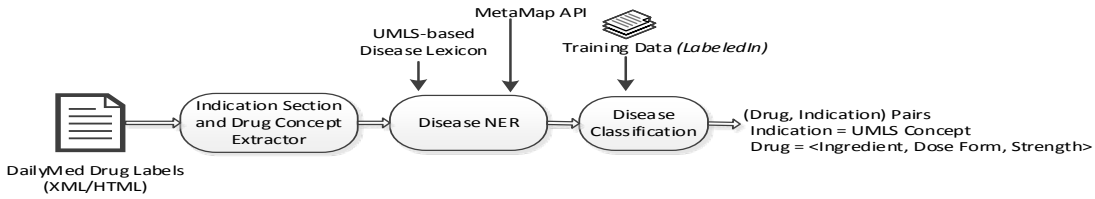


Figure 2. Overall Framework to Extract Drug Indications from FDA Drug Labels

Disease Named Entity Recognition

The goal of the method is to identify all the disease mentions, or indication candidates, from the textual descriptions of a given drug label. For this purpose, we prepared a disease lexicon using two seed ontologies, MeSH and SNOMED-CT, respectively useful for annotating scientific articles[17, 28, 29] and clinical documents[30-32]. The lexicon consists of 77,464 concepts taken from: (i) the disease branch in MeSH, and (ii) the 11 disorder semantic types (UMLS disorder semantic types excluding 'Finding') in the SNOMED-CT as recommended in a recent shared task [30]. As for the automatic tool, we applied MetaMap[16], a highly configurable program used for mapping biomedical texts to the UMLS identifying the mentions, offsets, and associated CUIs. We experimented with multiple settings of MetaMap, and the optimal setting for this study is illustrated in Figure 3.

The drug labels may contain overlapping disease mentions, e.g. the phrase "skin and soft tissue infections" denotes two specific diseases, "skin infections" and "soft tissue infections." While the final results by MetaMap do not return such overlapping mentions, these are captured in the intermediate results of MetaMap, known as the Metathesaurus candidates. Hence, we utilized these candidate concepts in our method. MetaMap provides two types of candidate

concepts: contiguous and dis-contiguous, e.g. in the phrase "skin and soft tissue infections", "soft tissue infections" is a contiguous candidate, and "skin+infections" is a dis-contiguous candidate. We found that MetaMap returns different sets of dis-contiguous candidates with and without the *term processing* feature. Hence, we conducted two runs of MetaMap for comprehensive results. Also, the *word sense disambiguation* feature was turned on to disambiguate mentions that may map to multiple CUIs, e.g. "depression." In order to restrict the returned candidates to specific semantic types from two vocabularies as mentioned above, we used a lookup against our custom disease lexicon as opposed to running multiple rounds of MetaMap for the two vocabularies. Finally, candidates with overlapping spans (e.g. "moderate to severe pain") were resolved in the following manner: (i) when both candidates were contiguous, the longer candidate was selected, (ii) when one candidate was dis-contiguous - (a) if the merged span contained conjunctions (e.g. "or," "and") or prepositions (e.g. "to"), then the merged span was pre-annotated and both CUIs were retained, e.g. the elliptical coordination in "skin and soft tissue infections," (b) if the two mentions were related by a parent-child UMLS relationship (e.g., the phrase "acute bacterial otitis media" maps to hierarchically related concepts "acute + otitis media" and "otitis"), then the longer mention was retained, else, the shorter mention was retained (e.g. the phrase "drug hypersensitivity reactions" maps to non-hierarchically related concepts "drug + reactions" and "hypersensitivity reactions").

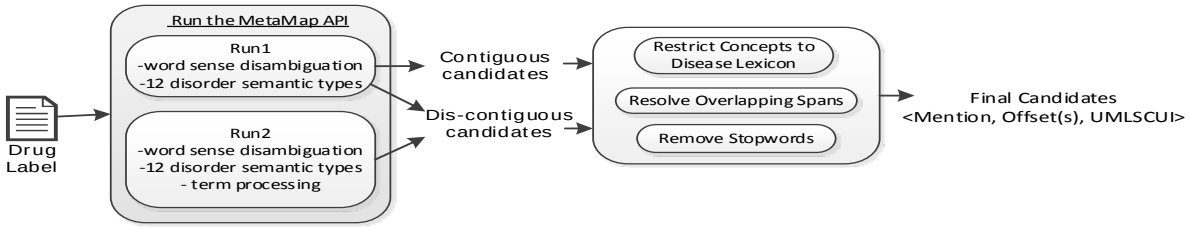


Figure 3. Disease Named Entity Recognition (NER) Method

Disease Classification

Next, we model indication extraction as a binary classification problem that judges whether a disease mention identified through NER method is an indicated usage of the drug. To build a high-quality gold standard, it is very important to judge and remove irrelevant disease mentions, e.g. "Difficulty falling asleep" from drug label A, and "cancer" from drug label B in Figure 1. This problem could also be perceived as another flavor of classifying diseases in clinical narratives into linguistic or temporal bins[33, 34].

Table 1. SVM Features for Disease Classification Problem

Feature (Description)	**Example1**		**Example2**	
	Data Point	**Feature Value(s)**	**Data Point**	**Feature Value**
Mention (Represented Surface-value)	Difficulty falling asleep (Drug Label A in Fig 1)	Difficulty falling asleep	Vomiting (Drug Label B in Fig 1)	Vomiting
Neighboring Tokens (left/right 5 tokens within sentence)		may, frequently, be, associated, with		the, prevention, of, -DISEASE-, and, associated, with, initial, and, repeat
Location (of sentence)		2		1
NDF-RT Match (the relationship of this disease concept with a concept catalogued in NDFRT[35] for this drug)		No Match		Exact Match
Semantic Category (of the UMLS concept corres.to mention)		Mental or Behavioral Dysfunction		Sign or Symptom

To address the classification problem, we used the support vector machines (SVM), a discriminative binary classifier known to be highly accurate and widely used in text classification. Given a set of data points (or examples) with known labels, the SVM assigns scores to each data point and finds a hyperplane that partitions the examples into two distinct bins, positive and negative based on the assigned score. For this problem, a data point is a disease mention at a given location, e.g. in Figure 1, the drug label A has five data points, and drug label B has three data points. We hand-picked the classification features based on our previous study with human experts, experimented with several features, and selected a set of linguistic, contextual, semantic, and dictionary-based features for this problem as described and exemplified in Table 1. Other features considered but not selected for the final classifier include the relative location of sentence in the drug label and the presence of drug name in the sentence containing the mention.

To train the SVM, we relied on a recently curated high-quality annotated text corpus, *LabeledIn*[24], that contains labeled or marketed indications for 250 frequently searched drug ingredients on PubMed Health. *LabeledIn* is unique among existing structured drug indication resources[21, 22, 27, 35, 36] in that it is human-validated, source-linked, and normalized to the most precise concepts in widely used UMLS vocabularies (SNOMEDCT, MeSH, and RxNorm). *LabeledIn* was created through double-annotation (88% kappa agreement) of 500 drug labels with assistance from a high-recall disease NER tool. In particular, for each drug label, the expert was presented with all disease mentions identified using an NER tool and was asked to assign a yes/no judgment. At the backend, the offsets of the disease mentions and the associated expert judgments were recorded. This helped in creating a fine-grained annotated text corpus that contained all disease mentions along with the associated human judgments. This becomes the training dataset for this problem as described in first row of Table 2. Since our goal is to automatically populate a computable gold standard of drug indications, we evaluate the results of the classifier at concept-level. For this purpose, we created an evaluation dataset that contains all the disease concepts (CUIs) present in a given drug label and the respective expert-determined yes/no assignments; this dataset is described in the second row of Table 2. There is a clear distinction between distribution of examples in both datasets because a mention identified as positive is likely to be repeated several times in a given drug label, e.g. the CUI represented by "RLS" has four occurrences in drug label A in Figure 1.

Table 2. Classification Data Points for 500 FDA Drug Labels

Data Point Definition	**#Data Points (+,-)**
Disease Mentions *with offset*	5,336 (70%, 30%)
Disease Concepts (UMLS CUIs)	3,013 (55%, 45%)

Results

We first measured the baseline performance on 500 drug labels from *LabeledIn* by assuming all diseases to be indications. The result is shown in the first row of Table 3. The result is a high recall and low precision. The recall is less than perfect because of the cases where the NER method identified a less-specific concept than annotated by the human experts, e.g., from the phrase "Acute Bacterial Otitis Media," the NER identified "Otitis Media" whereas annotators expanded the indication to the entire phrase.

Table 3. Indication Extraction Performance at Concept-level (micro-averaged)

Indication Extraction Method	**Precision(%)**	**Recall(%)**	**F1-measure(%)**
Disease NER *(baseline)*	54.99	93.93	69.37
Disease NER + NB	80.16	79.65	79.91
Disease NER + ME	85.02	**86.16**	85.59
Disease NER + SVM	**86.99**	85.58	**86.28**

We then conducted the experiment including a classification method in the framework. We experimented with three different classifiers, SVM, Maximum Entropy (ME), and Naïve Bayes (NB). We used an internal C++ implementation of the classifiers to conduct the experiments with 5,336 training examples and used 10-fold cross validation for evaluation against the evaluation dataset. To align the classifier results with our concept-level evaluation benchmark, we post-processed the results where multiple occurrences of the same concept were classified differently, e.g. SVM may classify the two occurrences of "RLS" in drug label A (Figure 1) as positive and negative, respectively. We

resolved this conflict by preferring the positive over negative classification. Finally, we selected SVM as it delivered the highest overall classification accuracy (optimized at score threshold 0.4) and hence leads to the highest indication extraction F1-measure on our data as shown in Table 3. The importance of each feature used for SVM is shown through the feature ablation experiment in Table 4 listing the features in decreasing order of their importance.

Table 4. Results of Feature Ablation Experiment (SVM)

Ablated (Removed) Feature	F1-measure(%)
Neighbors	79.75
Mention	83.09
Location	84.97
NDF-RT Match	85.11
Semantic Category	86.01

Using the results from the SVM-based extraction method, we randomly selected the 100 erroneous cases (50/220 false positives and 50/143 false negatives) for further analysis. Figure 4 shows the different categories of false positive errors that we observed. The two largest categories (shown separately) correspond to errors due to the NER component of the framework: either only partly identifies the correct concept (Boundary Error) or confuses with some other entities (Not a Disease Mention). About 36% of false positives represent boundary errors, e.g. from the phrase "carcinoma of the prostate," the disease tagger identified "carcinoma" and the classifier identified this mention as positive. This example is considered as a false positive as it does not match with our gold standard that contains the complete mention "carcinomas of the prostate." About 20% of false positives include non-disease mentions, such as part of an organization name (e.g. the "cancer" society), disease homonym ("seizure" used in a non-biological context), etc.

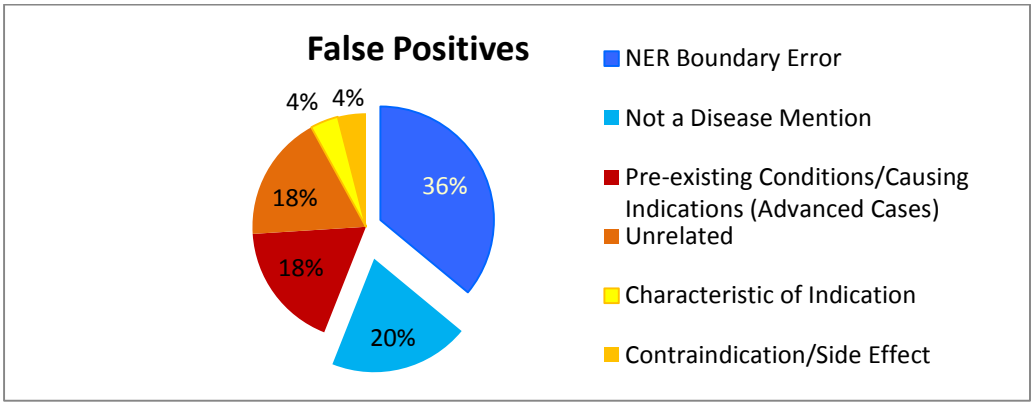


Figure 4. Analysis of 50 Randomly Selected False Positives

The remaining categories of false positive errors are due to indication misclassification as described next. About 18% include the diseases mentioned in some other context and unrelated to actual indications. Another 18% category represents the advanced cases, such as pre-existing conditions and conditions that cause indications. In our earlier study to create *LabeledIn* using expert annotations, we found some situations where experts needed to refer to additional references and drug properties to make their judgment. Consider the two drug labels in Figure 5: the "HIV infection" and "epilepsy" have similar feature values, e.g. neighbor tokens and semantic categories; however, "HIV infection" should be identified as a negative example and "epilepsy" as positive. Such cases are difficult to learn by a domain-independent classifier. Finally, the remaining few include characteristics of indications, contraindications, or side effects. In terms of false negatives, less than 5% were because of NER tool limitations, and the remaining represent misclassification errors because of poor structure of sentences (such as lists) and noisy neighbors (such as parenthesis), advanced cases, etc.

Since our final goal is to minimize human effort, we also studied how many concepts can be correctly classified by the SVM, and thereby eliminated from human judgment in a traditional machine-assisted annotation pipeline[26],

without incurring any errors. We investigated whether the scores (or ranks) assigned by SVM could be used to identify some perfectly classified examples. Figure 6 shows the variation of error rate with respect to the proportion of highest/lowest ranked concepts bypassed from further human validation; we found that we could safely eliminate 2.2% top-scored disease concepts and 5.1% bottom-scored disease concepts from further human validation.

Drug Label C

AZITHROMYCIN tablet, film coated

Azithromycin tablet, taken alone or in combination with rifabutin at its approved dose, is indicated for the prevention of disseminated Mycobacterium avium complex (MAC) disease in persons with advanced HIV infection.

Drug Label D

KEPPRA XR (levetiracetam) tablet, film coated, extended release

KEPPRA XR ™ is indicated as adjunctive therapy in the treatment of partial onset seizures in patients ≥16 years of age with epilepsy.

Figure 5. Example of Training Dataset showing an advanced case of classification ("HIV infection" and "epilepsy")

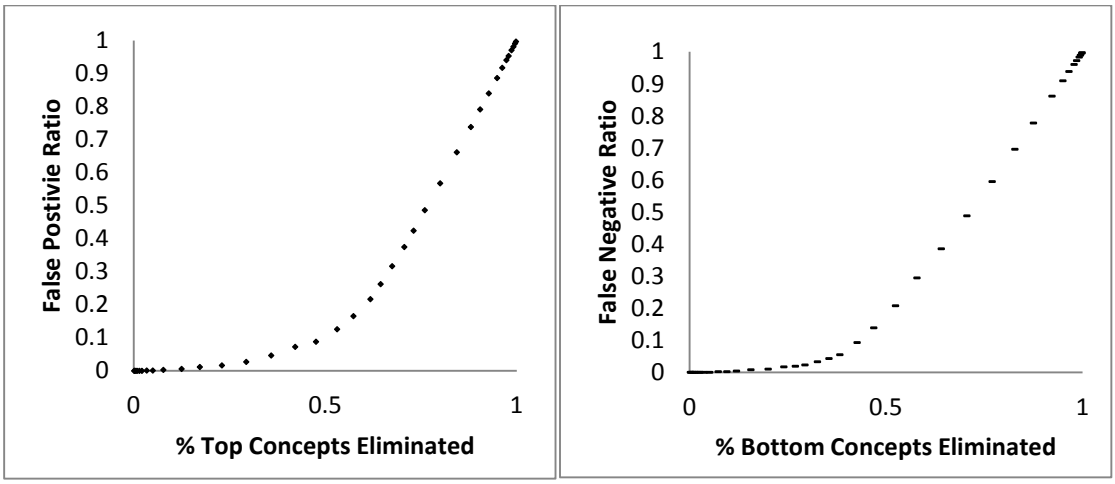


Figure 6. Effect of Eliminating Highest/Lowest Ranked Concepts on Error Rate

Conclusion

In this study, we have addressed the disease recognition and classification problems to prepare a comprehensive drug indication repository serving many critical applications. In the context of curation of a drug indication repository, there have been many important efforts primarily based on automatic recognition of disease mentions from drug narratives[20-24]. However, an important next step of classifying the automatically identified diseases as "indication"/"not indication" is largely missing from existing work, and hence the results from existing methods suffer from lower precision requiring further validation and cleaning by human experts. Taking a step forward, this study presents a framework that not only recognizes disease mentions from drug narratives, but further distinguishes indications from other disease mentions using an SVM-based classification method, thus minimizing the need to have humans in the pipeline. We experimented with 500 FDA drug labels corresponding to frequently-sought human drugs, and configured a disease NER tool to deliver an optimal recall on drug labels. We find that the combination (recognition + classification) framework achieves 32% improvement in precision and 17% improvement in F1-measure over a baseline method that is based only on recognition.

The proposed framework delivered an 86% F1-measure on 500 drug labels whereas two human experts in a previous study using similar NER mechanism[26] delivered 88% joint F1-measure on 100 drug labels after their first round of annotation. We consider these performances to be equivalent to each other, given the upper limit (94% recall) on classification performance due to natural language limitations of the NER tool. We also observed that over half of the

errors are due to the NER limitations. This clearly indicates that the proposed classifier can act as an independent or complementary annotator, saving annotation costs and time[26]. Furthermore, we find that about 7% of the training examples fed to the classifier could be altogether bypassed from further human judgments without incurring any errors. In terms of improvement, we found that several erroneous cases required advanced domain knowledge and are difficult to learn. Also, many false negative errors could be resolved by refining the neighbor calculation algorithm and adding formatting features to the classifier. Our future work also includes improving the technical aspects of the classifier such as automatic feature generation and handling imbalanced data. Lastly, we observed that certain annotated drug indications are specific to certain procedures/conditions (e.g. "nausea" and "vomiting" are associated with "cancer chemotherapy" in Figure 1 drug label B). Such information is not yet captured through NER or classification modules of our current pipeline, and requires further processing of drug labels.

Acknowledgements

This research was supported by the Intramural Research Program of the NIH - National Library of Medicine. The authors would like to thank Robert Leaman for proofreading the manuscript.

References

1. Islamaj Dogan R, Murray GC, Neveol A, Lu Z. Understanding PubMed user search behavior through log analysis. Database : the journal of biological databases and curation. 2009;2009:bap018. Epub 2010/02/17.
2. Neveol A, Islamaj Dogan R, Lu Z. Semi-automatic semantic annotation of PubMed queries: a study on quality, efficiency, satisfaction. Journal of biomedical informatics. 2011;44(2):310-8. Epub 2010/11/26.
3. Ely JW, Osheroff JA, Gorman PN, Ebell MH, Chambliss ML, Pifer EA, et al. A taxonomy of generic clinical questions: classification study. BMJ. 2000;321(7258):429-32. Epub 2000/08/11.
4. Lu Z, Agarwal P, Butte AJ. Computational Drug Repositioning - Session Introduction. Pacific Symposium on Biocomputing 2013. p. 1-4.
5. Li J, Lu Z. A New Method for Computational Drug Repositioning Using Drug Pairwise Similarity. IEEE International Conference on Bioinformatics and Biomedicine2012. p. 1-4.
6. Li J, Lu Z. Pathway-based drug repositioning using causal inference. BMC bioinformatics. 2013;14((Suppl 16):S3).
7. Chang RL, Xie L, Bourne PE, Palsson BO. Drug off-target effects predicted using structural analysis in the context of a metabolic network model. PLoS computational biology. 2010;6(9):e1000938. Epub 2010/10/20.
8. Khare R, An Y, Wolf S, Nyirjesy P, Liu L, Chou E. Understanding the EMR error control practices among gynecologic physicians. iConference 2013 Fort Worth, TX, 2013. p. 289-301.
9. Lesar TS. Prescribing errors involving medication dosage forms. Journal of general internal medicine. 2002;17(8):579-87. Epub 2002/09/06.
10. Ling Y, An Y, Liu M, Hu X. An error detecting and tagging framework for reducing data entry errors in electronic medical records (EMR) system. IEEE International Conference on Bioinformatics and Biomedicine (BIBM); Shanghai, China, 2013. p. 249-54.
11. Li J, Khare R, Lu Z. Improving Online Access to Drug-Related Information. IEEE Second International Conference on Healthcare Informatics, Imaging and Systems Biology; La Jolla, CA, 2012.
12. DailyMed: Current Medication Information. Available from: http://dailymed.nlm.nih.gov.
13. DrugBank: Open Data Drug and Drug Target Database. Available from: http://www.drugbank.ca/.
14. MedlinePlus: Trusted Health Information for You. Available from: http://www.nlm.nih.gov/medlineplus/.
15. MedicineNet: We Bring Doctor's Knowledge to You. Available from: http://www.medicinenet.com/script/main/hp.asp.
16. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proceedings / AMIA Annual Symposium AMIA Symposium. 2001:17-21. Epub 2002/02/05.
17. Leaman R, Islamaj Dogan R, Lu Z. DNorm: disease name normalization with pairwise learning to rank. Bioinformatics. 2013. Epub 2013/08/24.
18. Leaman R, Khare R, Lu Z. Automatic Disease Normalization in Clinical Notes with DNorm. Journal of the American Medical Informatics Association (JAMIA). Under Review.
19. Lab VBLP. KMCI - KnowledgeMap Concept Indexer. Available from: http://knowledgemap.mc.vanderbilt.edu/research/content/kmci-knowledgemap-concept-indexer.
20. Neveol A, Lu Z. Automatic Integration of Drug Indications from Multiple Health Resources. ACM International Health Informatics Symposium; Arlington, VA, 2010. p. 666-73.

21. Wei WQ, Cronin RM, Xu H, Lasko TA, Bastarache L, Denny JC. Development and evaluation of an ensemble resource linking medications to their indications. Journal of the American Medical Informatics Association : JAMIA. 2013;20(5):954-61. Epub 2013/04/12.
22. Fung KW, Jao CS, Demner-Fushman D. Extracting drug indication information from structured product labels using natural language processing. Journal of the American Medical Informatics Association : JAMIA. 2013;20(3):482-8. Epub 2013/03/12.
23. A side effect resource to capture phenotypic effects of drugs [database on the Internet]. 2010.
24. Khare R, Li J, Lu Z. LabeledIn: Cataloging Labeled Indications for Human Drugs. Journal of biomedical informatics. 2014.
25. Li Q, Deleger L, Lingren T, Zhai H, Kaiser M, Stoutenborough L, et al. Mining FDA drug labels for medical conditions. BMC medical informatics and decision making. 2013;13:53. Epub 2013/04/27.
26. Khare R, Li J, Lu Z. Toward Creating a Gold Standard of Drug Indications from FDA Drug Labels. IEEE International Conference on Health Informatics; September 09-11, 2013; Philadelphia, PA, 2013.
27. SIDER 2 Side Effect Resource. Available from: http://sideeffects.embl.de/.
28. Dogan RI, Lu Z. An improved corpus of disease mentions in PubMed citations. Workshop on Biomedical Natural Language Processing; Bethesda, MD, 2012. p. 91-9.
29. Huang M, Neveol A, Lu Z. Recommending MeSH terms for annotating biomedical articles. Journal of the American Medical Informatics Association : JAMIA. 2011;18(5):660-7. Epub 2011/05/27.
30. Leaman R, Khare R, Lu Z. NCBI at 2013 ShARe/CLEF eHealth Shared Task: Disorder Normalization in Clinical Notes with DNorm. Conference and Labs of the Evaluation Forum 2013 Working Notes. 2013.
31. Khare R, An Y, Li J, Song I-Y, Hu X. Exploiting semantic structure for mapping user-specified form terms to SNOMED CT concepts. ACM SIGHIT International Health Informatics Symposium; Miami, FL, 2012. p. 285-94.
32. An Y, Khare R, Hu X, Song I-Y. Bridging encounter forms and electronic medical record databases: Annotation, mapping, and integration. IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2012); October 04-07 2012; Philadelphia, PA, 2012.
33. Raghavan P, Fosler-Lussier E, Lai AM. Temporal Classification of Medical Events. . Workshop on Biomedical Natural Language Processing (BioNLP); Montreal, Quebec, Canada, 2012.
34. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal of biomedical informatics. 2001;34(5):301-10. Epub 2002/07/19.
35. 2012AA National Drug File - Reference Terminology Source Information. Available from: http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/NDFRT/.
36. Freebase: A Community-curated Database of well-known People, Places, and Things. Available from: http://www.freebase.com/.

Applications of Health Information Exchange Information to Public Health Practice

Patrick Kierkegaard, PHD[1] , Rainu Kaushal, MD, MPH[2,3,4,5,6], and Joshua R Vest, PhD, MPH[2,3]

[1]Department of Computer Science, University of Copenhagen, Denmark; [2]Center for Healthcare Informatics & Policy, Weill Cornell Medical College, New York NY; [3]Department of Healthcare Policy and Research, Weill Cornell Medical College, New York NY; [4]Department of Medicine, Weill Cornell Medical College, New York, NY; [5]Department of Pediatrics, Weill Cornell Medical College, New York, NY; [6]NewYork-Presbyterian Hospital, New York, NY

ABSTRACT

Increased information availability, timeliness, and comprehensiveness through health information exchange (HIE) can support public health practice. The potential benefits to disease monitoring, disaster response, and other public health activities served as an important justification for the US' investments in HIE. After several years of HIE implementation and funding, we sought to determine if any of the anticipated benefits of exchange participation were accruing to state and local public health practitioners participating in five different exchanges. Using qualitative interviews and template analyses, we identified public health efforts and activities that were improved by participation in HIE. HIE supported public health activities consistent with expectations in the literature. However, no single department realized all the potential benefits of HIE identified. These findings suggest ways to improve HIE usage in public health.

INTRODUCTION

Public and population health benefits are an important justification for the US' significant investment in health information exchange (HIE) and associated interoperable health information technologies.[1 2] During the past decade, the federal government has invested billions to support electronic health record adoption through the Meaningful Use program, which includes requirements for the capabilities to exchange patient information. In addition, the Office of the National Coordinator for Health Information Technology (ONC) has overseen the $547 million State Health Information Exchange Cooperative Agreement Program to establish information exchange activity at a state level.[3] Other federal agencies and individual states have also invested heavily in HIE infrastructures.[4-6] These investments were made with such assumptions that information exchange "improves public health activities and facilitates the early identification and rapid response to public health threats and emergencies,"[7] or that ensuring public health "will depend on the implementation of information technology systems."[8]

HIE, the process of electronically sharing patient information between different organizations, stands to benefit public health.[9] Public health agencies are data driven organizations and HIE increases access to previously difficult to obtain clinical and demographic data. Also, by drawing on data created and stored by multiple healthcare entities, HIE can create longitudinal descriptions of individual patient care and illness. With broad provider participation, HIE data can reflect the healthcare experiences and utilizations of whole communities and populations. This is an advantage over looking at data from a single institution or healthcare system, which is by definition a restricted and limited sample. As a result of these features, HIE is expected to benefit several areas of public health practice from disease surveillance, to disaster response, to healthcare service delivery.[10]

However, public health participation in HIE efforts has historically not been widespread.[11] While some notable examples of public health usages of HIE exist, the realization of the benefits of HIE remain largely undocumented.[12] We sought to determine if any of the anticipated benefits of exchange participation were accruing to public health practitioners. Because no extant survey or tracking tool examines this question and the fact that the potential benefits to public health are so varied, we undertook a qualitative study of public health practitioners' experiences with HIE.

Identifying any successful applications of HIE to benefit public health practice, may help spur wider adoption to support the assessment and assurance of the population's health.

METHODS

We interviewed public health professionals about their usage of HIE and experiences working with exchange facilitating organizations. All of the public health professionals worked at either a state or local health department that was an active participant in an HIE effort.

Sites & Participants

First, we approached active regional health information organizations (RHIOs) and state level exchange entities with public health members to participate in this study. This excluded all RHIO/HIE efforts still in planning phases or health departments without any active users of HIE systems. Prior and ongoing research projects, as well as contacts with other HIE researchers helped to identify the potential participants. We secured cooperation from three local HIE efforts: the Rochester Regional Health Information Organization (Rochester, NY), Southern Tier HealthLink (Binghamton, NY), and the Integrated Care Collaborative of Central Texas (Austin, TX). Two statewide exchanges agreed to participate as well: the Indiana Health Information Exchange and the New Mexico Health Information Collaborative. Each of these exchanges has been in existence for 7+ years and their participation, technology, usage and effectiveness have been previously described.[13-17]

Each HIE effort identified points of contact (such as an HIE committee member) or registered active users at their member health departments for participation. After each interview, we asked informants to recommend other potential interviewees who could provide additional insight about their site's experiences with HIE activities.[18] This approach ensures that the different point-of-views would be considered. We interviewed a total of 22 representatives from 5 local health departments and 2 state health agencies screening for either direct users of HIE systems or the ultimate consumers of HIE obtained information. The interviewees had the various job titles of: epidemiologist, medical director, public health nurse, disease surveillance, case manager, and director of information technology. We obtained individual consent from each participant. To protect confidentiality, we do not report the numbers of interviews or job titles by organization.

Data collection

For each exchange effort, we reviewed instruction manuals, recruitment materials, technical documents, published reports, and websites, or obtained an overview of the exchange effort and a demonstration of their HIE system's interface, data sources, and functionalities. In this way, we were familiarized with the capabilities and objectives of each effort prior to data collection.

Data were collected through a combination of in-person and telephone interviews (PK & JV). Interviews were conducted using a semi-structured protocol developed as part of a larger evaluation of HIE in the State of New York. The guide covered the areas of job responsibilities, HIE usage, workflow fit, and technology perceptions. Consistent with the semi-structured approach, we adapted question wording and probes to match the interviewee's work role, location, and HIE experiences. The interviews were timed to last approximately 30 minutes. All interviews were audio-recorded and transcribed. All data collection began in May 2013 and ended in January 2014. We ended data collection when we had reached thematic saturation[19] and had interviewed respondents from both the state and local agency levels.

Analysis

Because we were interested in a specific set of HIE benefits suggested by researchers, practitioners, and policy makers, we opted to analyze the transcripts using template analysis.[20, 21] This approach begins with an a priori identified set of codes, as opposed to deriving codes from the data themselves like in grounded theory. We selected a template analysis as the best match to our objective of determining if the anticipated benefits of exchange participation were accruing to public health practitioners.

To identify the codes for our template analysis, we reviewed the literature for the benefits of HIE to public health departments. Our search of the phrases "health information exchange" AND "public health" in PubMed yielded 85 citations. After reviewing the titles and abstracts, we selected 16 for full text review. A review of the references of these papers resulted in 2 additional articles for a total of 18 papers. Of these, 10 described realized, potential, or anticipated benefits of HIE to public health practitioners.[10, 22-30] We were focused on the application of HIE to public

health practice, and as such we did not include the potential benefits that could be accrued to the HIE efforts or private providers from data exchange with public health agencies.

Independently, we (PK & JV) abstracted 26 effects from the included articles (kappa = 0.74). We refined this raw list in two ways. First, we divided the identified effects into the activity benefited (e.g. "clinical care") from the reason for the benefit (e.g. "increased access to data"). Next, we merged related and synonymous concepts into single codes (Table 1). We independently coded our transcripts using this template (coding by PK and JV reviewed with RK). Template analysis allows for the identification of new or expected findings within the data, so we were attentive to new themes. Emergent themes were developed independently during our closed readings. To reconcile differences, we combined our independently derived lists of emergent themes, collapsing synonymous terms, and supplying a single definition. We independently recoded our transcripts with the single emergent theme list. We undertook joint readings to resolve differences between our coding schemes.

Table 1. Codes and definitions of health information benefits to public health practice for template analysis.

Category	Code	Definition
Activity benefited		**The service, action, effort, or activity of public health practice that is improved by health information exchange (i.e. the "what")**
	Assessment & planning	All activities of a strategic nature oriented towards community or organizational planning (e.g. quality measurement or community assessments)[10, 22, 30]
	Case management & care coordination	Activities surrounding managing the clinical care and multiple needs of individual patients[26, 27]
	Clinical care	Activities related to patient care or clinical service delivery[10, 26, 27, 30]
	Communication (public)	Information from the public health agency to providers or the public (including alerts about events)[10, 22, 28]
	Preparedness	All emergency or disaster preparation and response efforts[10]
	Surveillance	Collection, monitoring, and investigation of conditions, status, or events as required by law, including syndromic surveillance and outbreak detection[10, 22, 24-30]
	Workplace efficiency	Improved efficiency for public health professionals (i.e. productivity / effort as a function of real or opportunity costs)[22, 26, 27]
Reason for the benefit		**The feature, characteristic, or process by which health information exchange generates benefit (i.e. the "how")**
	Information completeness & quality	More complete and accurate information [22-29]
	Information availability	Access to new or more types of data (including clinical information)[10, 22, 24, 25, 27, 29, 30]
	More timely information	Improvements in the temporal aspects of information (e.g. how fast information can be obtained, how it is available sooner, or how it can be used faster)[24, 25, 27, 29, 30]

Consistent with our primary objective, we reported if, and how, HIE supported each of the identified activities in our template. We then summarized the salient emergent issues. To help ensure respondent confidentiality, we avoided reporting both the location and job title of the interviewee. Instead, we focus on reporting the state in which the interviewee worked and only reported the area of public health practice if that information was relevant to the quote and if we had interviewed multiple individuals in that state with the same job.

RESULTS

According to NVivo's guidelines, we achieved a good level of agreement in our initial independent coding.[31] The kappa score for the "activity benefited" category was 0.71 and 0.82 for the "reason for benefit" category. The overall agreement was 0.75. Differences were resolved by comparing coding and joint discussion.

We identified instances or examples of all 10 anticipated benefits of HIE to public health. All sites reported at least four benefits from HIE (max was 7). All public health agencies realized benefits in the areas of surveillance and workplace efficiency. While benefits were seen across many categories, the level of activity within each was not particularly deep (i.e. few users, integrated into few programs, or applied to a small number of projects).

Assessment & planning

Interviews revealed that some public health departments were using HIE systems to develop a better understanding of their communities. As expected, the assessment activities included traditional public health target areas like Indiana's interest in infant mortality. Likewise, departments in Texas were using HIE data to complete community health assessments and strategic plans and also to understand health disparities. However, public health professionals in Texas were also interested in healthcare utilization patterns and preventable emergency department visits, which are not traditional public health areas. HIE data also supported the practical activity of grant writing.

HIE systems supported these activities by increasing information availability and access to timely data. As one interviewee from Texas noted, "*We're able to get now what we really have not had any way to look at in the past…Because we've only guessed before in terms of how people – especially without health insurance – are using the healthcare system to try to get their primary care needs met…This has given us the ability to actually see that in a way that helps us change our strategies of intervention*." In Indiana, it was the timeliness of HIE data that was particularly useful, "*…98 percent of the decisions we're making now here at the state department of health are truly data driven…*"

Case management & care coordination

Access to HIE systems supported case management activities and care coordination by increasing data availability and the timeliness of information. For example, a nurse in New York noted how the local HIE system enabled her to monitor the health of tuberculosis (TB) patients: "*If I'm treating a case or a prevent case, it's 9 to 24 months, okay? So, they're under our umbrella for a long time. They could be in the hospital from everything from a major cardiac event to they needed stitches in their hand. That all pertains to us*." The nurse further used the HIE to follow more detailed clinical information like current medications. Also in New York, a cancer screening program used the system to receive imaging reports for enrolled clients. Likewise, case managers in Texas were able access the recent emergency department encounters of diabetics.

Clinical care

Only the health departments in New York provided direct clinical care to patients. Both New York sites had active HIE users in their TB clinics. One nurse noted how she relied on HIE to obtain imaging results, reports, and laboratory values. "*That aids me. Immediately, I can see whether the results are normal or not, and if they're abnormal, we need to act quickly in determining if, if it's abnormal due to active TB disease…It's something we can act on right away when we get that information*."

Communication

In all of the other potential benefits areas, the value to public health is in the movement of information to the public health agencies. Alternatively, information and data could flow the other direction: from public health to the community or to providers. This was really only occurring in Indiana, and even then, that was narrow use case. Provider access to the immunization registry data was via local exchanges (i.e. RHIOs). Therefore, providers could use HIE to populate their own local information systems with the state's immunization data. However, this is more like bi-directional data sharing than communication as defined in the literature.

Preparedness

Two departments reported using HIE for preparedness activities, broadly defined to include both preparation and response efforts. New Mexico reported using HIE to help monitor for health complications associated with special events like Native American Tribal gatherings within the state. In New York, however, one public health department had actually put HIE to use during a natural disaster. "*When we had our flood here, 25,000 people were evacuated,*

and they didn't have meds. They didn't know who their physicians were, where they were gonna get their meds. The [HIE] came in and set up their computers and gave access to physicians from both hospitals …so it enabled them to be continued to do primary care…We sheltered people for 99 days."

Surveillance

As would be expected of systems that compile healthcare utilization, patient characteristics, and diagnoses, HIE systems were widely used for surveillance activities. Because they are a continuum of activities and tasks, this code included both those dealing with the detection of public health events and the steps public health professional undertake to respond to cases of disease. In terms of detection, HIE helped in two ways. First, HIE systems increased provider reporting. A local health department in Indiana "*we're doing a better job of capturing [reportable conditions]*," because electronic laboratory were reported through the HIE. Another professional in Indiana echoed this sentiment: "*We find a lot of cases that were not being reported to us before. Probably the largest number would be Hepatitis C cases, but we see other cases too that we don't get the reports from physicians or sometimes they weren't being reported by the labs to us*." The second application of HIE to the process of disease detection was in syndromic surveillance efforts. Both state health agencies interviewed, Indiana and New Mexico, reported using HIE for syndromic surveillance.

We also considered public health investigation, the activities public health practitioners undertake in response to identified cases or events, as part of the broader surveillance code. HIE was beneficial to these activities as well. For example, surveillance staff in Texas reported using HIE systems to collect demographic information that was not included on laboratory reports or to access historical test results to determine if cases were chronic or acute. Surveillance staff in New York also used HIE for similar information, but also noted the importance of being able to collect basic information like patient addresses, phone numbers, or dates (and locations) of hospitalizations. In these instances, the HIE systems served to provide access to information that was not available in current public health information systems, information which providers or laboratories tended not to report, or information that took more time and effort to collect (e.g. telephone calls to providers or reviews of charts).

Workplace efficiency

While respondents generally thought systems could be improved or further developed, access to HIE resulted in numerous increases in workplace efficiency. These efficiencies cut across different public health activities and, generally, the improvement was in time saved. For example, one respondent working in a cancer program in New York noted: "*This is just a lifesaver because it eases the burden on [providers] and eases the burden on us...We're not waiting and waiting and waiting and calling them five times and annoying them. We just go [to the HIE] and we get the information we want…It just streamlines the process*." HIE saved time for surveillance staff by helping them locate case information like demographics or prior test results and by automating the delivery of disease and laboratory reports. As one respondent noted, her job was easier because, "*instead of calling every hospital, I pull up in the [HIE]*." Respondents various noted things like avoiding "*delays*," being able to "*act faster*," not having to send "*someone out in the field*," not "*having to wait*," and being "*quicker*."

Respondents from New Mexico also reported operational efficiency in terms of data management. Without the services of the HIE, the state health agency would have to create individual data feeds with each hospital in the state. That approach "*is cumbersome and you have to then work with every single different system and figure out formats and all that other stuff*." In New Mexico electronic laboratory reports and emergency department data are routed through the state's HIE. In addition, a public health professional from Indiana also noted that the HIE was able to provide services and technologies that his agency was not capable to provide in-house.

Emergent Themes

During the coding phase, several themes emerged (Table 2) that primarily consisted of the concerns public health professionals faced in terms of HIE usage and the requirements necessary for the technology to reach its full potential in terms of usefulness. We grouped emerging themes into the three categories: non-technical factors, technology-related deficiencies, and requested features.

Emergent theme: Non-technical factors

Most of the public health professionals we interviewed were end users with little or no influence over the design and functionality of the HIE systems. The emergent theme of non-technical factors represented all the influences beyond the sphere of the users' interaction with the HIE that led to implementation and integration challenges. Some were unique to a particular area of public health practice, but others were common across our study sample. Financial

limitations were one such cross-cutting challenge. One public health professional summarized his state's reality succinctly: *"Our state funding is very limited"* and *"we can't spend money on infrastructure development."* Respondents from other states noted that limited financial support impeded HIE development: "*[There are] not a lot of funding opportunities for areas to begin implementing an HIE system,"* and that organizations do not have *"the financial wherewithal or the technological wherewithal right now"* to provide support for such an initiative.

Variations in the level of data comprehensiveness were evident in all states. While HIE did increase the amount of data available, universally, public health professionals wanted more. They desired more data types and more sources of data contributing to the exchange effort. However, it was non-participating providers that were seen as the primary drivers of data comprehensiveness. While those engaged in patient care noted that non-participating providers created gaps in their clinical pictures, the non-participants were particularly troublesome for disease surveillance staff. Surveillance staff and epidemiologists talked about the need to get "*the whole population,*" "*a representative sample*," "*a statewide perspective*," or the "*denominator*." Without the participation of all providers in their community, they felt the data were limited, or at least highly qualified, for surveillance purposes. Additionally, data comprehensiveness had real implications for usage. One nurse said that she was unable to find any data on a suspected case *"probably 98 percent"* times she accessed the HIE system and another professional's concerns meant that there was hesitancy to use the system: *"I have found the majority of times when I go to [HIE system] that I can't find the patient. So I don't use it very much at all."*

During interviews we noted that data access policies were variable between states and could lead to challenges. Data appeared to be most accessible in New York. A nurse reported that she was simply able to mark HIE queries as "*No Consent Needed, because it's public health needed*". However, in Texas, access of non-consented patients for public health purposes was avoided, because it "*sends out the sirens.*" Another example of a policy-generated limitation occurred in Indiana. According to interviewees, "*(S)tate law basically says, 'You can see these results when it's a reportable positive test.' And so that's become an issue is that we don't see the negative tests in the database. And so, you know, it's hard to calculate incidence or prevalence off of that when you don't see the negative.*"

Emergent theme: Technology-related factors

Technology-related factors represented the technical-orientated challenges associated with the HIE. As an example, data standards and coding reportedly created challenges for one state: "*(O)ur big reference laboratories have their own local codes, which then we have to translate into something that's meaningful, and that takes a lot of time."* A more widespread challenge was the fact that HIE systems were generally not interoperable with any other systems in use at the health departments. The HIE systems might facilitate the timely access of information, but public health practitioners still had to print out patient records and reports from their respective HIE system and hand enter the information into various public health systems. Non-integration also appeared in terms of multiple methods of exchange. A respondent from Indiana noted that not only were there five local exchanges within the state, which prevented them from getting "*the full picture,*" they also had providers "*…that want to establish direct connections with us because they don't want to go through an HIE."*

Emergent theme: Requested features

Our study also revealed that respondents shared similarities in terms of how they would like their HIE system to improve. As expected, these requests were in direct response to the limitations or challenges above or were general comments about system functionality problems. Specifically, when asked what would make HIE better support their daily work requirements, in most cases, interviewees desired more data. Access to physician notes were among the prevalent feature requests as respondents generally agreed that it would *"be nice if the doctor's office notes were in there"* and that having access to them *"would be infinitely helpful."* Similarly, interviewees also expressed an awareness of the legal basis for public health activities and wanted HIE systems to be in line with those requirements. Respondents in Indiana and New York tied their wants for data to compliance with state reporting requirements. For example, in Indiana interviewee specifically wanted the HIE to include reports on other notifiable conditions to support the "*cancer registry for TB, HIV, and some of the other state mandated reporting."*

Within the feature request theme, we also identified different levels of interest in how HIE data could be directly queried by the health departments. In our sample, it was common that any direct analysis or custom reporting queries were often handled by the HIE facilitating organization and not the health departments. The state health agency in New Mexico reported the ability to directly query HIE systems on their own, but local health departments reportedly did not have direct access to the underlying raw data. However, local epidemiologists expressed an interest in being able to do their own queries. In Indiana, one respondent stated, *"If it's something that we could do*

ourselves and not have to rely on them again, it would be just that much faster." A similar sentiment was expressed in Texas: *"epidemiology staff would love to be able to do some things themselves."* Likewise a New York surveillance staff expressed her desire as: "*...let us be public health and figure out what to do with [the data]*."

Table 2. Emergent themes and definitions.

Category	Code	Definition
Non-technical factors		**Challenges caused by factors not related to the direct user-interaction with the health information exchange technologies**
	Financial	The ability or inability of public health or healthcare organizations to fund health information exchange
	Data comprehensiveness	Limitations on 1) data types needed for day-to-day activities (e.g. physician notes, radiology data, community care charts, physicians phone number, etc), and 2) sources of data (e.g. more providers) that would increase representativeness or utility
	Policy complications	The use of health information exchange for public health practice was incompatible or was challenged by organizational or state policies.
Technology-related factors		**Challenges directly related to the limitations of the technology**
	Non-integrated systems	Systems are unable to partially or fully exchange data, resulting in other of data exchange and processing such as paper-based artifacts and end-to-end technology.
	Standards and coding	Issues in the presentation, management, and transmission of data elements and meaning.
	System functionality	Users experience trouble with the technology providing up-to-date data due to system down-time, glitches, or data load errors.
Requested features		
	Query privileges	Epidemiologist's ability to conduct queries themselves and reduce dependence on external parties to provide reports.
	Inclusiveness of state mandated reporting	Facilitation of data sharing in compliance with state mandating.
	Management system modules	Systems to aid in organizational workflow and tracking of cases and outbreaks.

DISCUSSION

Several public health activities can be strengthened by participation in HIE. In this multi-site, multi-state qualitative sample of public health professions, we documented benefits consistent with expectations in the literature. The above results indicate a positive impact of HIE on public health. Nevertheless, that impact is qualified in several key ways and HIE usage in public health is still open to improvement.

First, while each public health department reported benefits from HIE participation, no department was leveraging HIE in all seven of the activities identified in the literature. This lack of comprehensive application of HIE to public health is partially explained by the different capabilities and maturity levels of each HIE effort, the inclusion of both local and state public health agencies, and the different service offerings between the public health agencies.

However, it is also the case that usage within each agency was not widespread. Each public health agency tended to focus on a singular or narrow set of use cases. In our interviewed sites, usage tended to be focused in epidemiology / surveillance units or in clinical care, with less usage in other areas. Again, this may be a product of developing systems, but several of the HIE efforts included in this sample have long histories.

Additionally, many of the documented gains could realistically be replicated on greater scales. For example, HIE did support more comprehensive surveillance data and clearly saved public health employees a lot of time and effort. Efficiency gains are not to be dismissed and public health is always in need of better data, however, these are both intermediate steps to the ultimate goal of better population health. Likewise, public health agencies were using HIE to support community health assessments and planning, but descriptive reports and aggregated statistics are not the same as advanced analytics, modeling, or evaluation. Also, as indicated by the interviewees, not every public health agency with HIE attempted this type of aggregated analysis. Likewise, interviews indicated that little evidence for using HIE to improve communication existed; one state reported sharing data back with providers. However, examples of using shared information systems to alert providers of important public health events exist.[32] These observations are not meant to diminish any of the current accomplishments of public health agencies, but taken together with the limited usage of HIE by public health, clearly more benefits can be realized.

The types of data needed by interviewees indicated one easy fashion HIE efforts can demonstrate their value to public health. A common use of HIE by public health was to essentially "fill in" missing data, but this basic activity supported surveillance efforts and helped workflow efficiencies. Simply enabling public health practictioner access to HIE systems would result in at least these moderate wins. The information available in the HIE does not need to be overly complex, because interviewees indicated that something as simple as demographics or contact information is important to public health. Without contact information, disease control is severely hampered; without demographics, public health agencies cannot assess and attempt to remedy disparities in health. Unfortunately, these are exactly the types of data healthcare information systems have not historically collected well.[33] By pooling data from multiple sources HIE can improve data completeness[23] and be of immediate utility to public health.

The emergent themes in our data documented several areas for improvement. One clear need is better systems integration. The lack of integration and sharing of data between systems is a common problem in public health practice[34] and one that leads to many inefficiencies.[35] In too many instances, public health professionals were accessing an information system that did not directly share data with any of their other data repositories. While the HIE may have useful information, that information was printed, copied, or just remembered for entry into other systems. Also, a dominant need raised by the interviewees themselves was the desire for each exchange effort to be more comprehensive in terms of information and information sources. Public health professionals knew the difference between an information source that was a sample and that which was truly representative of the population. Consistently, the utility of HIE would be higher with more participants on board.

The emergent themes also raise some interesting questions for future research and policy discussions. For one, these findings leave unanswered the optimal approach to HIE analytics. Should health departments advocate for a larger role in analytics through direct access to HIE data as some interviewee suggested? On one hand, state and large local health departments typically employ epidemiologists and biostatisticians with sufficient skills to analyze large datasets. However, technology infrastructures in public health tend to lag behind healthcare organizations and during periods of financial constraints departments may not be able to allocate staff and financial resources necessary to manage and analyze very complex clinical datasets. From the other perspective, HIEs may see being able to offer analytic services themselves as an important revenue stream.[36] Also, as organizations that primarily serve the healthcare system, privacy concerns are paramount for RHIOs and HIEs. This could create natural hesitancies around increasing direct data access. Another question raised by these findings revolves around the consent policies for public health practices. We noted differences in the policies followed in two different states: one allowed for easier public health access and the other appeared to restrict information retrieval to only consented patients. Public health's powers to investigate tend to be broad and allow for the collection of personal health information without written patient authorization.[37] Understanding how individual public health agencies and states view access rules could be important to understanding the actual effectiveness of HIE in supporting public health.

Limitations

These findings are subject to limitations. The activities supported by HIE and the interviewees' experiences with HIE may not be reflective of all public health practitioners. The sample was limited to those experienced with HIE

and the sample represented more developed, functional exchange efforts. Furthermore, while we were able to document benefits, we were not able to quantify any of the reported gains in efficiency, organizational learning, better decision making, or health outcomes.

Conclusion

HIE supported public health activities consistent with expectations in the literature. However, no single department realized the all the potential benefits of HIE. The opportunity exists for public health to make use of HIE to a greater extent. Additionally, HIE efforts can support public health even further by increasing the number of participating providers and increasing the amount of information available.

ACKNOWLEDGEMENTS

We would like to thank the health information exchange efforts and the public health departments for their cooperation in data collection: Rochester Regional Health Information Organization (www.grrhio.org), Southern Tier HealthLink (www.sthlny.com), the Integrated Care Collaborative of Central Texas (icc-centex.org), the Indiana Health Information Exchange (www.ihie.com) and the New Mexico Health Information Collaborative (www.nmhic.org). This project was funded by the New York eHealth Collaborative as a part of the evaluation of the Office of the National Coordinator for Health Information Technology's State HIE Cooperative Agreement Program. The Institutional Review Board of Weill Cornell Medical College approved this study.

REFERENCES

1. Brailer DJ. Guiding The Health Information Technology Agenda. Health Affairs. 2010;29(4):586-95.
2. The White House. Transforming Health Care: The President's Health Information Technology Plan. 2004 [03 Feb 2014]; Available from: http://georgewbush-whitehouse.archives.gov/infocus/technology/economic_policy200404/chap3.html.
3. Office of the National Coordinator for Health Information Technology. State Health Information Exchange Cooperative Agreement Program. 2012 [13 SEP 2012]; Available from: http://healthit.hhs.gov/portal/server.pt?open=512&objID=1488&mode=2.
4. Kern LM, Barron Y, Abramson EL, Patel V, Kaushal R. HEAL NY: Promoting interoperable health information technology in New York State. Health Affairs. 2009;28(2):493-504. Epub 2009/03/12.
5. Delaware Health Information Network. About DHIN. 2012 [13 SEP 2012]; Available from: http://www.dhin.org/about.
6. Vest J, Gamm LD. Health information exchange: persistant challenges & new strategies. Journal of the American Medical Informatics Association. 2010;17(3):288-94.
7. American Recovery and Reinvestment Act of 2009, (2009).
8. An act to amend the public health law, in relation to establishing a statewide capital grant program to improve the quality and efficiency of the health care delivery system, New York State Assembly, 2003-2004 Sess. (2004).
9. President's Council of Advisors on Science and Technology. Report to the President. Realizing the Full Potential of Health Information Technology to Improve Healthcare for Americans: The Path Forward. Washington, DC: Executive Office of the President, 2010.
10. Shapiro JS, Mostashari F, Hripcsak G, Soulakis N, Kuperman G. Using health information exchange to improve public health. American journal of public health. 2011;101(4):616-23. Epub 2011/02/19.
11. National Association of County & City Health Officials. 2010 National Profile of Local Health Departments. Washington, DC: 2011.
12. Reeder B, Revere D, Hills RA, Baseman JG, Lober WB. Public Health Practice within a Health Information Exchange: Information Needs and Barriers to Disease Surveillance. Online journal of public health informatics. 2012;4(3). Epub 2012/01/01.
13. Vest J, Kern L, Campion TR, Jr., Silver M, Kaushal R, for the HITEC Investigators. Association between use of a health information exchange system and hospital admissions. Applied Clinical Informatics. 2014;5(1):219-31.
14. Campion TR, Jr., Ancker JS, Edwards AM, Patel VN, Kaushal R, Investigators H. Push and pull: physician usage of and satisfaction with health information exchange. AMIA Annu Symp Proc. 2012;2012:77-84. Epub 2013/01/11.

15. Vest J, Zhao H, Jasperson J, Gamm LD, Ohsfeldt RL. Factors motivating and affecting health information exchange usage. Journal of the American Medical Informatics Association. 2011;18(2):143-9.
16. Biondich PG, Grannis SJ. The Indiana Network for Patient Care: an integrated clinical information system infomed by over thirty years of experience. J Public Health Management Practice. 2004;November(Suppl):S81-S6.
17. Parsons W, Gunter M, Kroth P, Fillmore D. PS1-50: Implementation and Evaluation of a Health Information Exchange (HIE). Clinical Medicine & Research. 2012;10(3):164-5.
18. Crabtree BF, Miller WL, editors. Doing Qualitative Research. 2nd ed. Thousand Oaks, CA: Sage Publications; 1999.
19. Morse JM. The Significance of Saturation. Qualitative Health Research. 1995;5(2):147-9.
20. King N. Doing Template Analysis. In: Symon G, Cassell C, editors. Qualitative Organizational Research Core Methods and Current Challenges. London: Sage Publications; 2012.
21. King N. Template analysis. In: Symon G, Cassell C, editors. Qualitative methods and analysis in organizational research: A practical guide. Thousand Oaks, CA: Sage Publications Ltd; 1998. p. 118-34.
22. Shapiro JS. Evaluating public health uses of health information exchange. Journal of biomedical informatics. 2007;40(6 Suppl):S46-9. Epub 2007/10/09.
23. Dixon BE, McGowan JJ, Grannis SJ. Electronic laboratory data quality and the value of a health information exchange to support public health reporting processes. AMIA Annu Symp Proc. 2011;2011:322-30. Epub 2011/12/24.
24. Dobbs D, Trebatoski M, Revere D. The Northwest Public Health Information Exchange's Accomplishments in Connecting a Health Information Exchange with Public Health. Online journal of public health informatics. 2010;2(2). Epub 2010/01/01.
25. Grannis SJ, Stevens KC, Merriwether R. Leveraging health information exchange to support public health situational awareness: the indiana experience. Online journal of public health informatics. 2010;2(2). Epub 2010/01/01.
26. Hessler BJ, Soper P, Bondy J, Hanes P, Davidson A. Assessing the relationship between health information exchanges and public health agencies. Journal of public health management and practice : JPHMP. 2009;15(5):416-24. Epub 2009/08/26.
27. Kass-Hout TA, Gray SK, Massoudi BL, Immanuel GY, Dollacker M, Cothren R. NHIN, RHIOs, and Public Health. Journal of public health management and practice : JPHMP. 2007;13(1):31-4. Epub 2006/12/07.
28. Magruder C. Public health/ health information exchange collaborative: a model for advancing public health practice. Online journal of public health informatics. 2010;2(2). Epub 2010/01/01.
29. Nangle B, Xu W, Sundwall DN. Mission-driven priorities: public health in health information exchange. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2009;2009:468-72. Epub 2009/01/01.
30. Trebatoski M, Davies J, Revere D, Dobbs D. Methods for Leveraging a Health Information Exchange for Public Health: Lessons Learned from the NW-PHIE Experience. Online journal of public health informatics. 2010;2(2). Epub 2010/01/01.
31. QSR. NVivo10 for Windows Help. 2014 [12 FEB 2014]; Available from: http://help-nv10.qsrinternational.com/desktop/procedures/run_a_coding_comparison_query.htm.
32. Papadouka V, Metroka A, Zucker JR. Using an Immunization Information System to Facilitate a Vaccine Recall in New York City, 2007. Journal of Public Health Management and Practice. 2011;17(6):565-8.
33. Smith N, Iyer R, Langer-Gould A, Getahun D, Strickland D, Jacobsen S, et al. Health plan administrative records versus birth certificate records: quality of race and ethnicity information in children. BMC Health Services Research. 2010;10(1):316.
34. Vest JR, Issel LM. Factors Related to Public Health Data Sharing between Local and State Health Departments. Health Serv Res. 2014;49(1 Pt 2):373-91. Epub 2013/12/24.
35. Vest J, Issel L, Lee S. Experience of using information systems in public health practice: Findings from a qualitative study. Online Journal of Public Health Informatics. 2014;5(3):1-17.
36. Vest J, Campion TR, Jr., Kaushal R, investigators ftH. Challenges, alternatives, and paths to sustainability for health information exchange efforts. Journal of Medical Systems. 2013;37(6):9987.
37. Centers for Disease C, Prevention. HIPAA privacy rule and public health. Guidance from CDC and the U.S. Department of Health and Human Services. MMWR Morb Mortal Wkly Rep. 2003;52 Suppl:1-17, 9-20. Epub 2003/05/14.

Trends in Publication of Nursing Informatics Research

Hyeoneui Kim, RN, MPH, PhD, Lucila Ohno-Machado, MD, MBA, PhD,
Janet Oh, BS, Xiaoqian Jiang, PhD
Division of Biomedical Informatics, University of California San Diego, La Jolla, CA

Abstract

We analyzed 741 journal articles on nursing informatics published in 7 biomedical/nursing informatics journals and 6 nursing journals from 2005 to 2013 to begin to understand publication trends in nursing informatics research and identify gaps. We assigned a research theme to each article using AMIA 2014 theme categories and normalized the citation counts using time from publication. Overall, nursing informatics research covered a broad spectrum of research topics in biomedical informatics and publication topics seem to be well aligned with the high priority research agenda identified by the nursing informatics community. The research themes with highest volume of publication were Clinical Workflow and Human Factors, Consumer Informatics and Personal Health Records, and Clinical Informatics, for which an increasing trend in publication was noted. Articles on Informatics Education and Workforce Development; Data Mining, NLP, Information Extraction; and Clinical Informatics showed steady and high volume of citations.

Introduction

Information technologies were identified as key factors in achieving improved patient safety and quality of care and the published literature is starting to confirm this role[1–3]. Biomedical informatics, including nursing informatics, is a fast moving field that is heavily influenced by healthcare policy and clinical practice. Professional and/or academic organizations have been dedicating substantial effort to facilitate nursing informatics research and training that address current needs. For example, many nursing education programs offer training opportunities specialized in nursing informatics and also mandate an introductory nursing informatics class to all students as part of their core curriculum[4–7]. The Nursing Informatics Working Group (NIWG) at AMIA addresses various issues in policy, research, and educational aspects of nursing informatics. The NIWG also offers two award programs to recognize noteworthy research presented in the annual AMIA symposium to encourage participation of nurse scholars [8].

Scientific journals are a venue for dissemination of research to a large community of readers. The proportion of peer-reviewed articles published in a certain area of research (and their citations) can provide interesting insights on topic trends in informatics[9,10]. Although it may not constitute a perfect measure, the number of citations is considered a de facto standard for measuring the impact of a scientific publication. Based on this idea, we have previously analyzed the publication and citation volumes in biomedical informatics that were published relatively recently (2009-2012) in the J Amer Med Inform Assoc (JAMIA) to better understand research topic trends for biomedical informatics in general.

We conducted a similar analysis to describe the trends in published nursing informatics research by analyzing the articles published on selected major informatics and nursing journals during the past 9 years. We also compared the active research areas reflected in these articles against the nursing informatics research agenda proposed by nursing informatics leaders[11–13]. Through this analysis, we aimed at (1) checking the trajectory of published nursing informatics research for the past 9 years, and (2) identifying potential gaps in particular research areas.

Background

In 1993, the National Institute of Nursing Research (NINR) of National Institutes of Health (NIH) sponsored a group of nursing informatics scholars in the investigation of research needs in nursing informatics with the purpose of developing a nursing informatics research agenda[14]. Healthcare and biomedical research are changing rapidly due to the availability of advanced technologies to collect and analyze large volumes of data.

Many nursing informatics leaders have proposed an updated research agenda over the past decade. In 2007, the AMIA NIWG proposed comprehensive nursing informatics agenda in nursing practice, education, and research [13]. A year later, Bakken et, al. identified the areas that nursing informatics research needed to expand further to better

accommodate fast moving biomedical sciences[12]. In 2012, the Nursing Informatics International Research Network (NIIRN) conducted an international survey of 468 nursing informatics researchers across the globe soliciting opinions on research topics that should be prioritized[11]. The research agenda identified in these works is summarized in **Table 1**.

Recently, Carrington et, al. reported on informatics topics that were most actively researched, the types of research, and research settings by reviewing 69 peer-reviewed articles with a nurse as the first author, published between Aug 2011 and Aug 2012[15]. They identified three broad topics of research: (1) clinical informatics research that deals with various clinical information applications such as Electronic Medical Records (EMR) and bar-code medication administration, (2) human factors such as human computer interaction and communication, and (3) data interoperability such as terminology/standardization and care transition/handoffs[15]. This work provided an important snapshot view of the research trends in that given year and provided valuable insights on the status of nursing informatics research.

Table 1. Nursing informatics research agenda

AMIA-NIWG (2007)[13]	**Bakken, et al. (2008)**[12]	**NIIRN (2012)***[11]
• Secondary use of clinical data • Use of aggregated & de-identified data • Data privacy and confidentiality • Data security and related technical infrastructure • Population health • Development and use of standards	• User friendly tools for data analysis visualization, and modeling • Evaluation methods of human and organizational factors on health IT • Consumer empowerment • Use of genomics and environmental data • Reengineering of nursing practice	• Clinical information systems that provide real-time feedback • Evaluation of the impact of health IT system on patient outcome • Nursing decision support systems • Evaluation of the impact of health IT system on nursing workflow • Management of nursing data for research and patient care • Training nurses in health IT • Identifying nursing outcomes that are important to patients

* Only the top 7 highest priority items out of 20 are listed.

Materials and Methods

Article collection

We retrieved articles relevant to nursing informatics published between 2005 and 2013 using ISI's Web of Science[16]. Here, the "articles relevant to nursing informatics" was defined according to the disciplines covered by the selected journals, keywords found in the articles, and authors' affiliations. Articles published in nursing informatics journals were considered relevant. Articles published in general biomedical informatics journals were considered relevant when at least one of the authors was affiliated with a nursing institution (e.g., nursing schools, nursing research institutes, centers of nursing excellence, etc.). Additionally, articles published in non-informatics nursing journals were considered relevant if they contained the keyword "informatics" in the title or abstract.

Nursing informatics research is published in a large number of biomedical informatics and/or engineering journals. We included one nursing informatics journal (*CIN: Computers, Informatics, Nursing*) and seven general biomedical informatics journals (*Journal of Medical Internet Research, Journal of American Medical Informatics Association, Medical Decision Making, International Journal of Medical Informatics, Journal of Biomedical Informatics, BMC Medical Informatics and Decision Making, Method of Informatics in Medicine*) in this analysis. These journals encompass a wide range of informatics topics and are among the top 20 journals with highest 5-year impact factors in the category "Medical Informatics" in ISI's Web of Science [16]. Four[1] of these eight journals were also recognized in a prior work as the journals in which the most nursing informatics works were published [15].

As many nursing informatics scholars publish informatics relevant works on non-informatics journals, we also included six nursing journals (i.e., *International Journal of Nursing Studies, Journal of Advanced Nursing, Nursing Research, Journal of Nursing Scholarship, Nursing Outlook, Journal of Nursing Administration*) that publish

[1] CIN: Computers, Informatics, Nursing; Journal of American Medical Informatics Association; International Journal of Medical Informatics; Journal of Biomedical Informatics

general clinical nursing research as well as administrative, and managerial topics. These journals are among the top 20 nursing journals with highest 5-year impact scores under the category "Nursing" of ISI's Web of Science[16]. This journal selection approach undoubtedly has limitations, which are described in the discussion section.

We retrieved citation information of selected articles from the ISI's Web of Science[16] using its citation report function. A total of 741 articles were included in this analysis, which excluded white papers, opinion papers, duplicated articles, and articles without citation reports. The largest number of articles (N=406) was collected from the nursing informatics journal, 306 from biomedical informatics journals, and 29 from nursing journals. The paper selection process is summarized in **Figure 1**.

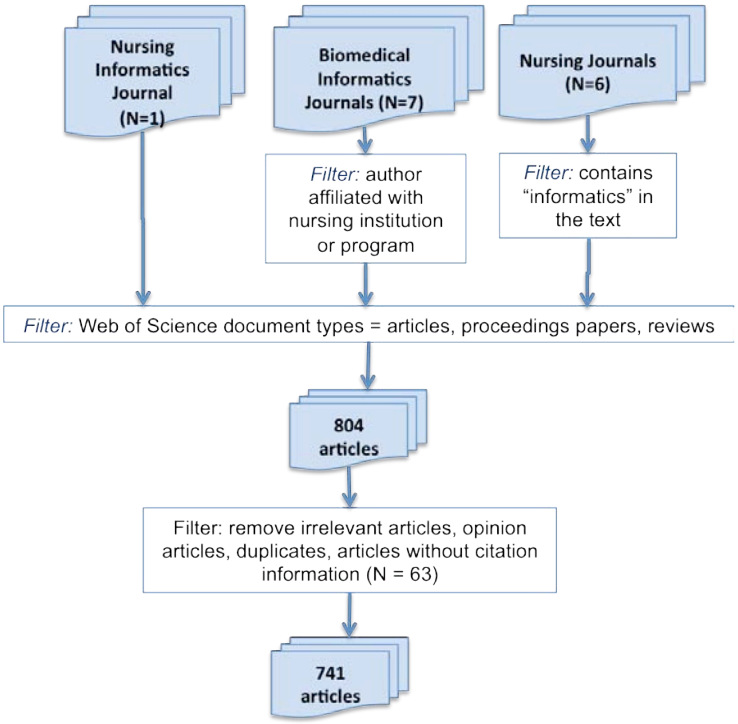


Figure 1. Article selection process

Citation score normalization

The citation report obtained from the ISI's Web of Science[16] shows the annual citation counts of an article for the indicated time period. Using the same method reported in the prior work[9,10] we produced normalized citation scores to each paper to minimize the bias introduced by the differences in publication dates. For instance, an article published in January of 2006 has more chances of being read and cited than an article published in December of 2011 as it has been available to research communities for a longer period of time. Therefore we used a unit of measure that reflects the number of citation per month (CPM) since publication, as adopted in the prior work [9,10].

The CPM scores were calculated separately for each year (i.e., annual CPM = number of citation in a given year / number of months an article was available in a given year) as well as for the entire period (i.e., overall CPM = total number of citations / total number of months an article was available). The two calculations were performed to account for the fact that articles that are old accrue a relatively small number of citations per month after several years of publication, and articles that are recent also accrue a relatively small number of citations in the first two years after publication. Hence articles published in the middle of the accounted period (i.e, 2009-2010) would account for higher overall CPMs than articles at the extremes. However, they should result in comparable annual CPM when the number of years after publication is taken into account.

Research theme assignment

We assigned a research topic category to each article using the AMIA themes proposed for the 2014 annual symposium[17]. Although the articles provide a list of keywords and/or are indexed with the terms from Medical Subject Heading (MeSH), they may not provide a single category label that represents the main topic of the study with sufficient expressivity for subdomains of informatics. We decided to use the AMIA 2014 themes after investigating various potential alternatives, such as chapter headings from biomedical informatics textbooks and subjects listed in informatics training curricula[18,19]. Although AMIA themes may focus on the contemporary topics of biomedical informatics research, they were deemed viable options for this work, as they are expressive labels designed to capture a single main topic of a research work for the purposes of directing a diverse set of attendees to the presentations that are more relevant to their interests.

We also developed annotation rules and a process shown in **Figure 2** to maximize annotation consistency and accuracy. The theme categories that were deemed straightforward to use (i.e., *Public Health Informatics and Biosurveillance, Translational Bioinformatics and Bioinformatics, Global eHealth, Imaging Informatics*) are not shown in **Figure 2** to simplify the display of the process. In addition, we did not use the theme category *Meaningful Use* as it can directly relate to many other informatics themes and was not relevant to older articles.

Two authors (HK, XJ) independently annotated the main research topics of randomly selected 100 articles using the AMIA 2014 themes. The annotation results were collaboratively reviewed and the discrepancies were resolved by reaching consensus. Also, the theme assignment rules and process were refined as necessary. One author (HK) who is trained in nursing informatics assigned themes to the remaining 641 articles following the refined rules and process. Metadata for the 741 articles with themes are available at https://idash-data.ucsd.edu/folder/504.

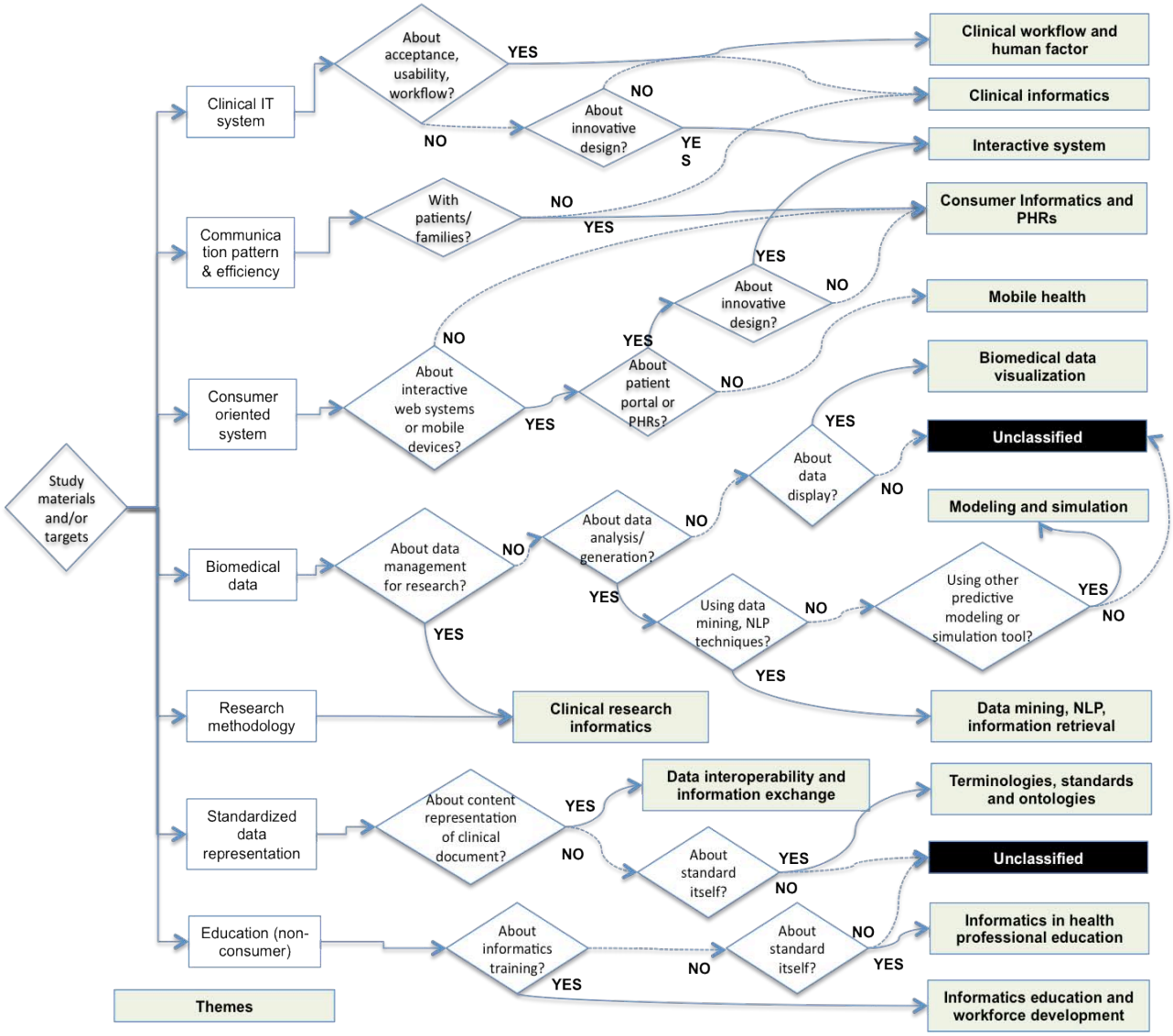


Figure 2. Topic assignment process

Trend analysis

We generated descriptive statistics on distributions of the research topics and their citation scores. We also tested whether these distributions differ significantly among journal types and whether they changed significantly over the years. SPSS (version 22) and Matlab were used to analyze the data.

Results

Among 741 articles collected, 701 were research articles and 40 were review articles. A total of 117 articles reported on work that targeted specific clinical information systems. The most frequently studied areas referred to Electronic Medical Records (EMR, N=61), Clinical Decision Support System (CDSS, N=33), Computerized Provider Order Entry (CPOE, N=10), and bar-coding medication administration (N=10).

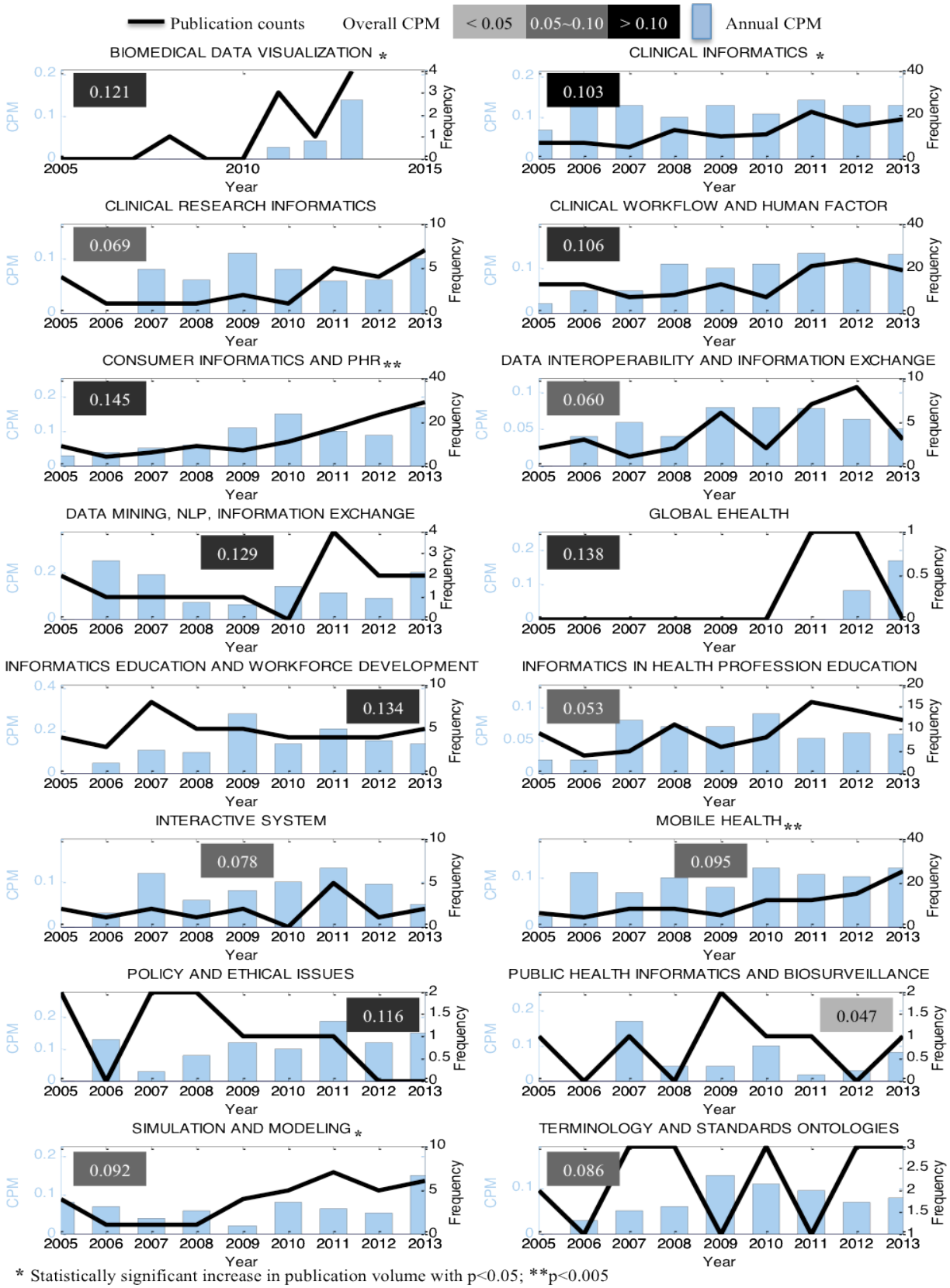


Figure 3 Annual article volume and Citation per Month (CPM), and overall CPM by Themes

Distribution of research themes

Most of the themes were utilized in the categorization of articles, except for the themes *Imaging Informatics* and *Translational Bioinformatics and Biomedicine* for which there were no articles. No article was left unassigned. The research theme most frequently appearing for the past 9 years was *Clinical Workflow and Human Factors*. *Global eHealth* was the least frequent (**Figure 3**). We analyzed the trend of publication frequencies for each research theme over the years using a parametric method. The article frequencies were used as the predictor variables and the publication year was used as the response variable to fit a linear regression model. Two parameters (constant and slope) were calculated and we conducted a t-test to check whether they were statistically significant. In **Figure 3**, the starred (*) themes are the ones that show statistically significant positive trends (i.e., increment of the article counts).

Publications focusing on *Consumer Informatics and PHR* and *Mobile Health* have been continuously growing. In addition to these two themes, *Clinical Informatics and Clinical Workflow and Human Factors* are the areas where the high volume of nursing informatics articles has been published for the past 3 years. Also it is noteworthy that the nursing informatics articles on *Biomedical Data Visualization* and *Global eHealth* started showing up relatively recently.

The articles published in informatics journals represented the broadest theme areas. Two most frequent themes from the nursing informatics journal were *Informatics in Health Professional Education*, and *Clinical Workflow and Human Factor*. *Consumer Informatics and PHRs*, and *Clinical Workflow and Human Factor* are the themes most frequently appearing in general biomedical informatics journals. Half of the articles from nursing journals were on *Informatics Education and Workforce Development* (**Figure 4**). The differences in theme distributions among journal types were statistically significant when tested with the Pearson Chi-Square test ($p<0.005$).

Table 2. Distribution of research themes by journal types

AMIA Themes	Nursing Informatics Journal	Nursing Journal	Biomedical Informatics Journal	Total
Clinical workflow and human factor	76	2	47	125
Consumer informatics and PHR	46	1	68	115
Clinical informatics	60	3	44	107
Mobile health	59	0	36	95
Informatics in health profession education	76	5	4	85
Informatics education and workforce development	22	15	5	42
Data interoperability and information exchange	21	0	14	35
Simulation and modeling	10	1	23	34
Clinical research informatics	12	0	14	26
Terminology and standards ontologies	5	1	14	20
Interactive system	10	0	6	16
Data mining, NLP, information extraction	5	0	9	14
Biomedical data visualization	2	0	7	9
Policy and ethical issues	0	1	8	9
Public health informatics and bio-surveillance	1	0	6	7
Global ehealth	1	0	1	2
Total	406	29	306	741

Research impact

The top 10 most highly cited articles (based on overall CPM) are about *Informatics Education and Workforce Development* (ranked 1st)[20], *Biomedical Data Visualization* (2nd)[21], *Clinical Workflow and Human Factor* (3rd)[22] and *Consumer Informatics and PHRs* (4th-10th). The first ranked article was published in a non-informatics nursing journal and the other 9 articles were published in general biomedical informatics journals.

Informatics Education and Workforce Development; Data Mining, NLP, Information Extraction; and Clinical Informatics were the research themes that were associated with the consistently high annual CPMs (**Figure 3**). However, *Consumer Informatics and PHRs* came out as having the highest overall CPM, followed by *Global eHeath* and *Informatics Education and Workforce Development.* It is noteworthy that the articles on *Global eHealth* and *Biomedical Data Visualization* themes showed quite high overall CPM scores in spite of their relatively short article available time. This indicates that these two themes are fastest growing research areas in nursing informatics but this observation needs to be revalidated when more data are accumulated. The overall CPMs for *Data Mining*, and *NLP, Information Extraction* and *Clinical Informatics* still remain high.

Alignment with the nursing informatics research agenda

We first aligned the nursing informatics research agenda items presented in **Table 1** with the AMIA themes. We then checked the number of articles published in those themes and the overall CPM and the most recent stable annual CPM scores as proxy information that reflects how active research in a particular agenda item has been. A research agenda item was assigned with only one theme even if it can be related to multiple themes. We excluded the agenda item "reengineering nursing practice" as it encompasses a full spectrum of research themes.

The agenda items were related to nine of the seventeen AMIA themes considered in this study. The eight of the nine mapped themes covered more than 60% of the articles analyzed in this study. The average overall CPM score of the articles on these eight themes was 0.102 (s.d. = 0.033). Of note, the average overall CPM score of the seventeen themes used in this study was 0.010 (s.d. = 0.027) (**Table 2**).

Discussion

We observed that the nursing informatics articles covered a broad spectrum of topics in biomedical informatics over a 9-year window. Highest volumes of articles were associated with *Clinical Workflow and Human Factor, Consumer Informatics and PHRs,* and *Clinical Informatics*, for which an increasing trend in publication was also noted. An increasing publication trend was also observed for the themes like *Mobile Health, Informatics in Health Profession Education, Simulation and Modeling,* and *Biomedical Data Visualization.*

In terms of citation counts, themes related to informatics training, clinical informatics and data mining dominated, possibly because they are established areas of research. When we adjust citation counts with publication dates (i.e., overall CPM), *Consumer Informatics and PHRs* and *Global eHealth* stood out as the most highly cited themes. However, the overall CPM of *Global eHealth* was calculated based on the 2 articles published only 2-3 years ago. Therefore, this high overall CPM may indicate temporary attention caused by the paucity of relevant publications thus a longer period of observation is needed to substantiate this trend.

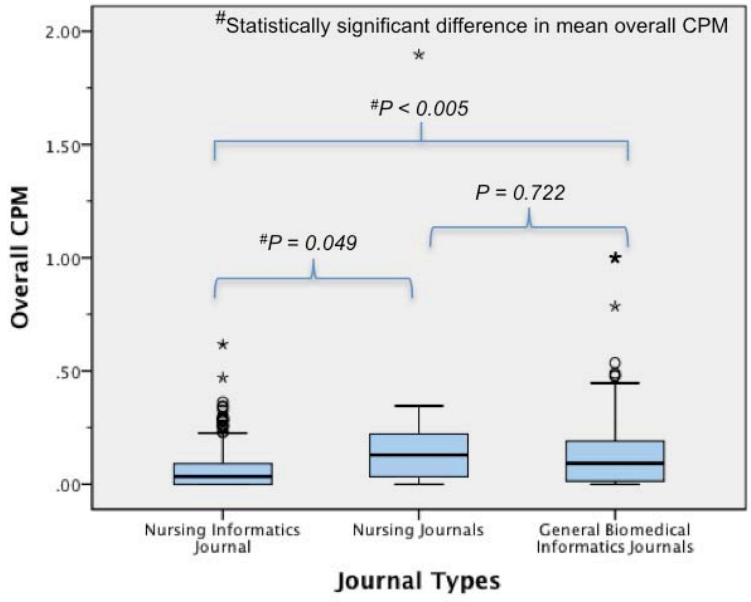


Figure 4. Overall Citations Per Month (CPM) distribution by journal type

We found that education topics published in non-informatics nursing journals are very highly cited compared to those published in informatics journals. In addition, we found that the articles published in the nursing informatics journal have relatively lower overall CPMs than the articles published in general biomedical informatics journals and non-informatics nursing journals. (**Figure 5**). This is probably related to the size of readership (the audience pool), which is much smaller for nursing informatics as it is for nursing in general, or for informatics in general.

Although most of the high priority areas of nursing informatics research were well covered, areas that may require more attention – related to public health informatics and bioinformatics – were also noted.

Table 2. Volume of articles, overall CPM, and most recent annual CPM (2012) by research agenda

Agenda items (proposed by)	Themes	Articles N (%)	CPMs: overall (2012)
Clinical information systems that provide real-time feedback (NIIRN)	Clinical Informatics	107 (14.44)	0.105 (0.130)
Nursing decision support system (NIIRN)			
Secondary use of clinical data (AMIA-NIWG)			
User friendly tools for data analysis visualization, and modeling (Bakken et, al.)			
Evaluation of the impact of health IT system on patient outcome (NIIRN)			
Evaluation methods of human and organizational factors on health IT (Bakken et, al.)	Clinical Workflow and Human Factor	125 (16.87)	0.109 (0.112)
Evaluation of the impact of health IT system on nursing workflow (NIIRN)			
Use of aggregated & de-identified data (AMIA-NIWG)	Clinical Research Informatics	26 (3.51)	0.069 (0.061)
Data privacy and confidentiality (AMIA-NIWG)			
Data security & related technical infrastructure (AMIA-NIWG)			
Management of nursing data for research patient care (NIIRN)			
Consumer empowerment (Bakken et, al.)	Consumer Informatics and PHRs	115 (15.52)	0.148 (0.087)
Identifying nursing outcomes important to patients (NIIRN)			
Population health (AMIA-NIWG)	Public Health Informatics and Biosurveillance	5 (0.94)	0.054 (0.028)
Use of genomics and environmental data (Bakken et, al.)	Translational Bioinformatics & Biomedicine	0	NA
Development and use of standards (AMIA-NIWG)	Terminology and Standards Ontologies	20 (2.70)	0.086 (0.070)
	Data interoperability and information exchange	35 (4.72)	0.059 (0.063)
Training nurses in Health IT (NIIRN)	Informatics Education and Workforce Development	27 (5.67)	0.137 (0.153)

It is noteworthy that *Clinical Workflow and Human Factor* is among the most frequently published and cited themes in nursing informatics research. This is not surprising considering the large amount of time that nurses spend working with clinical information system, especially EMRs[23]. This research trend may also imply that nurses can make significant contributions to mitigating various EMR usability issues. In July of 2013, the Office of National Coordinator (ONC) hosted a meeting with stakeholders to discuss usability issues in EMRs. Representatives from a healthcare informatics professional group, an EMR certification agency, and the American Medical Association were invited and had a chance to provide testimonials. Common criticism was that designs and workflows of many EMRs largely focused on meeting the Meaningful Use (MU) requirements rather than usability[24–26]. These testimonials addressed the usability issues commonly faced by various healthcare professionals. Also not many MU core/menu objectives are directly related to nursing documentation. Nonetheless it would be beneficial to actively seek nursing perspectives on the usability issues, considering that nurses are the major workforce for documenting and generating clinical data, as well as being heavily involved in the research on usability and workflow aspects of clinical information systems.

The findings of this study may not reflect the most comprehensive and accurate trends of nursing informatics research due to several limitations. Firstly, the operational definitions of nursing informatics research we adopted were arbitrary: research related to informatics published in nursing journals OR research done by a researcher affiliated with a nursing institution and published in an informatics journal. Ideally, nursing informatics research should be defined as any scholarly work that investigated informatics aspects relevant to nursing research, practice and education. However, operational definitions that could be readily applicable to literature search were

unavailable. Use of crude definitions of nursing informatics research may have caused either the inclusion of non-nursing informatics articles or exclusion of nursing informatics articles.

Secondly, we collected articles from a small subset of informatics and nursing journals. For example, we excluded conference papers with no citation information and specialty nursing journals, where many practical applied informatics articles are published, as they can be most beneficial to practicing nurses. Thirdly, although we used CPM to minimize the bias introduced by publication dates, CPM is still not completely free from the bias given that trends captured in the early age of articles might drastically change later. Finally, even if we found AMIA themes were robust and clear enough to label the articles collected for this study, there were cases that the authors struggled with assigning proper themes. Like in many studies that involve with human annotation of concepts, this study also carries subjectivity and inconsistency issues in theme assignment.

Conclusion

We conducted a descriptive analysis of nursing informatics articles published in the past 9 years to understand overall trends in publication of nursing informatics research in the peer-reviewed literature and to identify the research areas that may be relatively under-represented. Overall, we found that published nursing informatics research covered a broad spectrum of research topics in biomedical informatics and is aligned with high priority research agenda items advocated by the nursing informatics community. Despite limitations related to relatively narrow coverage of journals, the findings allow us to initiate a dialogue with journal editors and the nursing informatics community at large about the differences in maturity of research in each of the themes, and potentially guide authors and readers on which venue currently provides the highest number of articles related to a given theme.

Acknowledgement

This study was supported in part by NIH grant U54HL108460.

References

1. HITECH Act Enforcement Interim Final Rule. U.S. Department of Health and Human Services; [cited 2014 Mar 12]; Available from: http://www.hhs.gov/ocr/privacy/hipaa/administrative/enforcementrule/hitechenforcementifr.html
2. Meaningful_Use. 2013 Dec 6 [cited 2014 Mar 12]; Available from: http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/Meaningful_Use.html
3. The Future of Nursing: Leading Change, Advancing Health - Institute of Medicine [Internet]. [cited 2014 Mar 10]. Available from: http://www.iom.edu/Reports/2010/The-future-of-nursing-leading-change-advancing-health.aspx
4. Skiba DJ, Rizzolo MA. National League for Nursing's informatics agenda. Comput Inform Nurs [Internet]. [cited 2014 Mar 10];27(1):66–8.
5. Skiba DJ. Moving forward: the informatics agenda. Nurs Educ Perspect [Internet]. [cited 2014 Mar 10];29(5):300–1.
6. Carrington JM, Tiase VL, Estrada N, Shea KD. Nursing education focus of nursing informatics research in 2013. Nurs Adm Q [Internet]. [cited 2014 Mar 10];38(2):189–91.
7. TIGER Initiative [Internet]. [cited 2014 Mar 11]. Available from: http://www.thetigerinitiative.org/
8. Nursing Informatics | AMIA [Internet]. [cited 2014 Mar 12]. Available from: http://www.amia.org/programs/working-groups/nursing-informatics
9. Jiang X, Tse K, Wang S, Doan S, Kim H, Ohno-Machado L. Recent trends in biomedical informatics: a study based on JAMIA articles. J Am Med Inform Assoc [Internet]. 2013 Dec [cited 2014 Feb 20];20(e2):e198–205.
10. Kim H-E, Jiang X, Kim J, Ohno-Machado L. Trends in biomedical informatics: most cited topics from recent years. J Am Med Inform Assoc [Internet]. 2011 Dec [cited 2014 Feb 28];18 Suppl 1:i166–70.
11. Dowding DW, Currie LM, Borycki E, Clamp S, Favela J, Fitzpatrick G, et al. International priorities for research in nursing informatics for patient care. Stud Health Technol Inform [Internet]. 2013 Jan [cited 2014 Mar 10];192:372–6.

12. Bakken S, Stone PW, Larson EL. A nursing informatics research agenda for 2008-18: contextual influences and key components. 2008. Nurs Outlook [Internet]. [cited 2014 Mar 10];60(5):280–288.e3.
13. McCormick KA, Delaney CJ, Brennan PF, Effken JA, Kendrick K, Murphy J, et al. Guideposts to the future--an agenda for nursing informatics. J Am Med Inform Assoc [Internet]. [cited 2014 Feb 8];14(1):19–24.
14. NINR Priority Expert Panel on Nurisng Informatics. Nursing informatics: Enhancing patient care. 1993.
15. Carrington JM, Tiase VL. Nursing informatics year in review. Nurs Adm Q [Internet]. [cited 2014 Mar 10];37(2):136–43. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23454993
16. Web of Science [v.5.13.1] - Web of Science Core Collection Home [Internet]. [cited 2014 Mar 12]. Available from: http://apps.webofknowledge.com
17. AMIA 2014 Call for Participation | AMIA [Internet]. [cited 2014 Mar 12]. Available from: http://www.amia.org/amia2014/call-for-participation
18. Kampov-Polevoi J, Hemminger BM. A curricula-based comparison of biomedical and health informatics programs in the USA. J Am Med Inform Assoc [Internet]. [cited 2014 Jan 23];18(2):195–202.
19. Biomedical Informatics: Computer Applications in Health Care and Biomedicine [Internet]. Springer; 3rd edition; 2012 [cited 2014 Mar 13].
20. Cronenwett L, Sherwood G, Barnsteiner J, Disch J, Johnson J, Mitchell P, et al. Quality and Safety Education for Nurses. Nurs Outlook [Internet]. [cited 2014 Feb 26];55(3):122–31.
21. Demiris G, Thompson HJ, Reeder B, Wilamowska K, Zaslavsky O. Using informatics to capture older adults' wellness. Int J Med Inform [Internet]. 2013 Nov [cited 2014 Feb 20];82(11):e232–41.
22. Mei YY, Marquard J, Jacelon C, DeFeo AL. Designing and evaluating an electronic patient falls reporting system: perspectives for the implementation of health information technology in long-term residential care facilities. Int J Med Inform [Internet]. 2013 Nov [cited 2014 Feb 5];82(11):e294–306.
23. Kossman SP. Perceptions of impact of electronic health records on nurses' work. Stud Health Technol Inform [Internet]. 2006 Jan [cited 2014 Mar 13];122:337–41.
24. AMA. Health IT policy committee's workgroups on certification/adoption and implementation: Testimony of the American Medical Association - Implementation and Usability of Certified Electronic Health Records. 2013.
25. HIMSS. Testimony to the HIT Panels on Implementation/Usability Office of the National Coordinator [Internet]. 2013 [cited 2014 Mar 13]. Available from: http://www.healthit.gov/sites/default/files/archive/FACA Hearings/2013-07-23 Standards: Implementation, Meaningful Use, and Certification & Adoption WGs, Implementation & Usability Hearing/onc-staggers072213.pdf
26. CCHIT. Implementation and Usability of Meaningful Use: Usability Panel [Internet]. [cited 2014 Mar 13]. p. 2013. Available from: http://www.healthit.gov/sites/default/files/archive/FACA Hearings/2013-07-23 Standards: Implementation, Meaningful Use, and Certification & Adoption WGs, Implementation & Usability Hearing/iu_mu_raytestimony072313.pdf

Effectiveness of Evidence-Based Congestive Heart Failure (CHF) CPOE Order Sets Measured by Health Outcomes

Jacob Krive, PhD, MBA, MS[1,2,3], Joel S. Shoolin, DO, MBA[1], Steven D. Zink, PhD[2,4].
[1]Advocate Health Care, Oak Brook, IL; [2]Nova Southeastern University, Fort Lauderdale, FL; [3]University of Illinois at Chicago, Chicago, IL; [4]Nevada System of Higher Education, NV

Abstract

Objective: *Evidence-based order sets for treatment of patients with common conditions promise ordering efficiency and more consistent health outcomes. Despite ongoing utilization of order sets, quantitative evidence of their effectiveness is lacking. This study quantitatively analyzed benefits of CHF order sets as measured by mortality, readmission, and length of stay (LOS) outcomes.*
Methods*: Mortality and readmissions were analyzed by comparing "order set" and "free text" groups of adult patients using logistic regression, Pearson chi-squared, and Fisher's exact methods. LOS was calculated by applying One-Way ANOVA and Mann-Whitney tests, supplemented by comorbidity analysis via Charlson Comorbidity Index.*
Results*: CHF orders placed via sets were effective in reducing mortality [OR=1.818;95% CF 1.039-3.181;p=0.034] and LOS [F(1,10938)=8.352,p=0.013,4.75 days ("free text" group) vs. 5.46 days ("order set" group)], while readmission outcome was not significant [OR=0.913;95% CF 0.734-1.137;p=0.417].*
Conclusion*: Evidence-based medication ordering practices to treat CHF have potential to reduce mortality and LOS, without effect on readmissions.*

Keywords

Evidence-based medicine, medication order sets, health outcomes research, congestive heart failure, computerized physician order entry systems.

Introduction

For decades, hospitals have used templates for computerized physician order entry (CPOE) to standardize clinical workflow. Grounded in evidence-based care theories, physician instructions provide convenience, decrease delays, reduce errors, and improve inventory control. This resonates in the national debate over slowing growth of healthcare expenditures through more focused application of information technology not only in CPOE, but also in electronic medical records (EMR) and clinical decision support (CDS) systems to demonstrate meaningful use as mandated in the U.S. Patient Protection and Affordable Care Act. Aided by the increased sophistication of IT, expectations of CPOE benefits have moved far beyond efficiency. CPOE is now linked to reducing variability in treating well-researched patient conditions by sharpening diagnoses as well as tracking treatment plans throughout the continuum of patient care. Although expectations from CPOE may differ, they are widely deployed throughout the healthcare sector despite little quantitative evidence regarding their effectiveness in improving outcomes or cost reductions.

Early studies in treating chronic heart failure (CHF) found order sets effective in reducing mortality, readmissions, and length of hospital stay. However, none analyzed large data sets spanning several years of patient encounters. This study examines effectiveness of CHF order sets at a major community integrated healthcare delivery network using patient care history from five Advocate Health Care hospitals over five years. Advocate is the largest healthcare provider in the state of Illinois (USA), with 12 hospitals, 3500+ beds, and a total of over 250 sites of care.

Background and Significance

Specialty-specific order sets, engaged physicians, and large-scale CPOE system implementation are crucial to electronic medical records project success[1]. VA Puget Sound Health Care System in Seattle, Washington credits clinician engagement as a key success factor as they saw 50% of the order catalogs, 57% of the quick orders, and 13% of the order sets utilized within 6 months of CPOE implementation[2]. Order sets gain popularity due to their convenience, standardization, and adherence to evidence based practices features. Cowden et al.[3] conducted an order set quality improvement study to combine two orders and evaluate a combined chi-squared to predict order set correspondence to an ordering pattern. Performed at Ohio State University Medical Center, the study found that a large percentage of medication orders qualified for an order set. Dixon and Zafar[4] defined the theoretical foundation of order sets and their potential effectiveness in standardized treatment in a U.S. Department of Health and Human Services study in year 2008. Evidence suggests that order sets influence ordering practices as well as patient outcomes[5]. A study conducted at seven diverse sites across the United States showed many commonalities in physician ordering practices. Overwhelming popularity of a few sets yielded a short list of sets in common circulation[5].

Despite growing CPOE adoption and physician reliance on technology, complexity and specific concerns related to electronic medical records technologies remain high[6]. Resistance to clinical practice guidelines[7] rooted in evidence-based medicine presents a barrier to standardized order practices via sets. To address concerns over standardization and technical complexity issues, in order to take advantage of the potential patient safety benefits of evidence-based order sets, the Mayo Clinic initiated a NIH sponsored project that ensured participation of all clinicians to achieve greater respect, increased trust, and improved utilization of the order sets[8]. UCLA Children's Hospital reported success of their order sets by promulgating residents' 89% approval. Over half of the residents (58%) reported using reported using them >=90% of the time; 75% believed that medical errors were reduced[9].

Only two in-depth studies of the effectiveness of CHF order sets have been reported to date. An Advocate Christ Medical Center study evaluated the effect of improved organization and visual appearance of CHF order sets. Utilization increased for post intervention (72% vs. 9%), as well as compliance with CPG for angiotensin-converting enzyme (83% vs. 25%) and intravenous nitroglycerin (78% vs. 12%), while dosing, systolic blood pressure reduction, and urine output did not change[10]. An implementation of CHF set at Baylor University Medical Center led to increased utilization, increased core measures compliance (odds ratio=1.51), reduced inpatient mortality (odds ratio=0.49), and reduction in readmissions that approached significance[11].

Other studies have demonstrated evidence-based order sets importance in successful patient outcomes. An acute myocardial infarction study of order sets performed at seventy-eight acute care hospitals in Ontario, Canada[12] found increased core measure compliance in the fibrinolytics < 30min, primary percutaneous coronary intervention < 90min, fibrinolytics administration decided by ED physician, aspirin < 6hours of arrival, and lipid test < 24 hours of arrival categories. Patients who met the composite process of care measure had lower 30-day and 1-year mortality with odds ratio of 0.51. A pneumonia order set study at Baylor Health Care System[13] revealed reduced in-hospital mortality (odds ratio=0.73) and 30-day mortality (odds ratio=0.79), while core measure compliance remained unchanged. An earlier similar pneumonia study[14] at the same healthcare system also showed lower mortality and slightly higher core measures compliance (odd ratio=1.24). As part of the order set study with the goal of increasing prescription of calcium and vitamin D supplementation in patients receiving glucocorticoids[15], researchers found that prescriptions for calcium increased from 37% to 49% and vitamin D from 38% to 53% as a result of set implementation. In a study of septic shock management[16], researchers found that order sets were responsible for the after-intervention group receiving more intravenous fluids while in ED, more fluids of >20ml/kg body weight before vasopressor administration, and more likely to be treated with an appropriate initial antimicrobial regimen. These patients also had shorter length of stay and lower risk of 28-day mortality. Another bacteremic severe sepsis management study performed at Barnes-Jewish academic medical center[17] found similar benefits in fluids

administration, hospital length of stay, as well as reduced occurrence of renal and cardiovascular failures among patients in the intervention group.

The literature supports order set effectiveness in improving patient outcomes, with various measures leading to a number of positive conclusions. The number of studies that examined CHF order sets are more limited, and involved smaller patient populations.

Methods

This causal comparative study analyzed five years of patient data between 2007 and 2011, with the goal of determining order set effectiveness as applied to "order set" and "free text" groups of patients based on health outcomes, 30-day readmissions, length of hospital stay, and supported by comorbidity analysis using computed Charlson Comorbidity Index (CCI) score. CHF condition was originally selected due to a relatively large patient volume, higher order set utilization compared to other patient conditions, and organizational key results focus on quality metrics at the research site. As part of comparative study of the patient history, the "order set" group represented patient encounters where providers placed CHF orders using sets, and the "free text" group represented all other CHF treatment orders where physicians chose custom ordering methods and did not employ sets.

Mortality was confirmed by selecting discharge codes complying with ICD-9-CM diagnosis codes listed under AHRQ IQI-16 CHF Mortality definitions. Eligible "expired" codes were converted to a binary value of 1, with other codes converted to 0. Readmission was a 'yes/no' binary field, and length of stay (LOS) was a calculated field between the date/time values of admission and discharge. The total Charlson Comorbidity Index (CCI) score represented comorbidity. In this study, CCI could play a dual role: (1) explain LOS as an index of pre-existing conditions, thus introducing comorbidity adjustment, or (2) serve as a measure of post-treatment complications. Due to lack of a clear definition of comorbidity as a variable in this study, results are used only for discussion. Comorbidity calculations are also separate due to lack of reliable measure to adjust results by comorbidity, with CCI serving in the dual pre/post-treatment complication roles.

The main independent (cause) variable in this study was utilization of the order sets, forming two comparison groups. Health outcomes were defined as dependent (effect) variables. Race, age, and sex were mediating variables analyzed in conjunction with order set utilization to determine combined significance in predicting health outcomes. De-identified patient encounters were obtained via queries against the enterprise data warehouse containing records from CPOE and patient accounting applications. In order to increase size of the "order set" comparison group, outpatient CHF admission orders were pulled into the study, from those cases where patients visited Advocate provider on an outpatient basis, received diagnosis of CHF, and were ultimately admitted at one of the participating Advocate facilities. Under this scenario, the same inpatient hospital CHF sets were utilized for placing orders in advance of patient arrival, so there is not a mix-up of sets introduced to the study via inclusion of outpatients, who ultimately ended up being admitted. Majority of the patients were admitted to the hospital via emergency department or urgent care center. Data was manipulated in Excel to adhere to the following guidelines:

- Adults over the age of 18
- Patients with primary or secondary diagnosis of CHF
- Psych and OB patients excluded

CHF patients were assigned into two comparison groups, accounting for the above inclusion/exclusion criteria. Data was subsequently loaded into SPSS statistical analysis software to conduct statistical analysis. Binary logistic regression with chi-squared option was employed to compare groups in measuring mortality and readmissions outcomes. The Mann-Whitney non-parametric U test of independent samples was utilized to test the null hypothesis for LOS, while One-Way ANOVA was employed to compare the mean LOS scores. The latter two tests were used

for measuring comorbidity. Fisher's Exact was applied as a secondary method of measuring statistical significance, due to small "order set" group sizes.

Low utilization of the order sets in hospitals (around 7% of all CHF orders) is a challenge faced in many CPOE studies. This limits the categories in which statistically significant results can be claimed. In current practice, utilization of order sets is optional, contributing to low percentage of orders placed via sets and the disproportionally large size of the "free text" group. The choice of a method to place CHF orders is in the hands of providers. In this study, the numbers for all participating hospitals needed to be combined in order to produce sufficient data samples. Towards end of the studied date range, one system-wide CHF set was produced and released to all hospitals. All sets were evidence based and approved by clinicians to be available in CPOE for the purposes of treating CHF, with initial content differences not believed to introduce serious concerns regarding validity of outcomes from the study. An example of the content of CHF order set currently utilized at the system level is presented in Table 1.

Table 1. Example of Advocate CHF order set

Component	Order Details
LET Orders	
Code Status	No CPR for Cardiac Arrest
Code Status	Full Code
Admission Orders	
Admit Order	Diagnosis : Congestive Heart Failure
Consults	
Consult Physician	Cardiologist: CHF
Care Coordination	
CHF Nurse Consult	Congestive Heart Failure
Cardiac Rehab Inpatient	Congestive Heart Failure
Spiritual Care Consult	Emotional Distress
ZZ-Hospice Care Consult	Congestive Heart Failure
Palliative Care – Consult	Reason for Consult: Other, Congestive Heart Failure
Nutrition Consult Adult	Assessment, Education. CHF Patient
Physical Therapy Adult Eval and Treatment	Congestive Heart Failure
CHF Clinic Consult	Congestive Heart Failure
Discharge Planning Eval Adult	Specify in Special Instructions, Days till discharge: Undetermined, Congestive Heart Failure
Continuous Infusions	
Saline Lock Insertion	If not already done, replace field starts
Diagnostic Imaging	
XR Chest 1V	Routine, Transport Mode: Portable/CHF/etc (multiple)
XR Chest PA, Lateral 2V	Routine, T+1;0600, Transport Mode: WHEEL CHAIR, Reason For Exam: CHF
Diagnostic Tests	
EKG 12 Lead Adult	Routine: Heart Failure
Cardiac Imaging	
Echocardiogram – Adult	Routine: CHF
CD Echo 2D Complete W DOP and Color – Adult	Routine: Transport Mode: Cart; Reason for Exam: CHF
Laboratory	
Basic Metabolic Panel [BPNL]	Next Draw/Specimen Collection, T;N (DEF)* STAT, T;N
Comprehensive Metabolic Panel [CPNL]	Next Draw/Specimen Collection, T;N (DEF)* STAT, T;N
Magnesium Level [MG]	Next Draw/Specimen Collection, T;N (DEF)* STAT, T;N
Prothrombin Time [PTINR]	Next Draw/Specimen Collection, T;N (DEF)* STAT, T;N
CK [CPK]	Next Draw/Specimen Collection, T;N (DEF)* STAT, T;N

CK-MB [MMB]	Next Draw/Specimen Collection, T;N (DEF)* STAT, T;N
Troponin I Ultrasensitive [TROPI]	Next Draw/Specimen Collection, T;N (DEF)* STAT, T;N
B Type Natriuretic Peptide [BNPEP]	Next Draw/Specimen Collection, T;N (DEF)* STAT, T;N
Glucose - Fingerstick Bedside	T;N, One Time (unscheduled), Call for BS Less Than ____ Greater Than _____. (DEF)*T;N, QID [before meals & HS], Call for BS Less Than ____ Greater Than _____.
CBC with Automated Differential [CBCA]	Next Draw/Specimen Collection, T;N (DEF)* STAT, T;N
Urinalysis Complete W C/S if IND [UCOMCS]	Next Draw/Specimen Collection, T;N, Urine, Midstream
Digoxin Level [DIG]	Tomorrow AM Draw [4 to 6 AM], T+1;0600 (DEF)* Next Draw/Specimen Collection, T;N
Thyroid Stimulating Hormone W Reflex [TSHR]	Tomorrow AM Draw [4 to 6 AM], T+1;0600 (DEF)* Next Draw/Specimen Collection, T;N
Uric Acid Level Blood [URIC]	Tomorrow AM Draw [4 to 6 AM], T+1;0600 (DEF)* Next Draw/Specimen Collection, T;N
Lipid Panel W/O Reflex [LIPDPL]	Tomorrow AM Draw [4 to 6 AM], T+1;0600 (DEF)* Next Draw/Specimen Collection, T;N
Ferritin Level [FERR]	Tomorrow AM Draw [4 to 6 AM], T+1;0600 (DEF)* Next Draw/Specimen Collection, T;N
Medications	
Analgesics	
Tylenol	650 mg, Oral, Q4hr, PRN pain, Tab
No NSAIDs	
VTE Prophylaxis	
Heparin	Pharmacist to Dose (consult)
+120 Minutes Lovenox	Pharmacist to Dose (consult)
Heparin	5,000 unit, Subcutaneous, Q12hr, Injection
Lovenox	40 mg, Subcutaneous, Daily, Injection (DEF)*
	30 mg, Subcutaneous, Q12hr, Injection
Renal Dosing	(CrCl less than 30 mL/min)
Lovenox	30 mg, Subcutaneous, Daily, Injection
Antiembolytic Hose	T;N, Bilateral Thigh Length (DEF)*
	T;N, Bilateral Knee Length
Lipid Management	
Zocor	Multiple dose choices
Lipitor	Multiple dose choices
Pravachol	Multiple dose choices
Lovastatin	Multiple dose choices
NO ACE Inhibitors or ARBs	
ACE Contraindications	T;N, ACE Contraindications : Hypotension (DEF)*
	T;N, ACE Contraindications : Acutely Worsening Renal Insufficiency
ARB Contraindications	T;N, Hypotension (DEF)*
	T;N, Acutely Worsening Renal Insufficiency
Angiotensin-Converting Enzyme Inhibitors	
ACE Contraindications	T;N, ACE Contraindications : Acutely Worsening Renal Insufficiency (DEF)*
	T;N, ACE Contraindications : Cough
NO ACE Inhibitors	
Vasotec	Multiple dose choices
Zestril (Prinivil)	Multiple dose choices

Altace	Multiple dose choices
Angiotensin Receptor Blockers	
ARB Contraindications	T;N, Acutely Worsening Renal Insufficiency (DEF)*
	T;N, Hypotension
NO ARBs (Angiotensin Receptor Blockers)	
Diovan	Multiple dose choices
Avapro	Multiple dose choices
Cozaar	Multiple dose choices
Atacand	Multiple dose choices
Aldosterone Antagonists	
Aldactone	Multiple dose choices
Beta Blockers	
Lopressor (tartrate)	Multiple dose choices
Toprol XL (succinate)	Multiple dose choices
Bystolic	Multiple dose choices
Coreg	Multiple dose choices
Zebeta	Multiple dose choices
Ancillary Medications	
Lanoxin	Multiple dose choices
Diuretics	
furosemide (Lasix)	Multiple dose choices
Bumetanide (Bumex)	Multiple dose choices
Demadex	Multiple dose choices
Potassium Supplements	
Potassium Chloride	Multiple dose choices
Magnesium Supplements	
Magnesium Oxide	Multiple dose choices
Magnesium Sulfate	
Nitrates	
Imdur	Multiple dose choices
Medication Nitroglycerin PowerPlan	
Vasodilators	
Hydralazine	Multiple dose choices
Hydralazine-isosorbide dinitrate (BiDil)	1 tab, Oral, Q6hr, Tab
isosorbide dinitrate	20 mg, Oral, TID, Tab
Nursing Orders	
Monitoring	
Weigh Patient	T;N, Daily, Use same scale; weigh before breakfast
Vital Signs per Unit Routine	T;N, With Pulse Oximetry
Intake and Output	T;N, Q Shift (8 hr) (DEF)* Comments: Notify MD for urine output less than ____mL after first diuretic dose and/or urine output less than ___ml/hr.
	Comments: Notify MD for urine output less than ____mL after first diuretic dose and/or Urine Output less than ___ml/hr.
Cardiac Monitoring	T;N, Call for: Comments: Re-evaluate daily for continued need for telemetry
General	
Activity Patient	Multiple choices: Up ad Lib, Out of Bed in Cardiac Chair, Complete Bed Rest, Advance Activity to Baseline, Walk in halls 2x daily, Up in chair with all meals
Misc Nursing Orders	
Nursing Communication Order	Evaluate patient for sleep apnea
External Cardiac Defibrillator	
Nutrition	

Cardiac Diet	Multiple fluid restriction choices
Sodium	Multiple dose choices
Diabetic Calorie Diet	Multiple fluid restriction choices
Renal Diet	Multiple fluid restriction choices
Clear Liquid Diet	Multiple fluid restriction choices
NPO	
Respiratory Therapy	
Respiratory Oxygen PowerPlan	
Respiratory BiPAP/CPAP PowerPlan	

Results

1. ***Mortality***.

The goal of this study was to determine, quantitatively, whether utilization of CHF order sets leads to lower inpatient mortality. All tests were conducted with an assumption of null hypothesis and aimed at testing whether it could be rejected. The binary logistic regression method revealed that 1.8% of patients in the "order set" group died versus 3.2% in the "free text" group, OR = 1.818, 95% CF 1.039 – 3.181, $\chi^2 = 4.516$ ($p = 0.034$). Results were statistically significant. Population N of the "order set" group was 719, while N of the "free text" group was 10219. Due to small size of the "order set" group, a 2-tailed Fisher's Exact was calculated as an alternative to Pearson chi-squared, with statistically significant outcome of 0.04. Patients in the "order set" group, whose medications were ordered via pre-defined sets, had a nearly doubled chance of survival compared to patients in the "free text" group. The null hypothesis for CHF mortality was rejected. Mortality outcomes are summarized in Table 2.

Table 2. CHF mortality as a total for all Advocate hospitals participating in the order sets study

Outcomes and Measures	**Free Text Group**	**Order Set Group**	**Pearson Chi-Squared (χ^2)**	**2-sided Fisher's Exact**	**1-sided Fisher's Exact**	**EXP (B) – Binary Logistic Regression**
Mortality = yes	331	13				
Mortality = no	9888	706				
Percent (Mortality=yes) / total	3.2%	**1.8%**	4.516 (p= 0.034)	0.040	0.022	1.818 [1.039 & 3.181]

2. ***Readmissions.***

Another goal of this study was to determine whether placing CHF orders via sets helped reduce the 30-day hospital readmission rate. As in the mortality examination, statistical manipulations were performed to test the null hypothesis. There was insufficient information on readmissions available in the enterprise data warehouse for this study, and results were not statistically significant. Population N of the "order set" group was 538, while population of the "free text" group was 7583. Order sets made no difference on patients' chances to avoid hospital readmission within 30 days of being discharged with CHF diagnosis. Readmissions outcomes are summarized in Table 3.

Table 3. Readmissions among CHF patients as a total for all Advocate hospitals participating in the order sets study

Outcomes and Measures	Free Text Group	Order Set Group	Pearson Chi-Squared (χ^2)	2-sided Fisher's Exact	1-sided Fisher's Exact	EXP (B) – Binary Logistic Regression
Readmission = yes	1415	108				
Readmission = no	6168	430				
Percent (Readmission =yes) / total	**19%**	20%	0.659 (p= 0.417)	0.424	0.224	0.913 [0.734 & 1.137]

3. ***Length of Stay*.**

Two statistical tests, One-Way ANOVA and Mann-Whitney binary test of independent samples, were employed to address whether application of order sets in the clinical settings can help reduce the length of hospital stay. The mean length of stay among the "order set" group patients was 4.75 days vs. 5.46 days for patients in the "free text" group, indicating that CHF patients who received orders via sets stayed in the hospital almost a day less compared to patients whose orders were placed manually. Population N of the "free text" group was 10219 and 719 in the "order set" group. The One-Way ANOVA test was significant, F(1,10938) = 8.352, p=0.004. The Mann-Whitney test of independent samples for binary values confirmed rejection of the null hypothesis: [U(1,N=10937) = 3,472,652; p = 0.013], indicating the benefit of shorter stay among the "order set" group patients.

Expired patients could have contributed to shorter LOS. In order to adjust for mortality, the same calculations were repeated with expired patients excluded. Results remained consistent, showing shorter LOS for patients in the "order set" group, although the gap between comparison groups has narrowed to still significant 0.63 days compared to nearly a day for all patients: mean 4.76 ("order set") vs. 5.39 ("free text"), with N = 706 and 9888 within these respective groups. The One-Way ANOVA test was significant, F(1,10593) = 7.131, p=0.008. The Mann-Whitney test of independent samples for binary values confirmed rejection of the null hypothesis: [U(1,N=10594) = 3,301,394.500; p = 0.015].

4. ***Comorbidities/Complications***.

The mean CCI score among the "order set" group patients was 3.64 vs. 3.68 within the "free text" group. CCI is a total score attached to each patient encounter in history (where available) that is computed by adding individual weights of complications found in a patient. This score does not indicate whether a patient arrived with some of these conditions or acquired them during hospital stay referenced in the encounter. Population N of the "free text" group was 7232 and 525 in the "order set" group. Per One-Way ANOVA test, results were not significant: F(1,7757) = 0.298, p=0.585. The Mann-Whitney test of independent samples for binary values also failed to reject the null hypothesis, as follows: U(1,N=7757) = 1,840,535.5; p = 0.23. Utilization of CHF order sets did not make a difference on comorbidity rates for CHF patients, and/or pre-existing conditions did not influence health outcomes measured in this study.

Discussion

The process of patient history analysis does not allow utilizing traditional attributes of controlled or randomized clinical trials, such as variable control or manipulation of the order set content. Additionally, most EMR/CPOE applications are not designed with health outcomes research in mind, making it difficult to track patient encounters for search of patterns and causal relationships on an aggregate basis. However, multi-year analysis of large patient history data sets offers a unique opportunity to explore impact of healthcare IT innovations and major CPOE application attributes such as order sets in a community healthcare setting where these tools/interventions are utilized for routine treatments, well beyond a small experimental study with a small number of patients.

The mortality study showed that patients whose medications were ordered using custom selection methods had a nearly doubled chance of death compared to patients who received orders via CPOE sets. Lack of ability to track all causal relationships within patient history from admission to discharge means that other factors possibly influenced mortality outcomes beyond standardization of ordering practices grounded in evidence-based medicine. However, comparison of the "order set" and "free text" groups and statistical significance of the mortality outcome point to the fact that CHF ordering via sets has potentially strong influence on this health outcome. Since results of the comorbidity study that utilized CCI score did not reveal statistical significance, it is reasonable to assume that pre-existing conditions did not affect mortality outcome, although the dual meaning of CCI score could also indicate that ordering via sets did not produce the desired improvement in reduction of post-treatment complications.

The study did not establish statistical link between utilization of CHF order sets and 30-day readmissions. Yet, the length of hospital stay was almost one day shorter for patients in the "order set" group, indicating wide implications of the study for the cost cutting and patient satisfaction improvement efforts – without a corresponding reduction in mortality. Even after mortality adjustment, the difference of 0.63 days for the non-expired patients represents significant outcome to consider. Regression analysis of factors influencing mortality also revealed that order sets play a larger role in predicting the outcome compared to basic demographic factors such as age, sex, and race of the patients. Despite limitations of the study in the areas of factor control, tracking all causal relationships during CHF treatments, and small size of the "order set" group due to low utilization of CHF order sets, results that were statistically significant show potentially strong correlation between evidence-based CPOE ordering practices and CHF health outcomes.

Utilization of the order sets to treat CHF is only one of many factors in the overarching variability of care reduction process aimed at improving quality and reducing cost of treatment. Placing orders in CPOE should not be considered the sole strategy for achieving these outcomes. Other variability reduction methods could include such initiatives as shared governance among clinicians to target specific quality improvement efforts, selection of the most effective evidence-based techniques applicable to treating patient conditions within the context of local culture and clinical settings, making shared decisions on selection and maintenance of the order sets, among other undertakings. Indeed, lack of governance around sets is likely to cause low buy-in, lower core measures compliance, and potentially dangerous side effects from improper bundling of medications and other orders.

Conclusion

In reporting improved mortality and length of stay outcomes, this study contributes to the literature supporting the use of evidence-based prescribing practices in the treatment of CHF patients. These outcomes could lead to patient safety and cost efficiency gains, which, even without reduced readmission rate benefit, are significant for hospitals when reviewing their evidence-based practices, CPOE utilization guidelines, and order set governance efforts. Extensive communication and shared decision-making, resulting in high physician buy-in, are paramount in a clearly defined strategy to reduce variability of care and to enable healthcare facilities to take control over their

prescribing practices. The study also encourages further study leading to more granular approaches to investigation of order set management practices and their effect on patient outcomes.

References

1. Ahmad A, Teater P, Bentley TD, Kuehn L, Kumar RR, Thomas A, et al. Key attributes of a successful physician order entry system implementation in a multi-hospital environment. J Am Med Inform Assoc 2002 Jan-Feb;9(1):16-24.
2. Payne TH, Hoey PJ, Nichol P, Lovis C. Preparation and use of preconstructed orders, order sets, and order menus in a computerized provider order entry system. J Am Med Inform Assoc 2003 Jul-Aug;10(4):322-329.
3. Cowden D, Barbacioru C, Kahwash E, Saltz J. Order sets utilization in a clinical order entry system. AMIA Annu Symp Proc 2003:819.
4. Dixon BE, Zafar MA. Inpatient Computerized Provider Order Entry (CPOE). Agency for Healthcare Research and Quality [Internet]. 2009 [cited 2014 Jul 5]. Available from: http://healthit.ahrq.gov/ahrq-funded-projects/emerging-lessons/computerized-provider-order-entry-inpatient/inpatient-computerized-provider-order-entry-cpoe
5. Wright A, Sittig DF, Carpenter JD, Krall MA, Pang JE, Middleton B. Order sets in computerized physician order entry systems: an analysis of seven sites. AMIA Annu Symp Proc 2010 Nov 13;2010:892-896.
6. McAlearney AS, Chisolm DJ, Schweikhart S, Medow MA, Kelleher K. The story behind the story: physician skepticism about relying on clinical information technologies to reduce medical errors. Int J Med Inform 2007 Nov-Dec;76(11-12):836-842.
7. Cabana MD, Rand CS, Powe NR, Wu AW, Wilson MH, Abboud PC, et al. Why don't physicians follow clinical practice guidelines? A framework for improvement. JAMA 1999;282(15):1458-1465.
8. Formea CM, Picha AF, Griffin MG, Schaller JA, Lee MR. Enhancing participant safety through electronically generated medication order sets in a clinical research environment: a medical informatics initiative. Clin Transl Sci 2010 Dec;3(6):312-315.
9. Bekmezian A, Chung PJ, Yazdani S. Standardized admission order set improves perceived quality of pediatric inpatient care. J Hosp Med 2009 Feb;4(2):90-96.
10. Reingold S, Kulstad E. Impact of human factor design on the use of order sets in the treatment of congestive heart failure. Acad Emerg Med 2007;14(11):1097-1105.
11. Ballard DJ, Ogola G, Fleming NS, Stauffer BD, Leonard BM, Khetan R, et al. Impact of a standardized heart failure order set on mortality, readmission, and quality and costs of care. Int J Qual Health Care 2010 Dec;22(6):437-444.
12. Abrahamyan L, Austin PC, Donovan LR, Tu JV. Standard admission orders can improve the management of acute myocardial infarction. Int J Qual Health Care 2012 Aug;24(4):425-432.
13. Fleming NS, Ogola G, Ballard DJ. Implementing a standardized order set for community-acquired pneumonia: impact on mortality and cost. Jt Comm J Qual Patient Saf 2009 Aug;35(8):414-421.
14. Ballard DJ, Ogola G, Fleming NS, Heck D, Gunderson J, Mehta R, et al. The Impact of Standardized Order Sets on Quality and Financial Outcomes. In: Henriksen K, Battles JB, Keyes MA, Grady ML, editors. Advances in Patient Safety: New Directions and Alternative Approaches (Vol. 2: Culture and Redesign) Rockville (MD); 2008.
15. Kohler MJ, Amezaga M, Drozd J, Crowley ST, Gulanski B, Anderson DR, et al. Use of a computerized order set to increase prescription of calcium and vitamin D supplementation in patients receiving glucocorticoids. Journal of general internal medicine 2013;28(6):825-829.
16. Micek ST, Roubinian N, Heuring T, Bode M, Williams J, Harrison C, et al. Before-after study of a standardized hospital order set for the management of septic shock. Crit Care Med 2006 Nov;34(11):2707-2713.
17. Thiel SW, Asghar MF, Micek ST, Reichley RM, Doherty JA, Kollef MH. Hospital-wide impact of a standardized order set for the management of bacteremic severe sepsis. Crit Care Med 2009 Mar;37(3):819-824.

Clinical Decision Support-based Quality Measurement (CDS-QM) Framework: Prototype Implementation, Evaluation, and Future Directions

Polina V Kukhareva, MPH[1], Kensaku Kawamoto, MD, PhD[1], David E Shields[1], Darryl T Barfuss[2], Anne M Halley[2], Tyler J Tippetts[1], Phillip B Warner, MS[1], Bruce E Bray, MD[1], Catherine J Staes, MPH, PhD[1]

[1]Department of Biomedical Informatics, University of Utah, and [2]Quality Improvement and Safety Department, University of Utah Healthcare, Salt Lake City, UT

Abstract

Electronic quality measurement (QM) and clinical decision support (CDS) are closely related but are typically implemented independently, resulting in significant duplication of effort. While it seems intuitive that technical approaches could be re-used across these two related use cases, such reuse is seldom reported in the literature, especially for standards-based approaches. Therefore, we evaluated the feasibility of using a standards-based CDS framework aligned with anticipated EHR certification criteria to implement electronic QM. The CDS-QM framework was used to automate a complex national quality measure (SCIP-VTE-2) at an academic healthcare system which had previously relied on time-consuming manual chart abstractions. Compared with 305 manually-reviewed reference cases, the recall of automated measurement was 100%. The precision was 96.3% (CI:92.6%-98.5%) for ascertaining the denominator and 96.2% (CI:92.3%-98.4%) for the numerator. We therefore validated that a standards-based CDS-QM framework can successfully enable automated QM, and we identified benefits and challenges with this approach.

Introduction

Overview of Clinical Quality Measurement (QM). Delivering quality healthcare is challenging due to ongoing and ubiquitous variation in health system processes that may lead to errors.[1] Measuring and reducing variation from evidence-based clinical best practices have been shown to improve quality and decrease costs of healthcare.[2] Despite multiple efforts undertaken to improve healthcare quality since the publication of the Institute of Medicine's reports entitled 'To Err Is Human'[3] and 'Crossing the Quality Chasm'[4], the quality of healthcare in the United States continues to be compromised by unnecessary variation in the implementation of clinical practice guidelines. Having a means to assess healthcare quality is essential for identifying deviations from evidence-based best practices and mitigating preventable errors.[3] Clinical quality measures are measures of processes, experience, and/or outcomes of patient care. There are increasing mandates and financial incentives to use electronic health records (EHRs) to measure quality as opposed to employing traditional manual processes for QM.[3,4] For example, the Meaningful Use (MU) recommendations issued by the federal Health Information Technology Policy Committee (HITPC) in November 2012 require the implementation of QM as well as decision support for high-priority conditions, and the use of related standards.[4,5] The practice and value of quality measurement has evolved over time. According to Meyer *et al.*, "in the last half century, the US has gone from defining quality, to measuring quality, to requiring providers to publicly report quality measures, and most recently, beginning to hold providers accountable for those results".[6] The National Quality Forum (NQF) was created as a public-private partnership to guide decisions regarding quality measure selection.[7] Until recently, quality measurement has relied mainly on the use of electronic claims data, manual chart abstraction, and patient surveys.[8] Currently, QM is required by public and private payers, regulators, accreditors and others that certify performance levels for consumers, patients and payers.[6] Current quality measurement systems in many hospitals include time-consuming manual paper and electronic record abstraction by a quality improvement specialist.[9,10]

At large academic medical centers such as University of Utah Health Care (UUHC), manual data abstraction is often followed by data analysis by an external organization such as the University HealthSystem Consortium.[11,12] University HealthSystem Consortium is an alliance of 120 academic medical centers and 299 of their affiliated hospitals representing academic medical centers with a focus on quality and safety excellence.[13,14] Manual chart abstraction at UUHC is performed by the Quality and Patient Safety Department, which has 28 employees, including 12 Quality Improvement Specialists.[15] There are several limitations with this process. For example, (a) three to six months may elapse between the time of a clinical procedure (e.g., a surgery) and the time when feedback

is given to a clinician; (b) human errors may be introduced during manual record review; and (c) only a subset of the clinical events is oftentimes selected for review, leading to gaps in quality assessment coverage.

Previous Work in Automating QM. One of the promises of implementing EHRs is the possibility for automatic generation of QM.[10] A MU-certified EHR must be able to export standardized quality reports, which can then "be fed into a calculation engine to compute various aggregate scores".[16] Following these recommendations, major EHR vendors such as Epic have started to integrate QM logic into their products.[17] Currently, however, only some EHR vendors offer quality measurements embedded in their system, and the scope of measures supported is not always comprehensive.[10,17] For example, KPHealth Connect has automated six of 13 Joint Commission measurement sets, and Epic has automated 44 NQF quality measures.[10,17] Vendor-based solutions may offer 'full sample' analysis, but the logic may be a 'black box.' Also, it is not always clear which version of each rule has been implemented or whether the quality measure logic is up-to-date. In addition, users may not have control over the logic to customize quality measurement. Even so, automated QM has the potential to provide quality reports on demand, may avoid human errors in manual abstraction, and can analyze 100% of patient encounters. Most ongoing efforts to produce automated quality measures are tied to a specific EHR system, and the executable logic for the quality measure is not sharable between different EHR systems.[10]

The Problem: Duplicative, Divergent Implementation of QM and CDS. It is intuitively obvious that CDS and QM are highly related, as QM focuses on who is eligible for a needed intervention (denominator identification) and who among them has received the needed intervention (numerator identification), whereas CDS focuses on who is eligible for a needed intervention and has not received the needed intervention (equivalent to numerator identification). However, to the best of our knowledge, there have been limited evaluation and validation in the literature of how technical approaches for one problem space can be re-used in the other, especially for standards-based approaches. This is important, because EHR certification criteria will likely drive much work in this field. It has been suggested that the two could be combined.[18] There have been efforts to combine CDS and QM logic, but the efforts were not standards-based and no conceptual framework was developed.[19,20]

Potential Solution: Leverage a Standards-based CDS Web Service across a Population for QM. Kawamoto *et al.* have previously suggested that a standards-based, service-oriented architecture could be used to make CDS logic sharable between different EHRs.[21] In this study, we hypothesized that this approach could be extended to encompass both CDS and QM given similarities in their functional requirements.

In pursuing this potential approach to CDS-based quality measurement (CDS-QM), a promising resource to leverage is an open-source, standards-based, service-oriented framework for CDS known as OpenCDS.[22] As shown in Figure 1, an EHR system can submit patient data to OpenCDS and obtain patient-specific assessments and recommendations that are provided to clinicians via alerts, reminders, or other CDS modalities.[23] OpenCDS is compliant with the HL7 Virtual Medical Record (vMR) and HL7 Decision Support Service (DSS) standards, and it leverages various open-source component resources, including the JBoss Drools knowledge management platform and Apelon Distributed Terminology System. Theoretically, then, OpenCDS could be used to measure quality as well as provide CDS. Moreover, the use of a CDS-based QM approach could potentially provide advantages for quality improvement compared to traditional approaches. Therefore, the objectives of this study were to: a) identify opportunities to enhance quality improvement using CDS-QM, b) design and implement a CDS-QM approach aligned with candidate CDS standards for Meaningful Use,[24] and c) evaluate the CDS-QM approach for a representative quality measure.

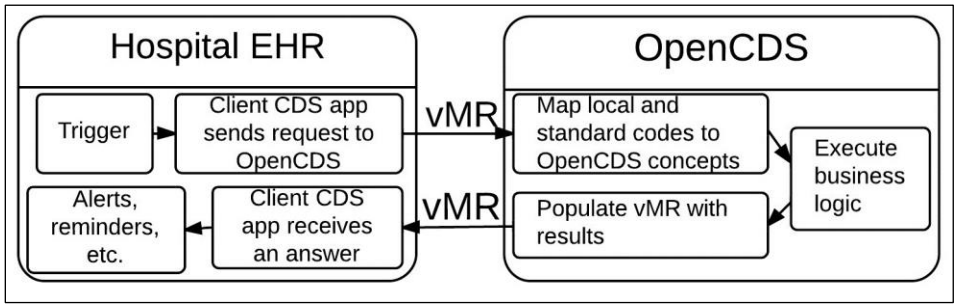


Figure 1. OpenCDS architecture: high-level interaction for CDS

Methods

Identification of opportunities to enhance quality improvement using CDS-QM

We engaged specialists from the Quality and Patient Safety Department to identify strengths and limitations of the CDS-QM approach. We documented the current process of quality assessment and reporting, and interviewed two of the 28 quality improvement specialists to identify possible ways in which the CDS-QM approach could enhance the institution's capabilities related to quality measurement and improvement.

Design and implementation of CDS-QM framework

Functional requirements

For the purposes of this implementation, our requirements were to enable the evaluation of national quality measures across the relevant patient populations in UUHC. The primary requirement was accurate evaluation of quality measure compliance.

Design principles

In designing the CDS-QM framework, a core design principle was *standards-based scalability*, so that the framework could potentially be leveraged in the context of various institutions and information systems. Related to this principle, a second principle was *open availability*, with open-source tooling used as to limit barriers to adoption related to licensing and intellectual property restrictions.

Scope and assumptions

Scope was limited to the analysis of structured data, as opposed to free text data requiring natural language processing. It was assumed that relevant patient data are available, such as in a data warehouse.

Tools and resources

In addition to OpenCDS, we leveraged the open-source Mirth Connect integration engine (v3.0.1). We also leveraged the UUHC data warehouse (DW), which contains data from the EHR systems and other ancillary clinical and administrative information systems at the institution.

Evaluation of CDS-QM approach for representative national quality measure

Quality measure

We chose the Joint Commission's Surgical Care Improvement Project (SCIP) Venous Thromboembolism 2 (SCIP-VTE-2) quality measure to evaluate the CDS-QM approach. This measure was chosen due to its technical complexity and its prioritization by the UUHC Quality and Patient Safety Department. VTE is a major cause of morbidity and mortality in hospitals.[25] In spite of evidence of their effectiveness, VTE prophylaxis by anti-coagulation and/or mechanical compression remains underutilized in US academic medical centers, particularly among surgical patients.[25] SCIP-VTE-2 is a well-established quality measure and is supported by level 1a evidence.[26,27] This measure is used to assess the percent of surgery patients that receive appropriate VTE prophylaxis within 24 hours of surgery.[28] For this evaluation, we implemented the SCIP-VTE-2 quality measure using the logic published for surgeries that occur in 2014 (version v4.2a).[28]

Data Sources

We used clinical data generated by inpatient surgeries that occurred at UUHC in 2013. A total of 8,924 cases were assessed using the CDS-QM automated method. The data elements required by the quality measure logic were documented in multiple source systems, including the inpatient EHR (Cerner) and two systems used to document anesthesia and nursing activities during the surgical event. The University of Utah Institutional Review Board performed an administrative review of this project and determined that IRB approval was not required because this effort was conducted for the purposes of quality improvement and does not meet the regulatory definition of human subject research.

Evaluation/Validation

As a reference standard for validation, we used quality measurement results produced by the University HealthSystem Consortium through an analysis of data extracted through manual chart abstraction by the Quality and Patient Safety Department. The sample used for validation was the 319 surgery cases randomly chosen for abstraction by the University HealthSystem Consortium for the first two quarters of 2013.

As the first step in our analysis, we evaluated the degree to which the data required for the evaluation was available in a structured format in the UUHC DW. Second, we compared the results from OpenCDS with the results from the reference standard approach (manual chart abstraction followed by University HealthSystem Consortium analysis). Observations were classified as true positive (TP), true negatives (TN), false positives (TP), and false negatives (FN) for denominator and numerator criteria separately. We calculated recall (sensitivity) of the OpenCDS-based process as the proportion of cases classified as positive by OpenCDS among the cases classified as positive by the reference standard (TP/(TP+FN)). We calculated precision (positive predictive value) of the OpenCDS-based process as the proportion of cases identified as positive by OpenCDS which was also classified as positive by the reference standard (TP/(TP+FP)). We assessed recall and precision for the classification of denominators, as well as recall and precision for the classification of numerators among cases that met denominator criteria. We also assessed the proportion of cases that yielded a complete match with the reference standard. Exact (Clopper-Pearson) confidence intervals (CI) were estimated for all binomial proportions. Statistical analysis was conducted using SAS version 9.3 (SAS Institute, Cary, North Carolina).

Results

Opportunities to enhance quality improvement using CDS-QM

As shown in Figure 2, the current quality assessment process at UUHC starts with data from the DW about clinical events (in this case major surgeries) being reported to the external quality benchmarking organization (the University HealthSystem Consortium), followed by this organization choosing a sample of surgery cases for manual chart abstraction. The UUHC quality specialists then perform manual abstraction of data from the EHR for the selected records. Finally, the UUHC specialist gives the information back to the external quality organization for summarization and reporting. The entire process from the time of surgery to quality reporting can take up to 6 months.

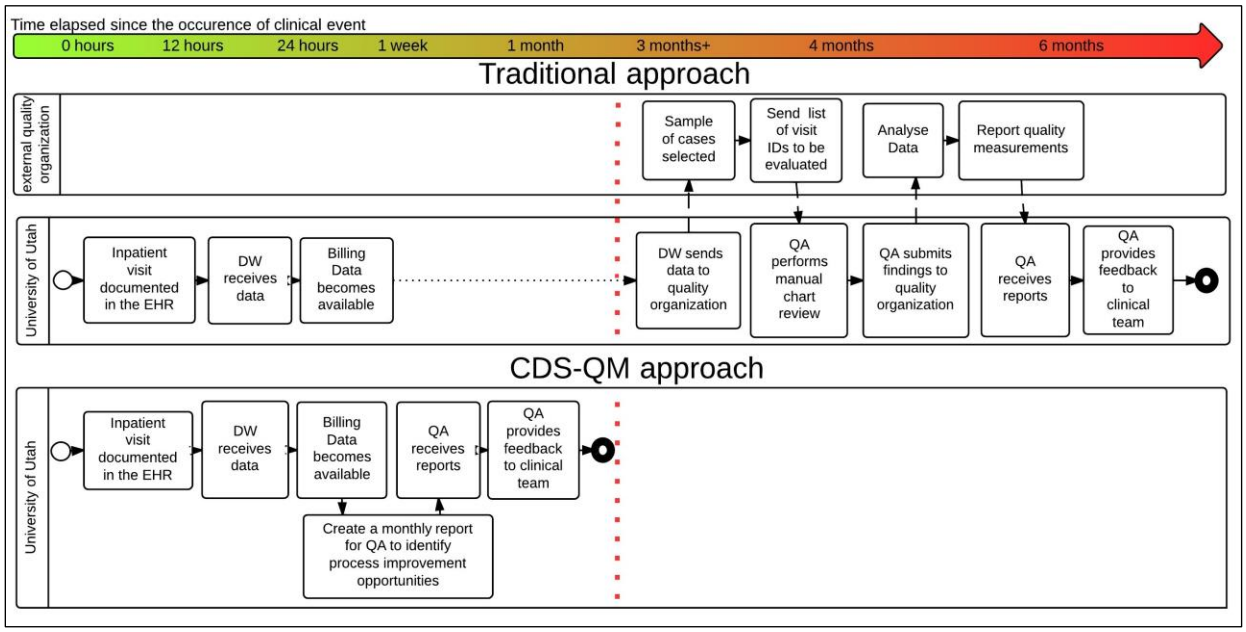


Figure 2. Comparison of traditional and CDS-QM approaches to quality measurement and improvement

Diagramming the traditional process for quality measurement and reporting was useful for identifying several opportunities using the CDS-QM approach to enhance workflow and impact clinical care. The informaticists and specialists with the Quality and Patient Safety Department identified several potential improvements. The automated approach using the CDS-QM strategy could:

- Improve the timeliness and completeness of feedback to the clinical stakeholders. The Quality Department holds monthly meetings with clinical stakeholders. Using the CDS-QM approach, a summary of the quality measure results from the previous month could be made available at these meetings to enable more rapid responses to identified quality deficiencies. The current reporting process allows such quality deficiencies to go unnoticed for potentially many months.
- Improve the completeness of assessment and feedback. While some records may still require manual review due to missing data (see Results), the vast majority of cases can be evaluated in an automated manner, as opposed to the baseline sampling approach. This more complete approach could potentially identify problem areas not yet sampled and therefore not yet identified by the quality team.
- Enable quality and clinical stakeholders to assess the impact of new rules or different versions of rules. This functionality is not available with the current process. At the same time, this functionality is critical for performing longitudinal analysis or for making predictions about future compliance.
- Provide additional useful information. A traditional quality measurement approach only identifies whether patients met denominator and/or numerator criteria. In contrast, the CDS-QM approach allows one to generate intermediate results that may be useful for better understanding the root cause underlying any deficiencies.

Design and implementation of CDS-QM framework

Figure 3 provides an overview of the CDS-QM approach. First, the Mirth Connect interface engine was used to identifying relevant surgery cases for analysis. Second, Mirth Connect was used to sequentially obtain relevant patient data from the DW. Third, the relevant patient data were converted into the HL7 vMR format used by OpenCDS. Then, Mirth Connect transmitted the vMR input data to OpenCDS using the SOAP Web service interface specified in the HL7 Decision Support Service (DSS) standard.

Within OpenCDS, the local and standard codes provided as the input were mapped to the internal concepts used by OpenCDS. The data were then evaluated by executing the relevant OpenCDS knowledge module, and the resulting patient-specific assessments were returned to Mirth Connect as vMR output objects. Finally, DW tables were populated with evaluation results by Mirth Connect.

Human workflows for developing the necessary business logic in Mirth Connect and OpenCDS were developed as well. These workflows involved identifying and characterizing the required input data, creating database queries in Mirth Connect, developing the required terminology mapping files, and developing the data processing algorithms in OpenCDS.

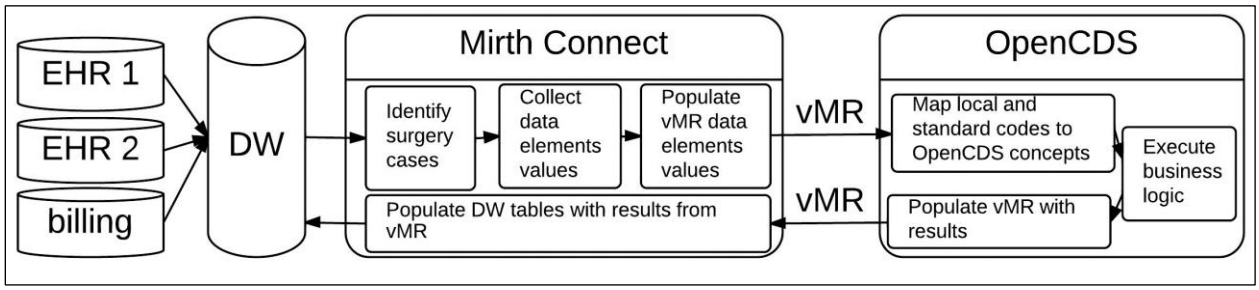


Figure 3. Major systems and processes involved in the CDS-QM approach

Evaluation of CDS-QM approach for representative national quality measure

The NQF-endorsed SCIP-VTE-2 quality measure was automated. Business logic was presented in the OpenCDS Web-based knowledge engineering platform, which uses the JBoss Guvnor platform. Figure 4 shows a sample screenshot from this knowledge engineering platform.

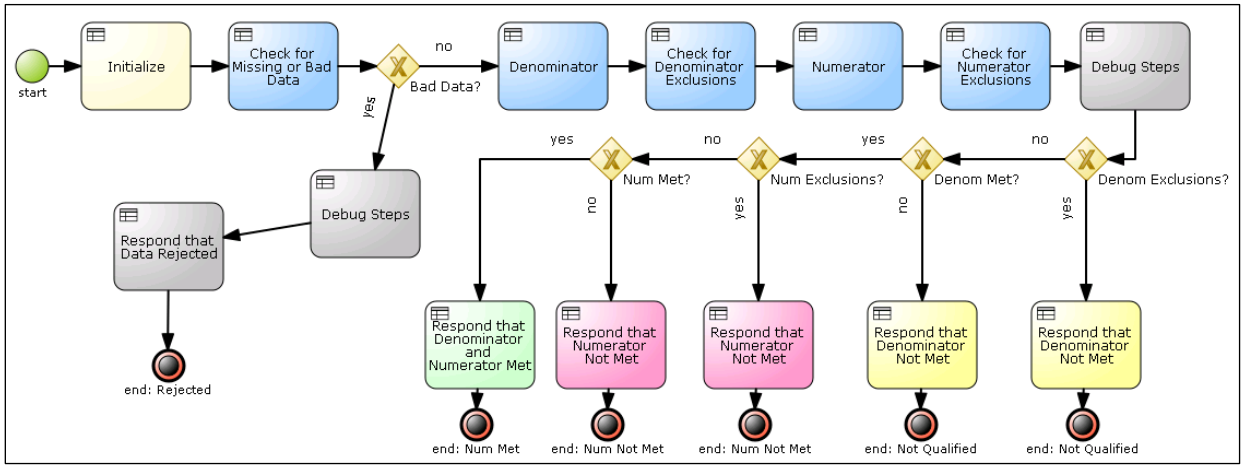


Figure 4. SCIP-VTE-2 Business Logic represented in Guvnor

Mirth queries were able to access all the necessary data elements except documentation about participation in clinical trials and the use of preadmission oral anticoagulant therapies. These two data elements were not recorded in the DW and would require manual review of the EHR text-based records for complete assessment. Among the 319 cases in the reference population selected by University HealthSystem Consortium, 14 did not have complete records stored in the DW and could not be used for automated QM. Completeness of the records was estimated as 95.6% (CI: 92.8%-97.6%). The remaining 305 cases were used to compare OpenCDS results with the reference standard.

When classifying denominators (i.e., identifying cases that should be included for quality measurement), the OpenCDS strategy yielded a recall of 100% (CI: 98%-100%) and precision of 96.3% (CI: 92.6%-98.5%) (see Table 1). A total of 183 cases among the 305 cases selected for review were found to meet the inclusion criteria. The seven cases included in the denominator using OpenCDS but excluded by the reference standard were cases with preadmission oral anticoagulant therapy. These cases were identified by the manual review process but not found by Mirth because the data is not currently saved in the DW.

Table 1. Comparison of the denominator (inclusion/exclusion) classification, using two methods (n=305)

CDS-QM approach (based on automated review using 2014 logic)	**Reference Standard (based on manual review using 2013 logic)**	
	include	exclude
include	TP=183	FP=7*
exclude	FN=0	TN=115

* had preadmission oral anticoagulants

Similarly, when classifying numerators (i.e., identifying cases that failed the quality measure among the 183 cases that met the denominator criteria), the OpenCDS strategy yielded a recall of 100% (CI: 97.9%-100%) and a precision of 96.2% (CI: 92.3%-98.4%) (see Table 2). The seven cases that passed following evaluation of the 2014 version of the SCIP-VTE-2 logic were cases that, in fact, failed according to the 2013 SCIP-VTE-2 specification version of the logic used for the manual review. These seven cases illustrate VTE-prophylaxis practices that were previously considered insufficient, but will be considered sufficient using the 2014 criteria. Thus, a complete match was found for 291 out of 319 records (91.2%; CI: 87.6%-94.1%).

Table 2. Comparison of the numerator (passed/failed) criteria, using two methods (n=183)

CDS-QM approach (based on automated review using 2014 logic)	**Reference Standard** (based on manual review using 2013 logic)	
	passed	failed
passed	TP=176	FP=7*
failed	FN=0	TN=0

* passed using 2014 logic

Discussion

We successfully prototyped an implementation of a CDS-QM-based system and demonstrated its feasibility. This use case demonstrates that, once the quality and CDS standards are fully aligned, the opportunity for meeting both goals using the same approach is quite feasible. We also identified special considerations for QM, such as the need to efficiently obtain and process data for a large cohort of patients, even while evaluation is conducted on a patient-by-patient basis.

As we engaged in the effort with the experts form the Quality and Patient Safety Department, we identified many opportunities to enhance quality improvement using CDS-QM. The currently used and proposed systems could complement one another. A CDS-QM system has the potential to be a cheaper and more efficient alternative to the analysis of quality measures relying on manual chart abstractions. However, it has a high initial cost for translating rules into executable format. Automation of the process of translating the measures into an executable format is a next logical step. Once translated into executable format, the logic can be shared with other institutions, and the logic can be modified to meet clinical needs to assess alternative or future quality measure specifications. The quality experts were particularly interested in the opportunity to apply the new SCIP VTE-2 logic specifications for 2014 to data generated from clinical practice occurring in 2013, which revealed that the new logic would reclassify their previous 'failures' as passing the new quality criteria. If the new logic had the opposite effect, it would be extremely helpful for a quality program to be able to anticipate 'failures' before they get reported six month later, as would occur using the traditional approach of manual abstraction and review by an external quality organization.

The CDS-QM approach is aligned with candidate CDS standards for Meaningful Use Stage 3. Standards proposed for 2015 voluntary EHR specification criteria were used[24]. We were able to use OpenCDS to implement quality measure logic published by the National Quality Forum, and we generated results that were either an exact match or could be explained where differences were observed. Many institutions use SQL queries instead of a CDS-QM approach because it requires less initial effort. However, maintaining an external rules repository is easier than when rules include a direct reference to the EHR or DW data, as there are no dependencies on the local database structure (i.e., the 'curly braces problem').

To maximize benefits from the use of a CDS-QM approach, high-quality structured data are necessary. Validation of the SCIP VTE-2 quality measure implementation using the CDS-QM approach highlighted gaps in documentation and problems with the transfer of information from source systems and the UUHC DW. These findings have been

shared with the DW and EHR teams, and the feedback is being used to improve processes and data being extracted into a surgery data mart. These feedback loops are important for improving data completeness and concordance. For data that was available in the DW, our automated approach shows favorable results compare to other automated approaches. For example, Kern *et al.* report that recall of electronic reporting ranged from 46% to 98% per measure, and precision from 57% to 97%.[29]

The CDS-QM approach may have limitations. Quality measures are often dependent on claims data, which are not usually available in real-time. More analysis is required to evaluate the timeliness of the availability of all the data required to implement quality measure logic. In addition, this study was performed in only one setting and focused on only one rule which may limit the generalizability of the results. However, OpenCDS uses a standard HL7 data model (the Virtual Medical Record), which potentially would allow the quality measurement rules to be implemented across systems. Additional research is necessary to demonstrate the CDC-QM approach in other EHRs and for other institutions.

In the future, we will automate more quality measures and provide additional feedback to the Quality and Patient Safety Department at UUHC, and the output metrics will be incorporated into dashboards that assess the cost and quality of care at UUHC. We are also working on modifying the OpenCDS infrastructure to support population-based queries. Instead of building one vMR at time, we are in the process of assembling thousands of vMRs at the same time. This approach to building vMRs should enable our approach to scale to quality measurement involving much larger patient samples, such as health maintenance measures for the general outpatient population.

Conclusion

In this study, we presented a prototype of the CDS-QM approach which can add value to the traditional quality reporting approaches. The CDS-QM approach allows for full case coverage, prospective use, near real-time evaluation, and is based on standards. The benefits of using CDS-QM include sampling a higher proportion of data, avoiding human error, saving abstractor time, providing more control over measurement rules, and independence from the EHR or other data source systems. To the best of our knowledge, this is the first study illustrating a framework and an approach for using an open-source, system-agnostic, standards-based CDS tool for continuous quality measurement.

Acknowledgements

This study was funded by University of Utah Health Care. The funding source played no role in the design and conduct of the study. KK is currently or recently served as a consultant on CDS to the Office of the National Coordinator for Health IT, ARUP Laboratories, McKesson InterQual, ESAC, Inc., Inflexxion, Inc., Intelligent Automation, Inc., Partners HealthCare, and the RAND Corporation. KK receives royalties for a Duke University-owned CDS technology for infectious disease management known as CustomID that he helped develop. KK was formerly a consultant for Religent, Inc. and a co-owner and consultant for Clinica Software, Inc., both of which provide commercial CDS services, including through use of a CDS technology known as SEBASTIAN that KK developed. KK no longer has a financial relationship with either Religent or Clinica Software. KK has no competing interests related to OpenCDS, which is freely available to the community as an open-source resource. All other authors declare no conflict of interest.

References

1. Landrigan CP, Parry GJ, Bones CB, Hackbarth AD, Goldmann DA, Sharek PJ. Temporal trends in rates of patient harm resulting from medical care. *N Engl J Med [Internet]*. 2010 Nov 25 [cited 2014 Feb 6];363(22):2124–34. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21105794
2. James BC, Savitz LA. How Intermountain trimmed health care costs through robust quality improvement efforts. *Health Aff (Millwood) [Internet]*. 2011 Jul [cited 2014 Jan 31];30(6):1185–91. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21596758
3. Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. *N Engl J Med [Internet]*. 2010 Aug 5 [cited 2014 Feb 6];363(6):501–4. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20647183
4. Blumenthal D. Launching HITECH. *N Engl J Med [Internet]*. 2010 Feb 4 [cited 2014 Feb 6];362(5):382–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20042745

5. Joseph AM. The stage 3 meaningful use preliminary recommendations: concerns are being raised. *MLO Med Lab Obs [Internet]*. 2013 Jul [cited 2014 Feb 20];45(7):64. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24180080
6. Meyer GS, Nelson EC, Pryor DB, James B, Swensen SJ, Kaplan GS, et al. More quality measures versus measuring what matters: a call for balance and parsimony. *BMJ Qual Saf [Internet]*. 2012 Nov [cited 2014 Feb 13];21(11):964–8. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3594932&tool=pmcentrez&rendertype=abstract
7. Panzer RJ, Gitomer RS, Greene WH, Webster PR, Landry KR, Riccobono CA. Increasing demands for quality measurement. *JAMA [Internet]*. 2013 Nov 13 [cited 2014 Feb 12];310(18):1971–80. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24219953
8. Anderson KM, Marsh CA, Flemming AC, Isenstein H RJ. Quality Measurement Enabled By Health IT: Overview, Challenges, And Possibilities: An Environmental Snapshot [Internet]. 2012 [cited 2014 Feb 6]. Available from: http://healthit.ahrq.gov/sites/default/files/docs/page/final-hit-enabled-quality-measurement-snapshot.pdf
9. Silow-Carroll S, Edwards JN, Rodin D. Using electronic health records to improve quality and efficiency: the experiences of leading hospitals. *Issue Brief (Commonw Fund) [Internet]*. 2012 Jul [cited 2014 Feb 6];17:1–40. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22826903
10. Garrido T, Kumar S, Lekas J, Lindberg M, Kadiyala D, Whippy A, et al. e-Measures: insight into the challenges and opportunities of automating publicly reported quality measures. *J Am Med Inform Assoc [Internet]*. 2014 Jan 1 [cited 2014 Jan 24];21(1):181–4. Available from: http://jamia.bmj.com/content/21/1/181.long
11. University HealthSystem Consortium (UHC) [Internet]. [cited 2014 Jan 30];Available from: https://www.uhc.edu/
12. University of Utah Health Care - Salt Lake City, Utah [Internet]. [cited 2014 Jan 30];Available from: http://healthcare.utah.edu/about/
13. Xu R, Polk RE, Stencel L, Lowe DK, Guharoy R, Duggal RW, et al. Antibiogram compliance in University HealthSystem Consortium participating hospitals with Clinical and Laboratory Standards Institute guidelines. *Am J Health Syst Pharm [Internet]*. 2012 Apr 1 [cited 2014 Jan 29];69(7):598–606. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22441793
14. Together we stand, collaborating we excel: partnering to target common goals. A report on and abstracts of the UHC (University HealthSystem Consortium) 2009 Quality and Safety Fall Forum. October 1-2, 2009. Atlanta, Georgia, USA. *Am J Med Qual [Internet]*. [cited 2014 Jan 29];25(2 Suppl):20S–36S. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20220202
15. The University of Utah Basic Search - Campus Directory [Internet]. [cited 2014 Jan 30];Available from: http://people.utah.edu/uWho/basic.hml?did=2012
16. Dolin RH. Accuracy of electronically reported "meaningful use" clinical quality measures. *Ann Intern Med [Internet]*. 2013 Jul 2 [cited 2014 Jan 27];159(1):73. Available from: http://annals.org/article.aspx?articleid=1700652
17. Epic: Meaningful Use Stage 2 Certification Details [Internet]. [cited 2014 Feb 3];Available from: https://www.epic.com/software-certification.php
18. Lobach DF, Kawamoto K, Anstrom KJ, Kooy KR, Eisenstein EL, Silvey GM, et al. Proactive population health management in the context of a regional health information exchange using standards-based decision support. *AMIA Annu Symp Proc [Internet]*. 2007 Jan [cited 2014 Mar 3];:473–7. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2655792&tool=pmcentrez&rendertype=abstract
19. Raja AS, Gupta A, Ip IK, Mills AM, Khorasani R. The use of decision support to measure documented adherence to a national imaging quality measure. *Acad Radiol [Internet]*. 2014 Mar [cited 2014 Feb 13];21(3):378–83. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24507424
20. Lakshminarayan K, Rostambeigi N, Fuller CC, Peacock JM, Tsai AW. Impact of an electronic medical record-based clinical decision support tool for dysphagia screening on care quality. *Stroke [Internet]*. 2012 Dec [cited 2014 Jan 27];43(12):3399–401. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3508397&tool=pmcentrez&rendertype=abstract
21. Kawamoto K, Jacobs J, Welch BM, Huser V, Paterno MD, Del Fiol G, et al. Clinical information system services and capabilities desired for scalable, standards-based, service-oriented decision support: consensus assessment of the Health Level 7 clinical decision support Work Group. *AMIA Annu Symp Proc [Internet]*. 2012 Jan [cited 2014 Jan 22];2012:446–55. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3540445&tool=pmcentrez&rendertype=abstract

22. OpenCDS Home [Internet]. [cited 2014 Jan 30];Available from: http://www.opencds.org/
23. Kawamoto K, Shields D, Del Fiol G. OpenCDS: Enabling Clinical Decision Support at Scale through Open-Source, Standards-Based Software and Resources [Internet]. *AMIA poster*. 2011 [cited 2014 Feb 24];:1830. Available from: http://assets.conferencespot.org/fileserver/file/121836/filename/1lqa1c.pdf
24. Department of Health and Human Services (HHS). Voluntary 2015 Edition Electronic Health Record Certification Criteria: Interoperability Updates and Regulatory Improvements [Internet]. 2014 [cited 2014 Mar 7];Available from: http://www.regulations.gov/#!documentDetail;D=HHS-OS-2014-0002-0001
25. Schleyer AM, Schreuder AB, Jarman KM, Logerfo JP, Goss JR. Adherence to guideline-directed venous thromboembolism prophylaxis among medical and surgical inpatients at 33 academic medical centers in the United States. *Am J Med Qual [Internet]*. 2011 [cited 2014 Jan 29];26(3):174–80. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21490270
26. Turpie AGG, Lassen MR, Davidson BL, Bauer KA, Gent M, Kwong LM, et al. Rivaroxaban versus enoxaparin for thromboprophylaxis after total knee arthroplasty (RECORD4): a randomised trial. *Lancet [Internet]*. 2009 May 16 [cited 2014 Jan 30];373(9676):1673–80. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19411100
27. Palmer AJ, Schramm W, Kirchhof B, Bergemann R. Low molecular weight heparin and unfractionated heparin for prevention of thrombo-embolism in general surgery: a meta-analysis of randomised clinical trials. *Haemostasis [Internet]*. [cited 2014 Jan 30];27(2):65–74. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9212354
28. National Quality Measures Clearinghouse | Surgical care improvement project: percent of surgery patients who received appropriate VTE prophylaxis within 24 hours prior to Anesthesia Start Time to 24 hours after Anesthesia End Time. [Internet]. [cited 2014 Jan 30];Available from: http://www.qualitymeasures.ahrq.gov/content.aspx?id=35538
29. Kern LM, Malhotra S, Barrón Y, Quaresimo J, Dhopeshwarkar R, Pichardo M, et al. Accuracy of electronically reported "meaningful use" clinical quality measures: a cross-sectional study. *Ann Intern Med [Internet]*. 2013 Jan 15 [cited 2014 Feb 20];158(2):77–83. Available from: http://annals.org/article.aspx?articleid=1556360

A Framework for Incorporating Patient Preferences to Deliver Participatory Medicine via Interdisciplinary Healthcare Teams

Craig Kuziemsky, Ph.D.[1], Davood Astaraky, MSc[1], Szymon Wilk, Ph.D.[2], Wojtek Michalowski, Ph.D.[1], Pavel Andreev, Ph.D.[1]
[1]University of Ottawa, Ottawa, ON, Canada; [2]Poznan University of Technology, Poznan, Poland

Abstract

Participatory medicine refers to the equal participation of patients and interdisciplinary healthcare team (IHT) members as part of care delivery. Facilitating workflow execution is a significant challenge for participatory medicine because of the need to integrate IHT members into a common workflow. A further challenge is that patient preferences should be considered when executing a workflow. To date there is limited research on supporting patient workflow as part of participatory medicine practices. To address that shortcoming we used a two-phase approach to develop a framework for participatory medicine that integrates different IHT members and workflows including the incorporation of patient preferences about care delivery options. Our framework uses a domain ontology to define the patient, IHT concepts and relations, as well as a workflow for operationalizing participatory medicine via an IHT. Proof of concept of the proposed framework is illustrated with a palliative care pain management case study.

Introduction

An increased prevalence of chronic illnesses and the identification of the benefits of team-based healthcare delivery are resulting in more care delivery by interdisciplinary healthcare teams (IHTs) [1,2]. One of the big challenges of care delivery via an IHT is integrating the various workflows and information flows of the different IHT members [3]. While models exist for supporting clinical workflows of individual providers, these models are often unsatisfactory when they are scaled up to support IHTs [4,5]. Therefore new approaches are needed to develop distributed models that represent a meso perspective of how clinical workflows and IHTs are integrated as part of care delivery.

Participatory medicine refers to the equal participation of patients and/or their family members and clinical IHT members in decisions about care delivery [6]. Developing solutions to support participatory medicine is largely about supporting distributed IHT workflows; including having the appropriate IHT member profile for specific tasks [6]. Patient participation in care delivery can have great benefits that include reducing adverse events [8], decreasing hospital lengths of stay [9], increasing patient satisfaction, and improving outcomes such as pain management [10]. However it is a challenge to incorporate patient oriented workflows as part of delivery participatory medicine. [4,7].

In theory, health information systems (HISs) are well suited to support participatory medicine via IHTs as they can coordinate the diverse processes and information flows across different IHT members. A shortcoming with existing HISs, however, is that they often focus on individual workflows and not on the distributed workflow that defines an IHT [7,11]. Before we can design HISs to support participatory medicine we need to develop better models of participatory medicine and the distributed workflows that define it [7]. In particular there is a need for models that incorporate patients as active members of an IHT. To date most HIS applications to support patient participation are generic and focus largely on information access and not on how to support patient participation in processes like decision-making. Examples of existing patient participatory applications include healthcare specific applications like 'Patients Like Me' and generic social media platforms like Facebook or Twitter. A new term, "healthicant," has been suggested to describe technology-enabled applications to enable patient participation in their healthcare delivery [12]. HIS design to support "healthicant" activities differs from current patient-participatory applications (i.e. Patients Like Me) in that the focus of HISs needs to be on eliciting patient preferences and using them in the creation and execution of patient-oriented workflows [12].

Although the need for care delivery via IHTs and participatory medicine has been well described there is a lack of frameworks showing how to integrate patient preferences and IHT workflow to deliver such deliver care. Existing models of IHTs are conceptual and does not provide details of how to operationalize them in clinical settings [13]. Moreover most of them focus just on isolated tasks conducted by individual providers [3]. Furthermore there is a shortcoming of frameworks showing how to incorporate patients as part of participatory medicine. Patient participation is a workflow issue for clinicians who are unprepared for the impact that patient participation will have on their workflow [14].

We believe that there are three key areas related to IHTs and participatory medicine that require more research attention. First, we need more formal approaches for developing distributed models of IHTs that support the dynamic integration an IHT and a workflow including assignment of tasks and team leadership and maintenance. Second, we need to incorporate patient preferences into the models to enable true participatory medicine. Third, we need approaches to enable us to operationalize the model in clinical settings. This paper addresses these three research needs by developing a framework for participatory medicine that integrates different IHT members and workflows including the incorporation of patient preferences about care delivery options. We present a two-phase modeling approach to develop a framework for participatory medicine by an IHT that includes patient participation. We then illustrate operationalization of our framework with a case study of palliative care pain management. We conclude the paper with a discussion and implications of our research.

Method

Ontological engineering was used as the research approach for modeling participatory medicine via an IHT. Ontologies formalize knowledge by means of concepts, attributes to characterize them, and relations between the concepts [15]. In ontological engineering a domain ontology, often referred to as a knowledge base, is developed first. Application ontologies are then developed by defining instances of classes from the domain ontology for the purpose of solving specific problems [16]. Ontologies have been used in healthcare for various tasks such as systems design, execution of clinical guidelines and representation of workflow for point of care decision support system design [17, 18]. Ontologies are particularly valuable for modeling concepts that are dynamic or flexible such as healthcare teams [19].

Proposed Framework

We followed a two-phase process to develop a participatory medicine-Interdisciplinary health Team (PM-IHT) framework. In the first phase we developed a conceptual model in the form of a textual description of important concepts needed for modeling an IHT (e.g., team, patient, preferences, and workflow), relations between them, and strategies to operationalize these concepts (e.g., a strategy to form and maintain a team, integrate patient preferences). In the second phase we formalized the conceptual model into a domain ontology and a set of algorithms detailing the operationalization strategies. In the following sections we describe the results from both phases with an emphasis on the domain ontology, as it is the central component of our framework.

The conceptual modeling phase (phase 1) assumes that an IHT manages a patient according to a case-specific workflow. We define a case as a disease or specific trauma that is managed by a workflow. The IHT is formed when the disease specific workflow is identified and patient management starts and is disbanded when the execution of the workflow (and related patient management) is completed. The team has a *leader* responsible for overseeing the execution of the workflow, for handling exceptional situations and for assigning workflow tasks to appropriate team members (i.e., members who are capable of performing the task). Drawing upon the principles of participatory medicine [6], the leader may consult with the patient prior to making any relevant decisions (e.g., selecting a therapy, or selecting a team member). The team has one leader at a time but s/he can change during workflow execution (e.g., at one point a physician might be the leader, while at another this role is transferred to a nurse). Finally, an IHT can be modeled as static entity, where a team is formed and remains in place for the entire workflow, or a dynamic entity, where members are recruited as needed and discharged after the task is executed. Given that care delivery in itself is a dynamic process, our conceptual model employs a hybrid approach where the IHT always has a leader (static approach), while other team members are recruited as needed to execute tasks from the workflow and then are dismissed from the team after task completion (dynamic approach). However, it is important to stress here that the leadership role is static but the team member assigned to this role can change.

Phase 2 formalizes the conceptual model into a domain ontology that serves as the PM-IHT framework (Figure 1). The framework can be divided into three areas that define the concepts and relations associated with the team, patient and workflow respectively. These three areas are not mutually exclusive but rather there are two concepts, *valued capability* and *presentation,* that are shared between the three areas. The concept of valued capability is fundamental for our framework. It couples a capability, understood as an ability to perform a certain clinical task [20] (e.g., ability to conduct a particular assessment), with an additional score that indicates the competency associated with this capability (for example, it may correspond to the level of clinical expertise). The other shared concept, presentation, represents any presenting complaint by the patient (e.g., a disease, symptom, or other problems).

We now describe each of the three main areas (team, patient, and workflow) of our PM-IHT framework.

Team-related Concepts and Relations

Team-related concepts are *team,* representing an IHT, and *practitioner,* representing any care provider (physician, nurse, pharmacist, dietitian, etc.) who is a member of the IHT. Practitioners are characterized as having valued capabilities, which provides the fine-grained description that is necessary to allocate team members to tasks [20]. The team manages a patient case with a certain presentation by executing a workflow specific for this presentation. The workflow is formed upon the patient's arrival and dissolved once the execution of the workflow has been terminated. Team members – practitioners – are recruited dynamically based on their valued capabilities (see the description of the workflow-related concepts below) and dismissed after they have completed their delegated tasks.

The team has a leader who is responsible for considering and following the patient's preferences when executing a workflow, for delegating tasks to team members, and for handling exceptional situations (e.g., inability to reconcile patient preferences). The leader is appointed just after forming the team and the leadership may change throughout the workflow execution depending on the requirements prescribed in the workflow (see the description of workflow related concepts below).

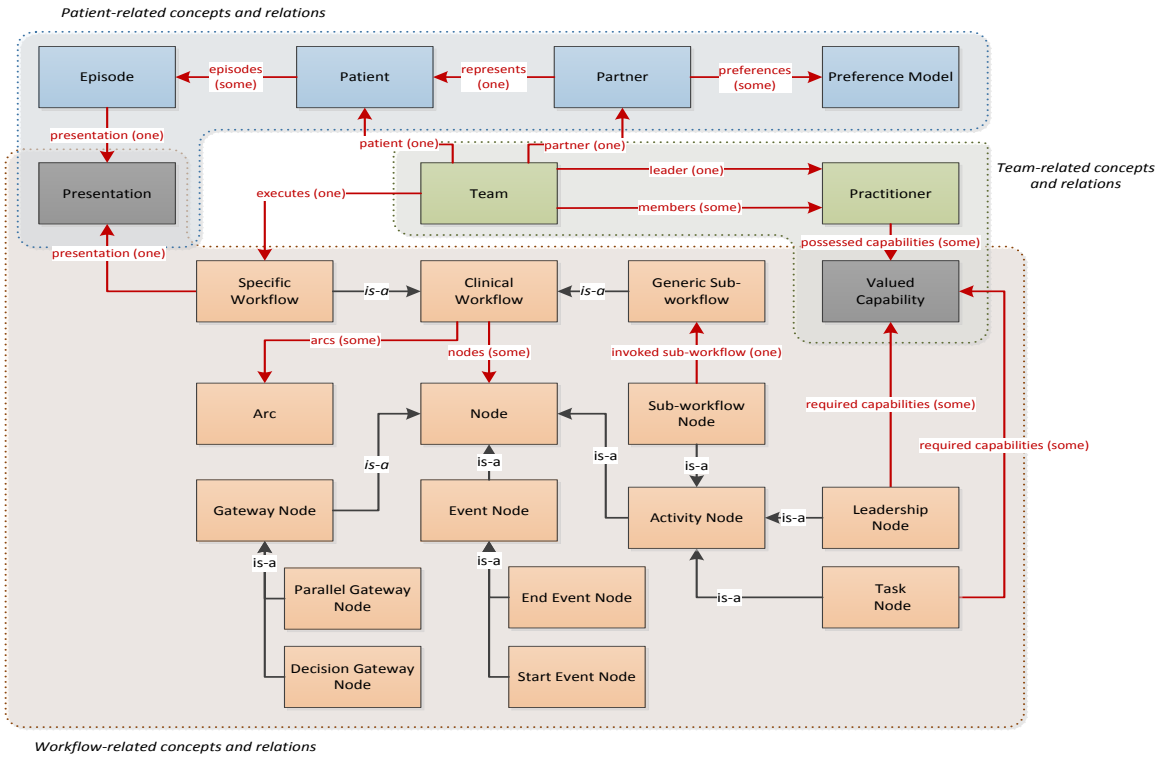


Figure 1. Domain ontology for the PM-IHT framework

Patient-related Concepts and Relations

The patient-related concepts include *patient*, *partner*, *preference model,* and *episode*. The first two concepts complement each other – patient represents a receiver of care and "provider" of clinical data (associated with an episode of specific presentation), and a partner represents the patient as an active participant in all decision-making activities that affect care delivery. The partner concept is crucial from the perspective of participatory medicine.

A partner has one or more preference models that are consulted before making various types of decisions (e.g., selecting a specific intervention among different possibilities). While there are different approaches to represent preference models, we advocate using functional models with additive value functions. Such models are well established and widely used in clinical practice in the form of various score or measures. A drawback of functional preference models is the difficulty associated with their elicitation. However, several methods, e.g., GRIP [21], have been proposed to construct such models from examples of decisions in order to structure relevant preferences.

In terms of interactions between the partner and the team, it is the leader who is responsible for communicating with the partner, presenting possible decision options, and capturing suggestions. More specifically, before recruiting any new team member, the leader presents eligible candidates to the partner who applies the preference model to identify the most preferred one. Moreover, wherever a task involves any decision-making related to the patient (e.g., selecting a therapy or a procedure), the available choices are communicated by the leader to the partner who consults a preference model in order to identify the best choice.

While in most cases the patient and partner is the same person, there are situations where it necessary to have a distinction between the two concepts because they may be two different people. For example, during the care of a pediatric patient it is parent or guardian who becomes the partner. In palliative care it may be a family member who is the partner if the patient is too ill or otherwise unable to make a decision. Thus, separation of these two concepts allows us to employ the principles of participatory medicine regardless of circumstances.

Workflow-related Concepts and Relations

The workflow-related concepts and relations between them have been inspired by Business Process and Model Notation (BPMN) [22]; however, they have been expanded to address the specificity of IHTs and participatory medicine. The central concept is *clinical workflow* that is specialized into *specific workflow* and *generic sub-workflow*. The former represents a "top-level" workflow that is associated with a specific presentation that is executed upon arrival of a patient with this presentation. The latter is a generic workflow (i.e., not associated with any presentation) that may be invoked by other workflows as a supporting workflow.

Each clinical workflow is composed of *arcs* and *nodes*, so it can be seen as a directed graph. The node concept is specialized into a *gateway node*, *event node* and *activity node*, depending on its purpose. The gateway node is further specialized into *decision gateway node* and *parallel gateway node*, which allows for conditional branching and parallel paths respectively. Event nodes are used to indicate starting and ending nodes in a workflow (via *start event node* and *end event node*). Finally, an activity node is specialized into *task node*, *sub-workflow node* and *leadership node,* corresponding to the three types of basic activities in a workflow – executing a clinical task, invoking a sub-workflow, and appointing the team leader, respectively. A clinical task can be delegated to a single practitioner who satisfies a requirement specified in terms of valued capabilities. Such a requirement is defined by associating valued capabilities with a task node. A practitioner has to possess all the required capabilities and his/her competency scores of possessed capabilities should be equal or better than competency scores of the required capabilities. The same mechanism for specifying and checking requirements is applied to leadership nodes – the only difference is that selection of the leader has a more permanent nature as the selected leader is retained until the next leadership node is encountered. Moreover, if a new leader is appointed in an invoked sub-workflow, the previous leader is retained and restored once the sub-workflow has been completed. As we explained above, the leader is selected after forming the team and the team has to have a leader at all times. This is achieved by introducing at least one leadership node into a workflow right after the start event node.

Case study: Palliative Care Pain Management

We used a case study of a palliative care pain management derived from a local palliative care unit. Palliative care aims to improve quality of life of patients and their families through the prevention and relief of suffering by means of early identification and assessment and treatment of pain and other symptoms including physical, psychosocial, and spiritual symptoms [23]. Palliative care is well suited for illustrating the interplay between participatory medicine, workflow execution, and patient management because a basic tenet of palliative care is that patients (or family members) are involved in care decisions and processes that implement decisions.

Pain management is a significant component of any palliative care management protocol as it is estimated that up to 70% of advanced cancer patients experience pain [24]. Pain in palliative care patients is complex and can include physical, spiritual, and psychosocial dimensions [25]. These different dimensions necessitates that palliative care be delivered by an IHT composed of palliative care physicians, psychologists, physiotherapists, nurses, and occupational therapists, among others [26]. The complexity of palliative pain also means that there is a range of possible pain management therapies (i.e. medical, spiritual, psychological) depending on the type and level of pain. A key aspect of palliative care pain management is that patient therapeutic preferences need to be incorporated into the decision-making and subsequent delivery of care services. For example, some patients prefer no pain and are willing to accept the side effects of opioid medications (e.g. drowsiness) to achieve that outcome. Other patients prefer to spend quality end of lifetime with family and friends and would prefer therapies that combine opioid and non-pharmacological therapies in order to manage drowsiness. Thus palliative care pain management needs to be tailored to the preference of each individual patient [27].

The process of palliative care pain management starts with assembling a palliative care team according to the framework presented earlier in order to execute the appropriate workflow. Drawing on existing pain management guidelines [25, 28] and in consultation with domain experts, we developed a set of workflows (a specific workflow and several sub-workflows) to sequence and structure all activities associated with assessing, diagnosing, and managing a palliative care patient's pain. Figure 2 shows a specific (top-level) workflow and selected sub-workflows. We used a standardized notation inspired by BMPN (Business Process Model and Notation) [22] to represent the workflows. In order to distinguish between particular types of activity nodes (represented as rounded rectangles) we use colors – sub-workflow nodes are white, task-nodes are light grey, and leadership nodes are dark grey. All task and leadership nodes are associated with required valued capabilities with competency scores expressed on the Likert scale (1 indicates the competency of a beginner/resident, 2 – the competency of a staff clinician, and 3 – the competency of an expert). The competency scores (and also the valued capabilities in table 1) are for the illustrative purposes only. They were determined from the pain management guideline in consultation with the palliative care clinicians consulting us on this research. Finally, we use the "consult with partner" tag to highlight the task nodes where the partner's preferences need to be incorporated and reconciled.

Patient management starts with the task of establishing an IHT leader who possesses the capabilities specified in the workflow. An initial assessment of the patient's condition follows. Completion of this activity requires executing an assessment sub-workflow that involves clinical and psychosocial and assessments and pain evaluation. While clinical and psychosocial assessments are tasks to be completed by appropriate team members (therefore they are ascribed with the required valued capabilities that control recruitment and selection of practitioners), pain evaluation is guided by a separate sub-workflow. This sub-workflow involves pain evaluation using the PQRST (provoking, quality, region, severity, and timing) pain evaluation tool for pain classification [29]. Once the initial assessment is completed, execution of the workflow continues. Palliative care pain management is an ongoing process that involves frequent reassessments and adjustments of the therapies in response to changing state of the patient [28]. The workflow terminates when therapy is no longer revisable and a new therapeutic workflow is needed.

In order to illustrate the versatility of the proposed framework, we will use two examples showing how the IHT is operationalized. One example involves the transition of a leader role, and the other illustrates how a partner's (patient, family member, etc.) preferences are explicitly taken into account when the therapy-planning task is executed.

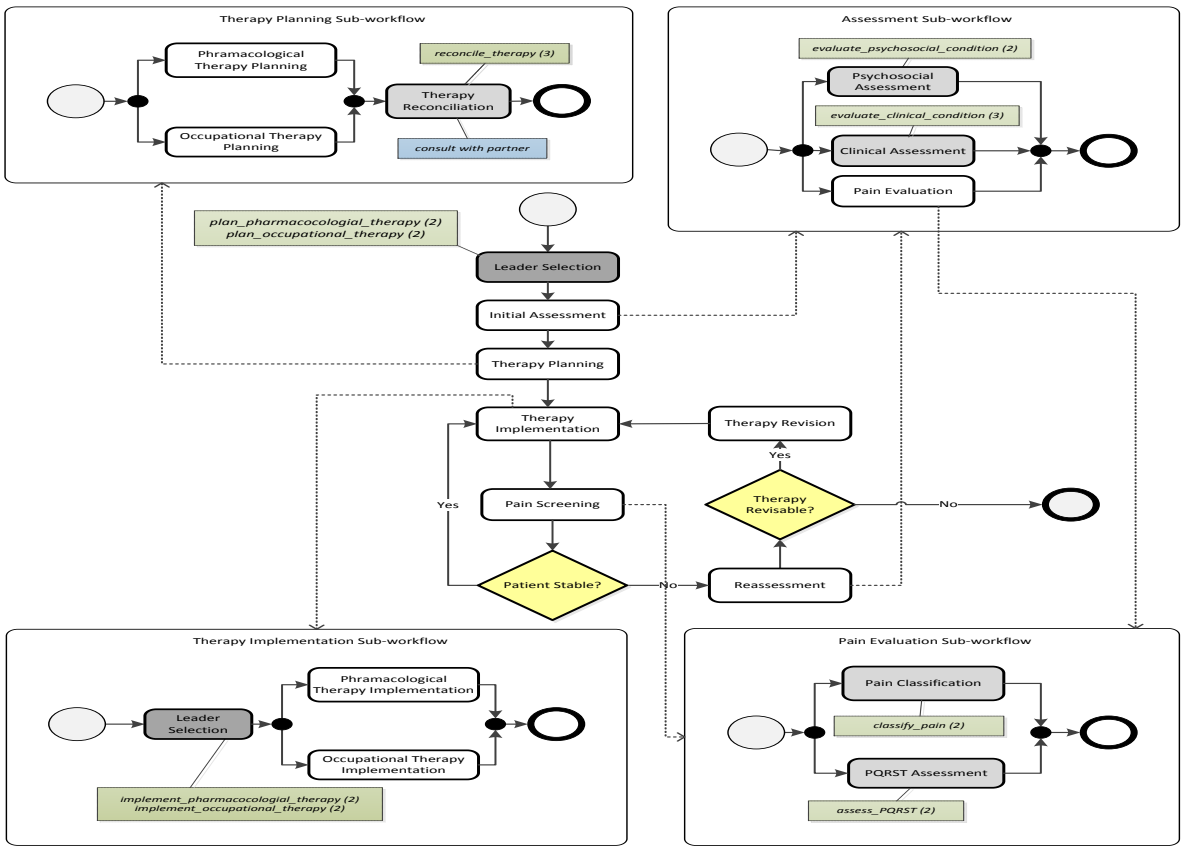


Figure 2. Overall IHT workflow for palliative care pain management (scores in brackets are competency scores for the task)

Example 1: Dynamic Selection of the Team Leader

Table 1 provides information about the practitioners who could be IHT members executing the palliative care workflow from Figure **2**. At the outset of palliative care pain management, a team leader is identified that is capable of planning the pharmacological and occupational therapies (i.e. the leader is chosen based on the competency scores ascribed to the tasks in Fig 2).

Table 1. Team members and possessed valued capabilities

Practitioner	**Possessed valued capabilities**
Palliative care physician	*access_PQRST* (2) *classify_pain* (2), *evaluate_clinical_condition* (3) *reconcile_therapy* (3) *plan_pharmacological _therapy* (3) *plan_occupational therapy* (2)
Clinical nurse specialist	*assess_PQRST* (3) *classify_pain* (3) *evaluate_clinical_condition* (2) *evaluate_psychosocial_condition* (2) *implement_pharmaceutical _therapy* (2) *implement_occupational _therapy* (3)
Nurse practitioner	*evaluate_clinical_condition* (1) *implement_pharmaceutical _therapy* (2) *implement_occupational_therapy* (1)

According to the data from Table 1, the only practitioner who meets the leadership requirement is a palliative care physician, who therefore becomes the team leader. The leadership remains unchanged until the therapy implementation sub-workflow needs to be executed. At that time new leadership requirements are encountered and in this case a clinical nurse specialist becomes the IHT leader. Once the therapy implementation sub-workflow is completed, the leadership goes back to the palliative care physician.

Example 2: Considering Partner's Preferences in Therapy Planning

Participatory medicine advocates that a patient is an active participant in decision-making and care plan development. According to the IHT framework presented in the paper, we incorporate the tenet of participatory medicine through the concept of a "partner" – who can be a patient, family member, or legal custodian. We illustrate how patient preferences interplay with a tasks' execution by the practitioners by using an example of therapy planning for pain management (specifically we focus on the therapy reconciliation task).

The patient's pain was assessed and determined to be in step 1 on the WHO Analgesic Ladder [30]. This corresponds to an assessment of a mild to moderate pain that should be managed with non-opioid analgesics, possibly combined with an adjuvant such as anticonvulsant or antidepressant medication. Assessment of the patient's psychosocial and clinical condition identified the possibility of combining pharmacological management with a non-pharmacological approach for pain management. The following three therapeutic options were identified by the IHT: (1) a standard therapy, (2) a therapy with alternative route of administration (transdermal) and (3) a less frequently applied therapy using cannabinoids (they have been demonstrated to have significant analgesic properties, but also are associated with a number of side effects like dizziness or nausea). The three therapies are given in detail in Table 2.

Table 2. Identified therapeutic options

	Pharmacological	**Adjuvant**	**Non-pharmacological**
Therapy 1	Acetaminophen administered orally	Antidepressant medication	Superficial heat & cold method
Therapy 2	NSAID with transdermal administration	Antidepressant medication	Guided imagery
Therapy 3	Cannabinoids by oral mucosal spray	None	None

The patient preference model for therapy selection in this particular case was constructed using the Generalized Regression with Intensities of Preferences (GRIP method) [21]. First, possible therapy options were characterized using two criteria: complexity of a therapy and associated drowsiness. The complexity of a therapy was defined by the compliance burden associated with this therapy – for example, a therapy that involves taking one dosage of medication per day has low complexity, while a therapy that involves multiple doses using different modes of administration has high complexity. Then, we elicit patient preferences by asking a patient to compare pairs of possible therapies. Once this preferential information is collected, it is processed by GRIP to calculate the so-called marginal value functions for the patient. These functions represent patient preferences associated with individual criteria (there is a separate function for each criterion) and constitute the patient preference model (see Table 3). It includes two marginal value functions (complexity and drowsiness), each represented as a set of numerical scores (marginal values) for all possible evaluation criteria. Higher marginal values are preferred by the patient. The patient preference model shows that the patient prefers therapies of low complexity that cause minimal drowsiness.

Table 3. Patient preference model for assessing therapies

Complexity		**Drowsiness**	
Evaluation	Marginal value	Evaluation	Marginal value
low	0.4	minimal	0.6
medium	0.2	moderate	0.2
high	0.0	maximal	0.0

The patient preference model not only gives insight into patient preferences, but it can be also used to evaluate therapies according to these preferences. The overall value of a therapy is computed by adding the marginal values

for each individual criterion. The best therapy would get the overall value of 1.0, while the worst one would get a value of 0.0. The preference model is then used to assess the potential therapies (see Table 2). Table 4 presents the evaluations for each therapy for the two criteria, as well as the overall value. Based upon the patient's preferences, therapy 3 is the most preferred therapy (overall value of 1.0) - it involves just one medication type, is simple to administer, and also causes minimal drowsiness. Therapy 3 is therefore selected to manage the patient's pain.

Table 4. Assessment of identified therapies

	Complexity	**Drowsiness**	**Overall value**
Therapy 1	medium	moderate	0.4
Therapy 2	low	moderate	0.6
Therapy 3	low	minimal	1.0

Discussion

While participatory medicine is a common healthcare objective there is a shortcoming of studies that illustrate how to operationalize it through an IHT. Further, despite calls for patient centered workflows there is a dearth of studies showing how to implement them in clinical settings. In this paper we presented a framework for IHTs that integrate multiple workflows to support participatory medicine that includes patient participation. The novelty of our research lies in the detail we provide on how the IHT is assembled for a particular situation. A patient's presenting condition drives the creation of a workflow specific for that presentation. We then use the notion of capabilities and competencies to assign IHT members to tasks as part of the workflow. We also assign an IHT leader role to the workflow that is responsible for considering and incorporating patient preferences into the workflow and for delegating tasks to IHT members based upon capabilities. Finally, we define a series of patient related concepts that enable the active incorporation of patient preferences into the IHT workflow. A preference model was developed to formalize the potential options to enable the patient or family member to make an informed decision about their preference.

To date much of the work on IHTs has focused on specific activities such as information seeking or group decision-making. However an IHT is not a set of isolated tasks but rather a distributed workflow that needs to integrate IHT members and task capabilities during workflow execution. IHT workflow is also very dynamic and a key part of workflow execution is coordinating team members to be in the right place at the right time to enable timely completion of workflow tasks and re-scoping of workflow if patient preferences change. Our framework provides the means of coordinating IHT members (through team leaders, capabilities and competencies) as well as re-tailoring the workflow if patient preferences change.

Our framework has implications for HIS design to support participatory medicine. A shortcoming of existing HISs is that they do not assume team-based management or allow for incorporating patient preferences, and therefore are not designed to support participatory medicine. The first step in HIS design is requirements definition and development of a conceptual model to drive HIS design. Our framework provides such a foundation by first developing a conceptual model of patient and IHT support as part of operationalizing participatory medicine. In phase 2 we formalize the conceptual model into an ontology that describes the concepts, relations, and workflows for participatory medicine. Two specific contributions from our framework are one, an attempt to implement participatory medicine principles by explicitly accounting for patient's preferences and two, the dynamic alignment of an IHT and a workflow, including support for assignment of tasks to practitioners, team leadership, and team maintenance. A key factor in HIS design to support participatory medicine is that system design cannot be done with a one size fits all mindset. IHT composition and the workflows that support it are dynamic and may change according to patient preferences. A key aspect of our model is that it enables a HIS to be designed to enable run time agility to support the dynamic nature of IHTs (i.e. leadership changes, evolving tasks) as part of implementing participatory medicine.

The next step in our research is translating our framework into a prototype multi-agent system (MAS). This will be done in two stages. First we first we assume that the practitioners are represented by the agents. Second, we will use the domain ontology (fig. 1) and follow ontology-driven design combined with O-MaSE [18] to design a prototype participatory medicine HIS. The HIS will be implemented using Workflows and Agent Development Environment (WADE) [31] for workflow execution.

In this paper we provided methodological foundations of our framework and illustrated them with a case study of palliative care pain management. We discussed how our framework enabled selection of team leader and how participatory medicine principle of patient involvement was explicitly modeled through inclusion of a "partner" team member and a preference model. While the research presented in this paper is illustrated using a specific case study of palliative care pain management, the proposed modeling framework is applicable for supporting other patient management situations involving IHTs.

Conclusion

There is an increasing body of knowledge calling for care delivery via participatory medicine that includes patient participation. However there are few studies that provide the means of operationalizing the workflow to deliver patient engaged participatory medicine. This paper provides a framework of how to operationalize participatory medicine, including patient preferences, via an IHT. Our framework can be used to design and evaluate informatics solutions for supporting participatory medicine.

Acknowledgements

We acknowledge funding support from the Natural Sciences and Engineering Research Council of Canada.

References

1. AAMC. Core competencies for interprofessional collaborative practice: Report of an expert panel. *Interprofessional Education Collaborative* Washington, D.C. 2011
2. Reddy MC, Gorman P, Bardram J. Special issue on Supporting Collaboration in Healthcare Settings: The Role of Informatics. Int J Med Inform. 2011;80:541–543.
3. Kuziemsky CE, Williams JB, Weber-Jahnke JH. Towards electronic health record support for collaborative processes. Proceedings of the 3rd Workshop on Software Engineering in Health Care (SEHC '11). ACM press, 2011. pp 32-39
4. Ozkaynak M, Brennan PF, Hanauer DA, Johnson S, Aarts J, Zheng K, et al. Patient-centered care requires a patient-oriented workflow model. J Am Med Inform Assoc 2013; 20 (e1): e14–6.
5. Unertl KM, Johnson KB, Lorenzi NM. Health information exchange technology on the front lines of healthcare: workflow factors and patterns of use. J Am Med Inform Assoc 2012; 19:392–400
6. Hood L, Auffray C. Participatory medicine: A driving force for revolutionizing healthcare. Participatory medicine: a driving force for revolutionizing healthcare. Genome Medicine 2013;5:110.
7. Ozkaynak M, Brennan P. An Observation Tool for Studying Patient-oriented Workflow in Hospital Emergency Departments, Methods Inf Med 2013; 52: 503–513
8. Jain M, Miller L, Belt D, King D, Berwick DM. Decline in ICU adverse events, nosocomial infections and cost through a quality improvement initiative focusing on teamwork and culture change. Qual Saf Health Care. 2006;15:235–239
9. Zwarenstein M, Goldman J, Reeves S. Interprofessional collaboration: effects of practice-based interventions on professional practice and healthcare outcomes. Cochrane Database Syst Rev. 2009;(3):CD000072
10. Rodriguez SML, Beaulieu MD, D'Amour D, Ferrada-Videla, M. The determinants of successful collaboration: a review of theoretical and empirical studies. Journal of Interprofessional Care 2005; supplement 1:132-147
11. Dorr DA, Jones SS, Wilcox, A. A framework for information system usage in collaborative care. Journal of Biomedical Informatics 2007;40(3): 282-287
12. Sherer SA. Patients Are Not Simply Health IT Users or Consumers: The Case for "e Healthicant" Applications. Communications of the Association for Information Systems 2014; 34, Article 17:351-364
13. Leggat, SG. Effective healthcare teams require effective team members: defining teamwork competencies. BMC Health Services Research 2007;7:17

14. http://cmaer.wordpress.com/2014/02/24/himss14-the-patient-has-no-clothes/ - Danny Sands. Last Accessed March 13, 2014
15. Leonardi G, Panzarasa S, Quaglini S, Stefanelli M, van der Aalst WM. Interacting agents through a web-based health serviceflow management system. Journal of Biomedical Informatics 2007; 40(5):486-499
16. Shaw M, Detwiler LT, Brinkley JF, Suciu D. Generating application ontologies from reference ontologies. AMIA Annu Symp Proc. 2008; 2008: 672–676
17. Isern D, Sánchez D, Moreno A. Ontology-driven execution of clinical guidelines. Computer Methods and Programs in Biomedicine 2012;107 (2):122-139.
18. Wilk SZ, Michalowski W, O'Sullivan D, Farion K, Kuziemsky C, and Sayyad-Shirabad JJ. Task-Based Architecture for Developing Point-of-Care Decision Support Systems for the Emergency Department. Methods of Information in Medicine 2013; 52(1): 18-32
19. Kuziemsky CE, Yazdi S. A methodology and supply chain management inspired reference ontology for modeling healthcare teams. Studies in Health Technology and Informatics 2011;169, 719-723
20. Astaraky D, Wilk S, Michalowski W, Andreev P, Kuziemsky C, Hadjiyannakis S. A Multiagent System to Support an Interdisciplinary Healthcare Team: A Case Study of Clinical Obesity Management in Children. Proceedings of the VIII Workshop on Agents Applied in Health Care. Murcia, 2013, 69-80.
21. Figueira J, Greco S, Słowiński R. Building a set of additive value functions representing a reference preorder and intensities of preference: GRIP method. Eur J Oper Res 2009; 195:460–486.
22. Scheuerlein H, Rauchfuss F, Dittmar Y, Molle R, Lehmann T, Pienkos N, et al. New methods for clinical pathways-Business Process Modeling Notation (BPMN) and Tangible Business Process Modeling (t.BPM). Langenbecks Arch Surg 2012;397(5):755-61
23. World Health Organization Definition of Palliative Care. Available at: http://www.who.int/cancer/palliative/definition/en/ Last accessed March 12, 2014
24. Van den Beuken-van Everdingen MHJ, De Rijke JM, Kessels AG et al. Prevalence of pain in patients with cancer: a systematic review of the past 40 years. Ann Oncol 2007;18:1437–1449
25. Hanks G, Cherny N, Kaasa S, et al., editors. Oxford textbook of palliative medicine. Section 10: The management of Common Symptoms and Disorders 10.1: The management of pain 4th ed. Oxford: Oxford University Press; 2009.
26. Kao CY, Hu WY, Chiu, TY, Chen CY. Effects of the hospital-based palliative care team on the care for cancer patients: An evaluation study. International Journal of Nursing Studies 2014; 51(2):226-235
27. García-Toyos, N., Escudero-Carretero, M. J., Sanz-Amores, R., Guerra-De Hoyos, J.-A., Melchor-Rodríguez, J.-M. and Tamayo-Velázquez, M.-I. Preferences of Caregivers and Patients Regarding Opioid Analgesic Use in Terminal Care. Pain Medicine 2014;15: 577–587.
28. Yamaguchi T, Shima Y, Morita T et al. Clinical guideline for pharmacological management of cancer pain: The Japanese Society of Palliative Medicine Recommendations. Jpn J Clin Oncol 2013; 43(9): 896–909.
29. McCarberg B, Stanos S. Key patient assessment tools and treatment strategies for pain management. Pain Pract 2008;8:423-32
30. WHO Analgesic Ladder for Adults. Taken from http://www.who.int/cancer/palliative/painladder/en/ Last accessed March 13, 2014
31. WADE: Workflows and Agents Development Environment. Available from: http://jade.tilab.com/wade/.

Medical Alert Management: A Real-Time Adaptive Decision Support Tool to Reduce Alert Fatigue

Eva K. Lee, PhD [*,1,2,3], Tsung-Lin Wu, PhD[1,2,3], Tal Senior, RN[4], James Jose, MD[4]
[1]Center for Operations Research in Medicine and HealthCare, [2]NSF I/UCRC Center for Health Organization Transformation, [3]Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, [4]Children's HealthCare of Atlanta, Atlanta, GA

Abstract

With the adoption of electronic medical records (EMRs), drug safety alerts are increasingly recognized as valuable tools for reducing adverse drug events and improving patient safety. However, even with proper tuning of the EMR alert parameters, the volume of unfiltered alerts can be overwhelming to users. In this paper, we design an adaptive decision support tool in which past cognitive overriding decisions of users are learned, adapted and used for filtering actions to be performed on current alerts. The filters are designed and learned based on a moving time window, number of alerts, overriding rates, and monthly overriding fluctuations. Using alerts from two separate years to derive filters and test performance, predictive accuracy rates of 91.3%-100% are achieved. The moving time window works better than a static training approach. It allows continuous learning and capturing of the most recent decision characteristics and seasonal variations in drug usage. The decision support system facilitates filtering of non-essential alerts and adaptively learns critical alerts and highlights them prominently to catch providers' attention. The tool can be plugged into an existing EMR system as an add-on, allowing real-time decision support to users without interfering with existing EMR functionalities. By automatically filtering the alerts, the decision support tool mitigates alert fatigue and allows users to focus resources on potentially vital alerts, thus reducing the occurrence of adverse drug events.

*Corresponding author: Eva K Lee, eva.lee@gatech.edu

Introduction

Much research has been performed on prescribers' views on alerts and their reasons for overriding alerts, such as Hsieh et al. 2004 for drug allergy alerts, Grizzle et al. 2007 and Ko et al. 2007 for drug-drug interaction alerts, and Shah et al. 2006 and Van der Sijs et al. 2006 for multiple types of alerts.

Too many alerts and too many alert overrides could lead to alert fatigue, which could threaten patient safety. Some research focuses on improving the specificity of alerts to reduce inappropriate alerts. Hsieh et al. (2004) studied characteristics of drug allergy alert overrides and made specific recommendations for increasing the specificity of alerting. Shah et al. (2006) suggested making the least severe alerts non-interruptive, not requiring a user action. Van der Sijs et al. (2008) interviewed prescribers about turning off frequently overridden drug-drug interaction alerts, finding that most of them wanted to reduce alert overload but they did not agree on which alerts could be safely turned off. Seidling et al. (2009) deduced the upper dose limits for statins from pharmacokinetic studies, incorporating dosage checking into alerts for dose-dependent drug-drug interaction alerts. Lee et al. (2010) designed a decision-learning framework to consistently identify override alerts, which could be filtered automatically to reduce alert fatigue. Riedmann et al. (2011) comprehensively studied factors which could be used to prioritize alerts in order to present them adequately. Twenty factors were identified and grouped into three categories: organization unit, patient/case, and alert itself.

This paper extends the study from Lee et al. (2010). In the previous study we built filters from three criteria to decide which alerts could possibly be automatically turned off or become non-interruptive. The

three criteria include (1) number of alerts, (2) override rate, and (3) range of monthly override rate. The effect of the filters was validated by applying them to real datasets. In this study we propose to use a moving time-window to build the filters for real-time adaptive decision support. We also analyze the effectiveness of the criteria via linear regression.

Methods

Based on the alert datasets from Children's Healthcare of Atlanta (CHOA), we focus on three types of alerts: dose alerts, drug allergy alerts, and drug-drug interaction alerts. Drug allergy alerts consist of four severities: level 1, 2, 3, and 4. Drug-drug interaction alerts consist of three severities: moderate interaction, severe interaction, and contraindicated drug combination. We sort out drugs or interactions under each alert type and severity. The filter is built on four criteria: (1) number of previous months used for training (size of time window), (2) number of alerts, (3) override rate, and (4) range of monthly override rates. We implement moving time-windows for continuous learning of the cognitive overriding decision of users. Specifically, to filter the alerts of the current month, alert data and their associated user decision patterns from a specified number of previous months serve as the training set to establish the filter decision rule. Such a dynamic design is both intuitive and appealing as it captures users' decision characteristics from previous months to help predict the alert action at the current time. For each drug/interaction, the automatic filter is established via the following criteria on the training set:

- the total number of alerts is greater than or equal to a threshold,
- the overall override rate is greater than or equal to a threshold, and
- the range of monthly override rates (i.e., the difference between the maximum and minimum monthly override rates) is less than or equal to a threshold.

The filtered drug/interaction is considered most likely to be overridden by users based on previous patterns and the nature of the drug alert.

To evaluate the predictive accuracy and effectiveness of each of the filtering criteria, we contrast the results against filters that are established via i) static time-window prediction; ii) a time gap between training set time to the blind prediction set; and iii) regression confidence interval analysis.

Medication order alert data and their override patterns by providers from CHOA are used to validate and test our methodology. The data cover January to September 2009 and January to July 2011, and include only unfiltered alerts output by the EMR system. These alerts require manual review and actions by providers. Using moving time-windows, we build the filter for each month based on previous months. Table 1 shows the total number of unfiltered drugs/interactions for different alert types and severities as obtained from the EMR system. Table 2 shows the total number of unfiltered alerts in different months based on alert types. The last row in Table 2 reports the percentage of unfiltered alerts that were eventually overridden by the providers.

Table 1: Number of drugs/interactions in different alert types and severities

Alert type	Severity	Number of drugs/interactions
Dose alert	N/A	851
Drug allergy alert	Level 1	110
	Level 2	48
	Level 3	206
	Level 4	89
Drug-drug	Moderate interaction	184

interaction alert	Severe interaction	130
	Contraindicated drug combination	37
Total		1655

Table 2a: Number of unfiltered alerts for the period January to September 2009

Alert Type	2009								
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep
Dose	1330	1382	1396	1323	1301	1344	1334	924	1213
Drug Allergy	580	604	585	499	463	492	549	575	682
Drug-Drug	892	1021	1178	1075	1141	892	1091	877	969
Total	2802	3007	3159	2897	2905	2728	2974	2376	2864
% overridden	65.8%	63.2%	63.5%	59.6%	64.7%	65.4%	63.9%	73.6%	74.6%

Table 2b: Number of unfiltered alerts for the period January to July 2011

Alert Type	2011						
	Jan	Feb	Mar	Apr	May	Jun	Jul
Dose	8516	8747	9878	8811	9174	8903	8736
Drug Allergy	1360	1349	1619	1281	1465	1403	1437
Drug-Drug	1293	1238	1453	1642	1305	1335	1260
Total	11169	11334	12950	11734	11944	11641	11433
% overridden	77.9%	79.1%	77.4%	77.1%	76.2%	76.8%	75.0%

There is a significant difference in the number of unfiltered alerts between these two periods as the hospital expanded the EMR alert system to cover more units. , In addition, clinical staff performs regular review and adjusts EMR alert parameters to reflect current clinical practice guidelines, thus resulting in different levels of alert reporting. . For those alerts being fired out (unfiltered), providers must review each of them manually and carefully to determine if the prescribed action is appropriate or if other actions should be taken. Those for which the current prescribed action is deemed appropriate result in alerts being overridden. Those that are not overridden require that providers take additional and/or alternative action. Reviewing all alerts can be a tedious and time-consuming task. Alert fatigue has been documented and reported on in numerous studies and can be detrimental to clinical outcome. Hence, a decision support system that can identify and appropriately handle non-essential alerts is critical because it enables providers to focus on vital ones and make proper decisions.

Results

In this section we describe a test implementation of our methodology at CHOA, a network of three pediatric hospitals in Atlanta, Georgia, which serves over half a million patients annually.

Predictive Performance of the Time Window Adaptive Filters

We consider the following criteria to construct filters:

a) *Time-Window*: Size of time window = 2, 3, 4, 5, 6 months
b) *Number*: Number of alerts ≥ 20, 30, 40, 50, 60
c) *Rate:* Override rate ≥ 90%, 92%, 94%, 96%, 98%
d) *Range*: Range of monthly override rates ≤ 10%, 12%, 14%, 16%, 18%

As an example, we use alert data from January and February 2009 for training (thus, size of time window is 2 months) and choose the number of alerts to be at least 20, with override rate of at least 90%, and override rate range to be within 10% to construct a filter, *Filter1*. *Filter1* is then applied to the March 2009 alerts, resulting in ten drugs/interactions being filtered. Figure 3a shows the type of alerts being filtered versus the actual action by the providers. Those filtered by us and overridden by providers are true negatives (that is, these are non-essential alerts). Those not overridden by providers but filtered by us are false negatives. All unfiltered alerts are reviewed manually by providers.

Counts of alerts (and associated percentages) filtered by us versus counts of alerts overridden by providers are shown in Table 3b. Of those that are manually overridden by providers, we filter 22.5% of them. Such automatic filtering can save precious time, enabling providers to tend to other important patient care issues. Of those alerts that are not overridden by providers but are filtered by us, we commit a 3.5% error. There may be serious medical implications to these errors. We pursue two actions to correct such errors: a) *adaptive learning* to incorporate these alert contents within our filtering algorithm; b) second-layer filtering with inclusion of patient risk factors to enrich our filtering algorithm's predictability. From Table 3b, we note that the predictive accuracy of this filter is 91.9% (452/(452+40)) . This high value indicates that if an alert is filtered (by our algorithm), there is a high level of confidence that the filtering result is correct in predicting the actions of the providers.

Table 3a: Drugs/interactions from March 2009 alerts that are filtered by *Filter1*.

Type	Severity	Drug name	Providers' actions	
			Overridden #	Non-overridden #
Dose	(N/A)	CHLOROTHIAZIDE SODIUM	2	1
Dose	(N/A)	FOSPHENYTOIN	8	0
Dose	(N/A)	KCL	120	11
Dose	(N/A)	ROCURONIUM	2	0
Drug-Allergy	Level 3	INSULIN REGULAR HUMAN	0	0
Drug-Allergy	Level 3	MORPHINE SULFATE	16	6
Drug-Drug	Severe Interaction	CYCLOSPORINE/AZOLE ANTIFUNGAL AGENTS	71	1
Drug-Drug	Severe Interaction	KETOROLAC / ANTICOAGULANTS	204	12
Drug-Drug	Severe Interaction	LAMOTRIGINE / VALPROIC ACID	18	1
Drug-Drug	Severe Interaction	METHOTREXATE / PENICILLINS	11	8
		Subtotal	452	40

Table 3b: Filtered alerts and associated rates for *Filter1* on March 2009 alerts

Filtering results / Providers' action		Counts		Percentage	
		Filtered by us (negative)	Not filtered by us (positive)	Filtered by us	Not filtered by us
Counts	Manual Overridden (negative)	452	1555	22.5% (*Specificity*)	77.5% (*False positive rate*)
	Non-overridden (positive)	40	1112	3.5% (*False negative rate, FNR*)	96.5% (*Sensitivity*)
Percentage	Manual Overridden (negative)	91.9% = 452/(452+40) (Negative	58.3%		

		predictive value, *NPV*)	
	Non-overridden (positive)	8.1%	41.7%

Table 3c: Predicted results of *Filter 1* -- *Time-Window*=2, *Number*≥20, *Rate*≥90%, *Range*≤10%

Predicted month / Metrics	2009							2011					Average
	Mar	Apr	May	Jun	Jul	Aug	Sep	Mar	Apr	May	Jun	Jul	
Specificity (% of overridden alerts filtered by us)	22.5%	30.4%	30.9%	33.1%	31.8%	30.3%	17.7%	14.5%	18.0%	20.6%	24.8%	26.3%	25.1%
FNR (% of non-overridden alerts filtered by us)	3.5%	4.4%	3.7%	2.3%	4.1%	7.5%	3.9%	3.3%	5.8%	5.3%	6.1%	6.9%	4.7%
NPV (% of predictive accuracy)	91.9%	91.0%	93.9%	96.4%	93.2%	91.9%	93.1%	93.7%	91.3%	92.6%	93.1%	91.9%	92.8%

To understand the tradeoffs among s*pecificity*, the false negative rate (*FNR*), and the negative predictive accuracy (*NPV*) with respect to different criteria, we derive 625 filters based on the four criteria, each taking on five potential values ($5^4 = 625$). Each filter is applied to all potential training sets and is used to predict the override pattern of the following month. For example, a filter with time window of size 5 uses alert data and associated decisions from five consecutive months to predict the override pattern of the following month. So, January to May alert data and decisions are used as a training set to establish the rule, and the override pattern of June is predicted. Similarly, February to June alert data and provider decisions are used to develop a rule that is used to predict the override pattern for July. And so forth. Performing this on all alert data from January 2009 to June 2011 leads to 5000 training sets and prediction results. The predictive status of each alert is compared against the actual providers' override action. These values are summarized via boxplots in Figures 1a-c, grouped with respect to specific values for each criterion. From these figures we observe that our filtering performance is most sensitive to *Rate*. As *Rate* increases (i.e., filters are set to be more stringent), the percentage of overridden alerts being filtered (*specificity*) decreases significantly (Figure 1a), fewer false negative errors are made by the filter (Figure 1b), and the predictive accuracy of the filter increases (Figure 1c). We also observe that *Time-Window* could have some effects on the filtering results, whereas *Number* and *Range* appear to have marginal impact.

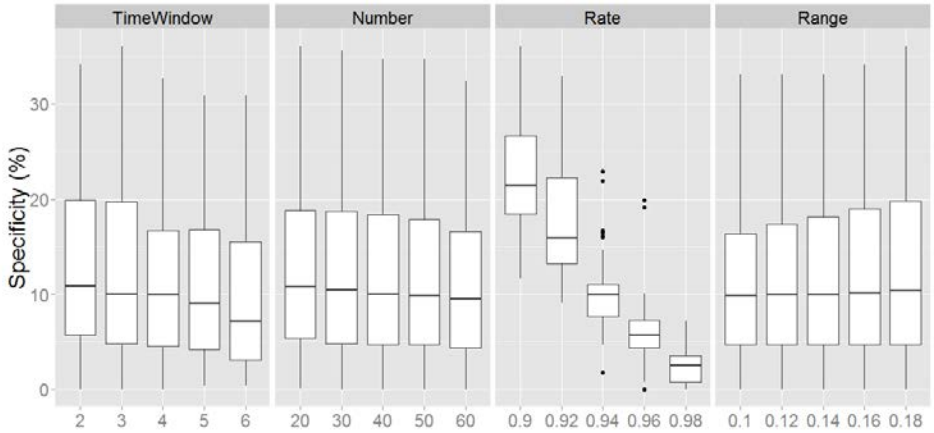


Figure 1a: Boxplots of % of overridden alerts being filtered (s*pecificity*) with respect to single-criterion value change.

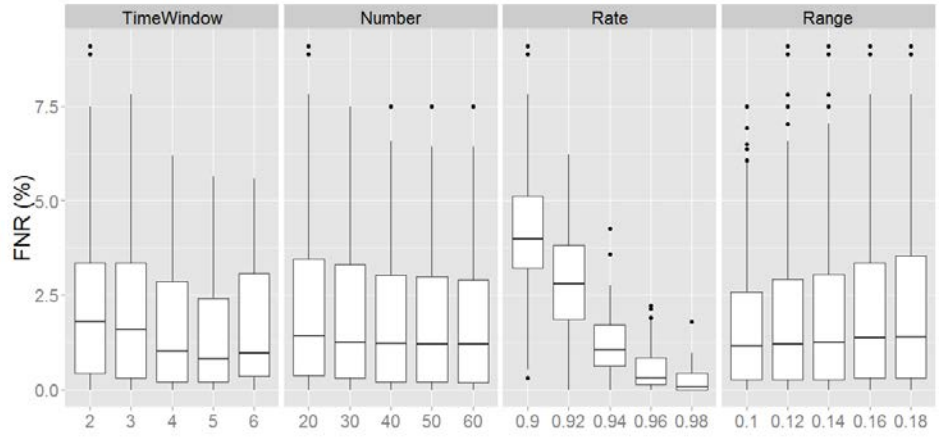


Figure 1b: Boxplots of % of non-override alerts being filtered (false negative rate) with respect to single-criterion value change.

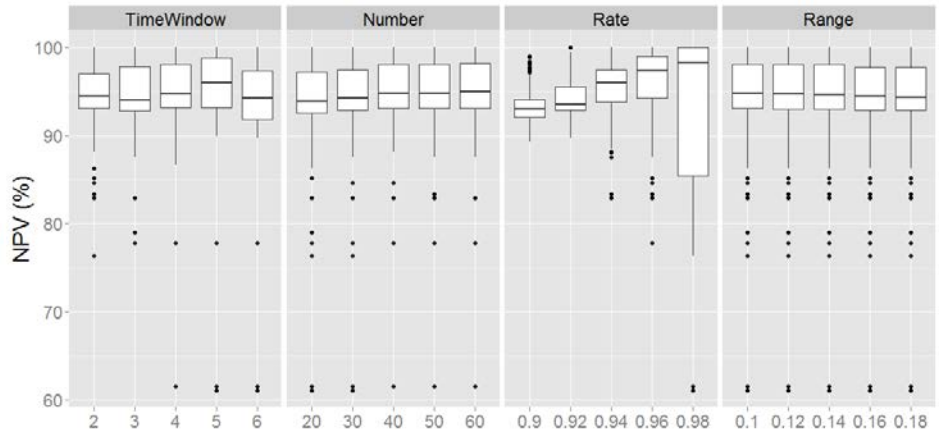


Figure 1c: Boxplots of predictive accuracy % among filtered alerts (overridden and filtered/total filtered) with respect to single-criterion value change.

Among the 625 filters, we present below the most conservative one. For each filter, the average values of S*pecificity*, *FNR*, and *NPV* over different training/prediction months are taken as consideration to measure the overall performance. The goal is to have a high-specificity, low-FNR, and high-NPV filter. However, these metrics are somewhat conflicting, as reflected in the scatterplot matrix in Figure 2. Here we pick a filter with the best average rank over these three metrics:

a) *Time-Window:* Size of time window = 5 months
b) *Number*: Number of alerts ≥ 40
c) *Rate*: Override rate ≥ 94%
d) *Range*: Range of monthly override rate ≤ 10%

Using this filter the *Specificity* ranges from 6.6%~10.8%, *FNR* ranges from 0%~1%, and *NPV* ranges from 93.8%~100%, as shown in Table 4.

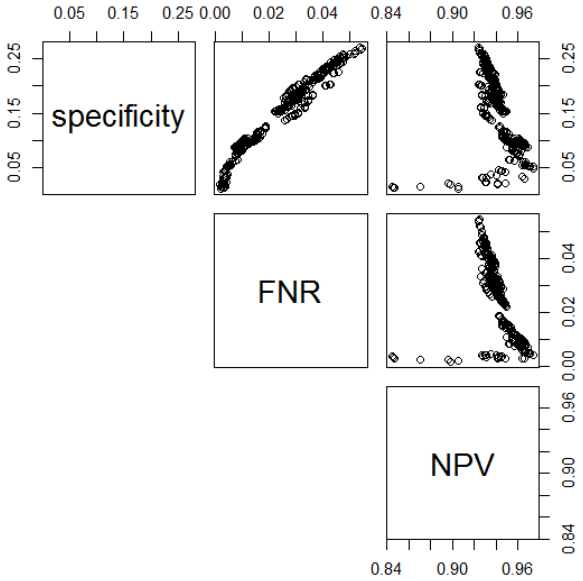


Figure 2: Scatterplot matrix of *Specificity*, *FNR*, and *NPV*

Table 4: Predicted results with *Time-Window*=5, *Number*≥40, *Rate*≥94%, *Range*≤10%

Metrics \ Predicted month	Jun '09	Jul '09	Aug '09	Sep '09	Jun '11	Jul '11	Average
Specificity (% of overridden alerts filtered by us)	9.1%	7.2%	8.2%	6.6%	10.6%	10.8%	8.8%
FNR (% of non-overridden alerts filtered by us)	0.0%	0.8%	0.6%	0.8%	1.0%	1.0%	0.7%
NPV (% of predictive accuracy)	100.0%	93.8%	97.3%	95.9%	97.3%	97.0%	96.9%

Effectiveness of the Filtering Criteria

We examine the effectiveness and significance of these criteria in more detail using regression analysis, focusing on understanding the importance of the moving time window, the allowance of time gap between the training months to the predictive months, and the degree of significance of each criterion. For brevity, we summarize the findings when the time window is set to 3.

Moving time-window versus static time-window First we compare the results of using moving time-windows as training set versus employing static time-windows. We perform the filtering by maintaining the same for all but the time window criterion. Specifically, the filters will include the parameters: number of alerts ≥ 20, 30, 40, 50, 60, override rate ≥ 90%, 92%, 94%, 96%, 98%, range of monthly override rate ≤10%, 12%, 14%, 16%, 18%.

We apply paired t-test to see whether there is significant outcome difference between that of moving time-windows and those from static time-windows. There are 5*5*5=125 filters, and for each filter we predict 8 different months (May-Sep 2009 and May-Jul 2011), so in this analysis the sample size is 125*8=1000. The average net *Specificity* of moving time-windows is 13.6% versus 11.3% from those obtained via the static time-window. The resulting p-value is less than 0.0001, indicating that using moving time-windows is more effective than static ones.

Gap between training months and predicting month Next we analyze how the filter performs if we allow a gap between the moving time-window and the month to be filtered (predicted). Under the same set of parameters in all criteria, we apply paired t-test to observe if there is any significant change in *Specificity* when using no gaps versus using gaps for the moving time-windows. The results are shown in Table 5. Regardless of the size of the gap (1, 2, or 3 months), the p-values are all less than 0.0001, indicating that using immediate previous months for training is more effective than using previous months with gaps. This reflects that the overriding patterns of immediate months reflect better the cognitive decision of the providers in the immediate next month.

Table 5: Comparison of net *Specificity* outcome between no gaps and with gaps. The gap is measured by the number of months between the months for training and the month for prediction

Size of gap	Sample size	Average net *Specificity*		p-value
		No gaps	With gaps	
1 month	1,000	13.6%	11.9%	<0.0001
2 months	750	13.9%	11.8%	<0.0001
3 months	500	13.9%	10.9%	<0.0001

Degree of Significance of Individual Criterion We conduct regression analysis to gauge the relative effectiveness of each criterion in our filtering schema. Specifically, *Specificity*, *FNR*, and *NPV* are used as the response variables separately, while size of *Time-Window*, *Number* of alerts, override *Rate*, *Range* of monthly override rate, and their pairwise interactions are the predictor variables. In the model where the response variable is *Specificity* or *FNR*, almost all predictor variables including their interactions have significantly small p-values; thus there are no nonsignificant ones. In the model where the response variable is *NPV*, only *Time-Window*, *Rate*, and their interaction term are significant, thus we can rule out *Number* and *Range*. Below is the regression equation to predict *NPV*, the predictive accuracy of the filter:

Predictive accuracy (*NPV*) ~ *0.053 + 0.265 Time-Window + 0.963 Rate – 0.287 Time-Window * Rate*

This finding validates the observations from the single-criterion boxplot results that *Rate* and *Time-Window* are the influencing criteria in determining the predictive accuracy (Figure 1c).

Discussion

With the adoption of electronic medical records (EMRs), drug safety alerts are increasingly recognized as valuable tools for reducing adverse drug events and improving patient safety. However, even with proper tuning of the EMR alert parameters, the volume of unfiltered alerts can be overwhelming to users. Studies have been performed to improve the specificity of alerts, to effectively prioritize and present the alerts for providers' review, and to automate the filtering of some unnecessary or least severe alerts. This remains a challenging problem as there is a tradeoff and disagreement between reducing alert overload versus the safety of turning off some alerts.

In this paper, we design an adaptive decision support tool in which past cognitive overriding decisions of users are learned, adapted and used for filtering actions to be performed on current alerts. The filters are designed and learned based on four criteria: moving time window, number of alerts, override rates, and monthly override fluctuations. Using alerts from two separate years to derive filters and test performance, prediction accuracy rates of 91.3% - 100% are achieved. This associates to 0.0% - 6.1% non-overridden alerts being filtered by our system. To reduce this false negative rate, we test the concept of *adaptive learning* by mining the key contents of these false negative alerts and incorporating them within our filtering algorithm. This helps to bring down the false negative rate to 0.0% - 1.5%. Further reduction may be possible by importing clinical and patient risk factors into our analysis. Added clinical features

may tag these to be non-filterable and thus correct the committed errors. Careful review of each such patient record is necessary. A follow-up study will have to be performd to validate its potential.

The moving time window works better than a static training approach. It allows continuous learning and capturing of the most recent overriding characteristics and seasonal variations in drug usage. Further, our study shows that the overriding patterns of immediate months reflect better the cognitive and clinical decision of the providers in the immediate next month. From regression analysis we conclude that the size of time window, number of alrts, override rate, range of monthly override rate, and their pairwise interaction are all critical factors in influencing specificity and false negative rate; whereas predictive accuracy is governed mostly by time window sizes and the override rates.

The clinical staff diligently reviews and fine-tunes EMR alert parameters to capture proper clinical practice guidelines. For those alerts being fired out, providers will manually review them to determine if the prescribed action is appropriate or if other actions should be taken. Those for which current prescribed action is deemed appropriate results in alerts being overridden. The non-overriding ones amount to important alerts that providers must take action on. Alerts fatigues have been reported by numerous studies and can be detrimental to clinical outcome. Our alert management decision support system helps to improve the specificity of alerts by reducing inappropriate alerts, thus enabling providers to focus their attention on important alerts and make proper decisions.

Among all the filters, some of the filtering levels are rather conservative, as roughly only 20% of overridden alerts are being filtered, with majority of the alerts remain unfiltered and require the attention of providers. Further, as the system “adapts” and ”learns” that certain drug interactions are critical and cannot be filtered and required specific actions, this knowledge will be transferred to the providers with alerts being tagged with knowledge of its importance. This provides an opportunity for training, in particular for those inexperienced providers who will benefit from the captured clinical knowledge.

From Table 2b, we observe that over 75% of alerts reviewed by providers are over-ridden. Using our system, about 20% of these are filtered. Using the estimate of 11,000 alerts per month and 1 minute per alert review, the estimated time savings is 27.8 provider-hours per month. More importantly, this enables providers to strategically allocate appropriate time to review vital alerts.

The decision tool can be plugged into an existing EMR system as an add-on, allowing real-time decision support to users without interference with existing EMR functionalities. We have tested it on EPIC. Specifically, the alerts directed out of EPIC are captured and funneled into the decision support tool. The tool grabs the data, establishes the filters, and performs the filtering. The tool then reports the alerts with status “filtered” and “not-filtered”. In the latter, it also highlights those “learnt” critical alerts. Providers can review them accordingly. By automatically filtering the alerts, the tool mitigates alert fatigue and allows users to focus resources on potentially vital alerts that require clinical decisions, thus reducing the occurrence of adverse drug events. Further, the learnt knowledge of critical alerts will facilitate the decision process. Clinical trials must be performed to fully evaluate the impact of this decision support tool on alert fatigue, patient safety and quality of care.

Acknowledgement

The study is partially supported by a grant from the National Science Foundation.

References

1. Hsieh TC, Kuperman GJ, Jaggi T, Hojnowski-Diaz P, Fiskio J, Williams DH, Bates DW, Gandhi TK. Characteristics and consequences of drug allergy alert overrides in a computerized physician

order entry system. *Journal of the American Medical Informatics Association*. 2004;11(6):482-491.

2. Grizzle AJ, Mahmood MH, Ko Y, Murphy JE, Armstrong EP, Skrepnek GH, Jones WN, Schepers GP, Nichol P, Houranieh A, Dare DC, Hoey CT, Malone DC. Reasons provided by prescribers when overriding drug-drug interaction alerts. *The American Journal of Managed Care*. 2007;13(10):573-580.
3. Ko Y, Abarca J, Malone DC, Dare DC, Geraets D, Houranieh A, Jones WN, Nichol WP, Schepers GP, Wilhardt M. Practitioners' views on computerized drug-drug interaction alerts in the VA system. *Journal of the American Medical Informatics Association*. 2007;14(1):56-64.
4. Shah NR, Seger AC, Seger DL, Fiskio JM, Kuperman GJ, Blumenfeld B, Recklet EG, Bates DW, Gandhi TK. Improving acceptance of computerized prescribing alerts in ambulatory care. *Journal of the American Medical Informatics Association*. 2006;13(1):5-11.
5. Van der Sijs H, Aarts J, Vulto A, Berg M. Overriding of drug safety alerts in computerized physician order entry. *Journal of the American Medical Informatics Association*. 2006;13(2):138-147.
6. Van der Sijs H, Aarts J, van Gelder T, Berg M, Vulto A. Turning off frequently overridden drug alerts: Limited opportunities for doing it safely. *Journal of the American Medical Informatics Association*. 2008;15(4):439-448.
7. Seidling HM, Storch CH, Bertsche T, Senger C, Kaltschmidt J, Walter-Sack I, Haefeli WE. Successful strategy to improve the specificity of electronic statin-drug interaction alerts. *European Journal of Clinical Pharmacology*. 2009;65:1149-1157.
8. Lee EK, Mejia AF, Senior T, Jose J. Improving patient safety through medical alert management: An automated decision tool to reduce alert fatigue. *AMIA Annual Symposium Proceedings 2010*. 2010:417-421.
9. Riedmann D, Jung M, Hackl WO, Stuhlinger W, van der Sijs H, Ammenwerth E. Development of a context model to prioritize drug safety alerts in CPOE systems. BMC Medical Informatics and Decision Making. 2011;11:35.

COPD Hospitalization Risk Increased with Distinct Patterns of Multiple Systems Comorbidities Unveiled by Network Modeling

Young Ji Lee, RN, PhD[1], Andrew D. Boyd, MD[2,3,4], Jianrong 'John' Li, MS[5], Vincent Gardeux, PhD[5], Colleen Kenost, MA[5,8], Don Saner, MS[7,8], Haiquan Li, PhD[5], Ivo Abraham, PhD, RN[6], Jerry A. Krishnan, MD, PhD[1,4,*], Yves A. Lussier, MD[5,7-9,‡,*]

[1]Department of Medicine, [2]Institute for Translational Health Informatics, [3]Departments of Biomedical and Health Information Sciences, [4]University of Illinois Hospital and Health Science System, University of Illinois at Chicago, Chicago, IL;[5]Department of Medicine, [6]Department of Pharmacy Practice and Science, [7]Cancer Center, [8]Biomedical Informatics Service Group, Arizona Health Science Center, [9]Interdisciplinary Program in Statistics, The University of Arizona, Tucson, AZ, USA;

[‡]The analysis of this study were conducted while at the University of Arizona;*corresponding authors

Abstract

Earlier studies on hospitalization risk are largely based on regression models. To our knowledge, network modeling of multiple comorbidities is novel and inherently enables multidimensional scoring and unbiased feature reduction. Network modeling was conducted using an independent validation design starting from 38,695 patients, 1,446,581 visits, and 430 distinct clinical facilities/hospitals. Odds ratios (OR) were calculated for every pair of comorbidity using patient counts and compared their tendency with hospitalization rates and ED visits. Network topology analyses were performed, defining significant comorbidity associations as having OR≥5 & False-Discovery-Rate≤10^{-7}. Four COPD-associated comorbidity sub-networks emerged, incorporating multiple clinical systems: (i) metabolic syndrome, (ii) substance abuse and mental disorder, (iii) pregnancy-associated conditions, and (iv) fall-related injury. The latter two have not been reported yet. Features prioritized from the network are predictive of hospitalizations in an independent set ($p<0.004$). Therefore, we suggest that network topology is a scalable and generalizable method predictive of hospitalization.

Introduction

Chronic Obstructive Pulmonary Disease (COPD) is the third leading cause of death in the United Sates and an important cause of disability and hospitalizations, particularly in aged populations[1,2]. Currently, about 12 million adults (aged 18 and over) have been diagnosed with COPD, but there are likely to be more who have yet to be diagnosed[3]. Deaths attributable to COPD in women are higher than men, with more than 6,000 deaths for women alone in 2010[4]. Hospitalizations for COPD account for a large economic burden to the patients and to society[5,6]. In that same year, COPD resulted in $49.9 billion in direct and indirect costs, and total costs incurred by COPD patients are approximately $6,000 higher than patients without COPD[6]. Due to its large burden on individuals and health care systems, it is crucial to find evidence-based strategies to identify those who are at the highest risk of being hospitalized in order to target preventive interventions[7,8].

Comorbidities such as hypertension, ischemic heart diseases, diabetes, and pneumonia have been known as contributing causes to COPD hospitalizations. Therefore, reducing COPD-associated comorbid conditions may decrease the hospitalization rate, which ultimately reduces the economic burden of COPD patients[7,9]. Multiple studies have developed predictive models including comorbidities of COPD hospitalizations and re-hospitalization; however, those models were constructed only on the use of regression models **(Table 1)**. These models aim at predicting COPD hospitalization from comorbidities data. However, they do not take into account the impact of all potential interactions between each comorbid condition, unless these and their interactions are specified *a priori* in the functional model. Also, COPD hospitalization risk predictions are based on labor-intensive scoring systems such as the Charlson Comorbidity Index (CCI)[10] or Elixhauser[11], which are the best-known comorbidity indices. However, these tools only provide pre-selected list of diseases[12] and therefore many diseases are not taken into account, and do not provide the level of granularity necessary to understand the clinical dynamics of COPD hospitalization risk. In other words, they exploit a biased and small number of features for predictions. As a result, comorbidities finding may be biased and have limitations for identifying novel patterns of comorbidities.

When investigating conditions or features that can predict high risks of admission in COPD patients, it is necessary to understand the association between COPD and its comorbidities. Studying the structure defined by the entire set of comorbidities with the novel methodology is required to understand the enriched association and unveil the hidden structure. Recently, network-based approaches have been applied to human disease and have revealed unknown connections between diseases, which shed new light on the clinical research realm (**Background**)[13]. For example, Hidalgo et al. introduced a Phenotypic Disease Network (PDN), which uses data obtained from the medical claims of more than 30 million patients to demonstrate that highly connected diseases are more lethal than barely connected ones[14]. However, results from these patterns have not been translated into classifiers of hospitalization.

In this study, we hypothesized that network topology modeling of COPD-associated comorbidities with higher risk of hospitalization and emergency department visits can predict future hospitalizations.

Table 1. Summary of three studies on the prediction of hospitalization among COPD patients

Author (yr)	Austin et al, 2012[7]	Fan et al, 2002[8]	Coventry et al, 2011[15]
Total patients	855,661	3,282	79
Validation			
Independent study		yes	yes
Cross-validation	yes		
Measurement of features			
Index name 1	Charlson	Charlson	Charlson
Count of features 1	19	19	19
Index name 2	Elixhauser		
Count of features 2	30		
Index name 3	Aggregated Diagnosis Groups (ADGs)		
Count of features 3	32		
Outcome measures			
Hospitalization for COPD	Yes	Yes	
Rehospitalization for COPD			Yes
Any hospitalization	Yes	Yes	
Analysis method	Logistic regression	Logistic regression	Logistic regression

Background on Network Analysis

These approaches aim to discover, map, and quantify complex relationships among variables that may not be revealed by correlational or clustering methods. Combining exploratory analytics with statistical decision points and graphical displays, these methods may discover complex networks of variables that were heretofore unknown, yield new insights into the dynamics among these variables, and link these networks to clinical outcomes.

Methodology

Sample: We collected data from the Illinois Health Connect Medicaid, which included patients from 430 distinct clinical facilities and hospitals from January 1, 2010 through December 31, 2010 and covered 1,446,581 visits. These clinical institutions have been chosen using the distinct provider ID from outpatient billing and inpatient billing. If a site had multiple locations and a single billing address, then it counted as one institution. Claims include all claims adjudicated for payment through April 29, 2011. Location (emergency department, inpatient, and outpatient), admission date, primary disease and secondary disease information with ICD-9-CM code were extracted for the network analysis. Although the system provided 5 digits level ICD9-CM code, we used the ICD-9-CM at the 3 digit level for the analysis. COPD was defined as ICD-9-CM code 490–492 and 496 based on the literature. COPD with asthma (ICD-9-CM code 493) was not included in this study since the COPD with asthma is under asthma category which is different from COPD. In total, 38,695 patients and 1,049 ICD-9-CM codes at the 3 digit level for COPD and comorbid diseases were included in the dataset.

Initially, 3,862 patients had a COPD diagnosis, and potentially other comorbid conditions (886 total comorbidities). We removed from further analysis ICD-9-CM codes associated to less than 20 patients, in order to prevent any future re-identification. This filtering resulted in a final dataset of 3,831 patients and 754 comorbidities. Among those data, 880 COPD patients were hospitalized at least once, and 2,711 patients visited the ED at least one time. Further, we created a subgroup of patients that we labeled as "higher risk of acute exacerbation (e.g. bronchospasms) (AE-risk) COPD patients", given the following inclusion criteria: 1) patients with COPD as either primary or secondary diagnosis, and 2) patients who visited emergency department (ED) 5 or more times and hospitalized at least once for the condition of COPD during 1/1/2010~12/31/2010. We identified 238 AE-risk COPD patients that met those two criteria (ED $\geq$ 5 AND hospitalization $\geq$ 1).

Data analysis: First, we randomly selected 25% of the "AE-risk COPD patients" to create a validation set of 60 patients and kept the remaining 178 patients (from the 238 AE-risk COPD patients) in the background. The aim was to identify comorbidities associated with an increased hospitalization risk. Specifically, we searched among the 754 comorbidities the ones that were significantly associated with the AE-risk patients, and the associations between these morbidities. We prioritized those comorbidities by using the associated Fisher's Exact Test (FET) result as the statistical criterion to determine whether a given comorbidity was retained in the network analysis (3,771 patients; **Figure 1)**. Our model can be viewed as an alternate feature selection procedure, where features are here comorbidities with the COPD condition. We created the association network of all comorbidities with the AE-risk COPD condition, and extracted the most significant ones and their interactions/associations. The levels of significance (p-values) by FET of all tested paired associations were adjusted by False Discovery Rate (FDR[16]) to correct for multiple comparisons for multiple comparisons and the possibility of finding a statistically significant result merely because of high statistical power. The analyses were performed using custom scripts written in Python[17] and R programs[18]. We used a stringent cutoff of significance for the association between diseases: Odds Ratio (OR)$\geq$5 of the hospitalization with and without the co-morbidities and FDR$\leq 10^{-7}$. For a one-year period, the number of hospitalizations and ED visits were computed for each patient using the MySQL Community Server 5.6. We bundled the high cost of recurrent ED visits with those of hospitalizations as proxies for high risk of COPD exacerbation associated to high overall health system costs.

Network modeling: To explore the pattern of COPD with multiple systems comorbidities, we constructed the network consisting of the significant associations (according to FET) of comorbidities of the AE-risk COPD patients **(Figure 2)**. The network has been constructed using Cytoscape[19]. Next, we investigated the tendency of hospitalization and ED visits according to the topology of the network. Our research group has extensive experience and pioneered network modeling in: (1) diseases[20-23], (ii) translational bioinformatics[24-39], and (iii) between multiple scales of molecules of life[40-43].

Evaluation: After the model was constructed from the background data, significant comorbidities were prioritized. We then retained the same comorbidities from the validation set (60 patients) and applied a clustering procedure in order to automatically separate those 60 patients into two subgroups, from their associated comorbidities. We used Partitioning Around Medoids (PAM) method for unsupervised clustering, as it is a well-established method for identifying subgroup (clusters) from data[44]. The R algorithm for Partitioning Around Medoids (PAM)[44] was utilized in a parameter-free way. It resulted in two clusters on which we compared Hospitalization and ED visits distribution. The comparison was statistically assessed using two-tailed non-parametric Mann-Whitney test (GraphPad PRISM v.5.0d). Of note, the PAM algorithm utilized exclusively comorbidity data and was not presented the hospitalization rates per patient nor ED visit.

IRB: The research project was approved by the University of Illinois Institutional Review Board id#2012-0150 (non-human subject, de-identified dataset).

Results

From the total 754 comorbidities found in COPD, 215 were significantly associated to rehospitalization (OR$\geq$5 and FDR$\leq 10^{-7}$, **Methods**) linked to patients at high risk of hospitalization and ED visits (AE-risk patients). These 215 comorbidities were regarded as prioritized features extracted from the training data (**Methods**). They were associated with each other, involving 280 interactions that we represented in a network (**Figure 2b**). Since all comorbidities were connected to COPD, we removed the COPD node from the network to explore further the associations between comorbidities. Of note, for a better readability of the network structures, only comorbidities containing associations with other comorbidities are shown in **Figure 2**. We provide the full network in supplement (**Supplement Figure S1**, http://lussierlab.org/publications/COPD_networks/SuppFigureS1.pdf).

We further annotated the network with the number of hospitalizations and ED visits for each comorbidity condition, in order to highlight the patterns they create in the network topology.

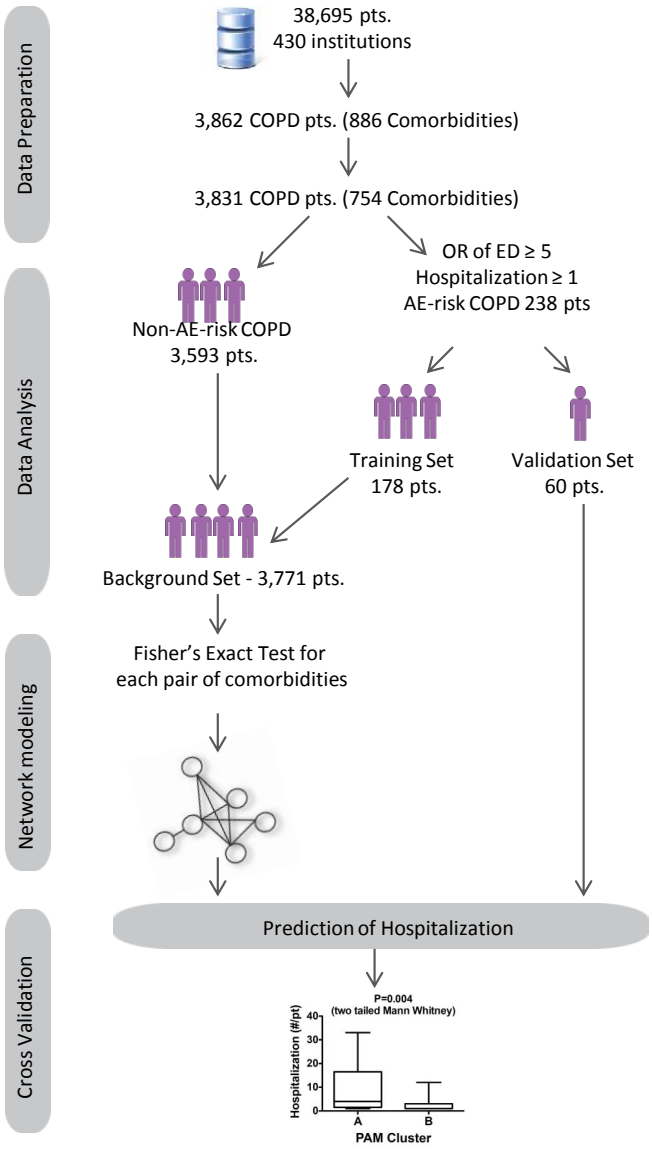


Figure 1. Summary of the study design

To express the multiplicity of information, we kept the original network but represented its nodes in three ways: 1) ICD-9-CM class of diseases and number of hospitalizations **(Figure 2, Panel A)**; 2) the number of ED visits per disease **(Figure 2, Panel B;** darker=more visits**)**; and 3) Odds ratios between AE-risk patients (≥1 hospitalization and ≥ 5 ED visits) and comorbidity **(Figure 2, Panel C;** darker for higher OR**)**. In **Panel A,** the size of the node corresponds to the number of hospitalizations, while it corresponds to the relative numbers in the other two panels. Moreover, in **Panel A**, the link thickness corresponds to the strength of the association (OR) between two comorbidity diseases (thicker link=higher OR). Of note, if patients visited the ED or were admitted to the hospital more than once, each visit or admission was counted separately. The network topology showed that among the total 754 comorbidities, the prioritized 215 comorbidities had an increased risk for hospitalizations and ED visits (FDR≤0.05 **Panel B**).

The clinical dynamics of this network analysis become evident from a pattern analysis of **Figure 2, Panel A.** Visual inspection reveals four distinct sub-networks: I = multisystem and metabolic syndrome, II = drug abuse and mental disorder, III = pregnancy-associated conditions and IV = fall-related injury. The sub-network I comprises hypertension, obesity, myocardial infarction, diabetes mellitus, disorders of lipoid metabolism, angina pectoris, and chronic ischemic heart diseases. The metabolic syndrome has already been defined as "a cluster of cardiovascular risk factors that is frequently associated with insulin resistance."[45] In this network each comorbidity disease is connected to the others, forming a complex sub-network. The metabolic syndrome is part of the largest network (**Figure 1, Panel A**), while many additional metabolic diseases are agglomerated to this multi-system. Considering the sub-network II, we can see that mental disorder and drug abuse are connected to each other (regarding their related comorbidities). This study also shows that falls, injuries and pregnancy-associated complications seem to be associated with COPD and its hospitalization (**Figure 2 sub-networks III and IV**).

The analysis of the sub-networks reveals that sub-network I is centered on the metabolic syndrome and is very complex in terms of clinical systems. Sub-networks I, II and IV each comprise diseases associated with some of the highest odds ratio of re-hospitalization and ED visits (**Figure 1, Panel C; Table 2**). Sub-network III is centered on pregnancy and while its odds ratio of re-hospitalization and ED visits is statistically and clinically significant, none of its morbidities are among the top 20 shown in **Table 2** and merit further investigation of studying separately the ED visits and hospitalizations. Additionally, Contusion to the trunk (**Sub-network IV**) stands out as a high risk for ED visits and hospitalization in COPD populations that may merit subsequent investigations for potential preventive measures.

In **Figure 2, Panel B**, the comorbidity conditions with the highest number of ED visit is 'Essential hypertension' followed by 'Nondependent abuse of drugs' and 'Diabetes Mellitus', which represents 625, 481 and 449 visits, respectively (**Table 2**). However, the number of ED visit seems not correlated to the OR; their corresponding odds ratio is "modestly" increased by 2.8. **Figure 2, Panel C** shows that 'Alcohol dependence syndrome', 'Disorders of mineral metabolism' and 'Contusion of trunk' are the top three comorbidities with highest OR (9.3, 8.2 and 8.0 respectively; **Table 2**), while their number of ED visits are low.

Table 2. Top COPD's co-morbidities associated to increased hospitalizations and ED visits as measured by OR. Of note, none of the top ORs was part of sub-network III (**Figure 1**).

Sub-network of Figure 1*	ICD-9 CM Code	Odds Ratio≥ 5 ED visits & 1 hospitalization	# patients	#ED visits	# Hosp.
I	**250 Diabetes mellitus**	2.9	**65**	**449**	229
	272 Disorders of lipoid metabolism	3.0	72	197	132
	275 Disorders of mineral metabolism	**8.2**	21	18	21
	276 Disorders of fluid, electrolyte, and acid-base balance	5.3	92	135	112
	285 Other and unspecified anemia	4.8	83	139	113
	288 Diseases of white blood cells	7.2	28	30	32
	V12 Personal history of certain other diseases	**6.0**	84	159	72
	401 Essential hypertension	2.8	**126**	**625**	308
	453 Other venous embolism and thrombosis	6.3	24	20	21
II	295 Schizophrenic disorders	4.3	32	108	124
	296 Episodic mood disorders	3.8	79	135	126
	298 Other nonorganic psychoses	**8.7**	33	19	13
	303 Alcohol dependence syndrome	**9.3**	34	60	49
	305 Nondependent abuse of drugs	3.9	**133**	**481**	288
	V15 Other personal history presenting hazards to health	5.3	84	164	134
IV	920 Contusion of face, scalp, and neck except eye(s)	4.6	19	16	1
	922 Contusion of trunk	**8.0**	21	22	5
	959 Injury, other and unspecified	3.2	76	25	2
	V62 Other psychosocial circumstances	5.5	35	62	56

- I = Multisystem and Metabolic syndrome, II = drug abuse and mental disorder, IV=fall-related injury

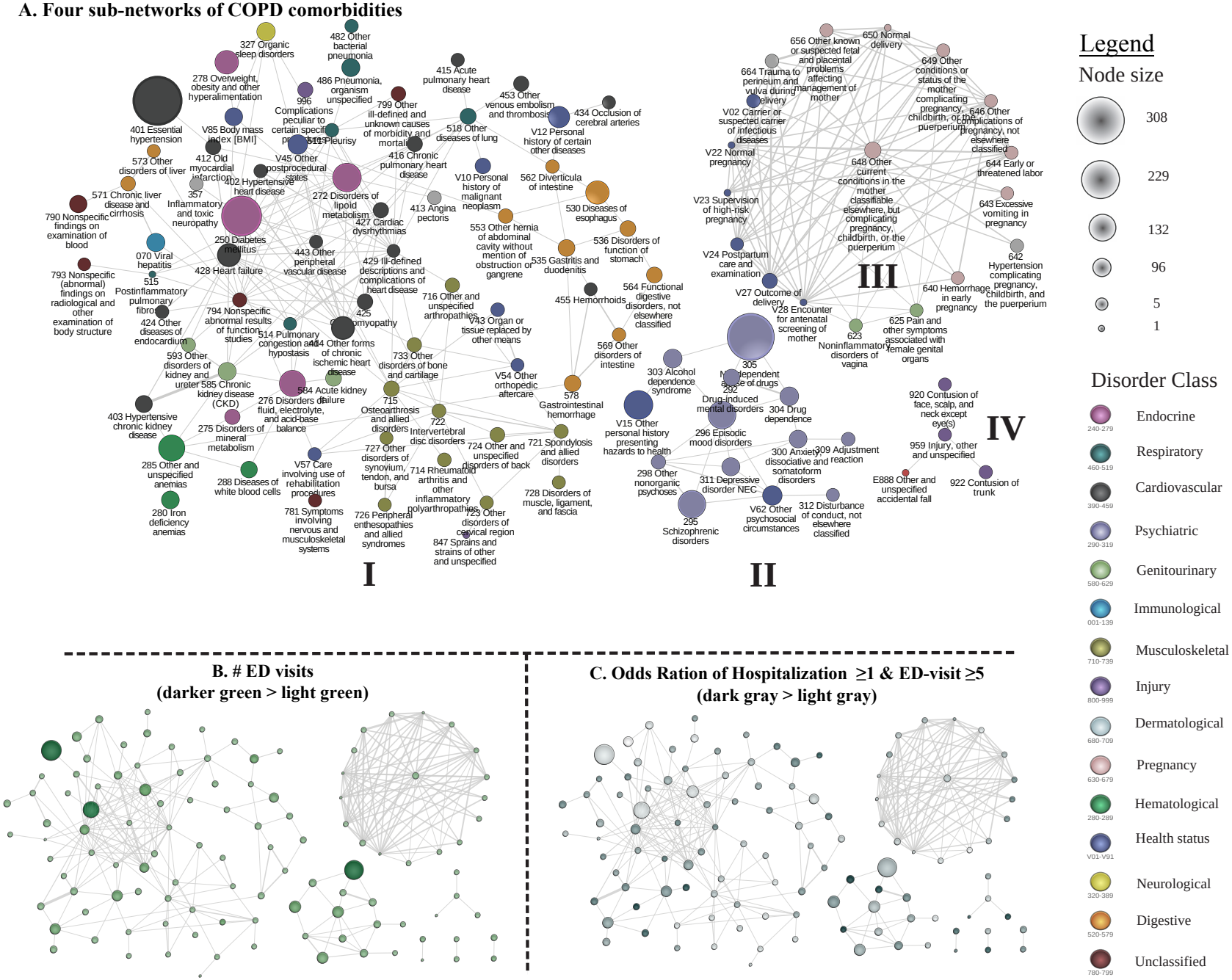


Figure 2. COPD-derived comorbidity network. Four sub-networks are shown in **Panel A** (details in **Table 2**). **Panels B and C** are highlighted for the number of ED visits and OR of hospitalizations respectively.

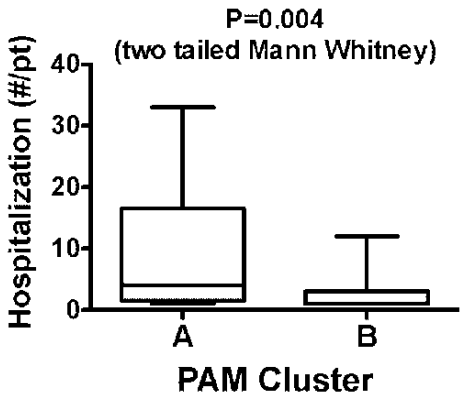


Figure 3. Evaluation in Independent Dataset Using Disease-Features Prioritized In Learning Set. Using the co-morbidities identified in the learning set (**Figure 2, panel A),** we then clustered the independent validation set using PAM clustering (restricted to two clusters), an unsupervised vectorial quantization method. Of note, PAM analysis did not have access to the hospitalization rates of the patients in the validation set. As shown here, the validation of the identified clusters was conducted by comparing the hospitalization rate of the identified clusters of the validation set.

Figure 3 describes how the 215 prioritized comorbidities are predictive of hospitalization in an independent data set. The same 215 comorbidities were retained from the 60 patients of the validation set (**Figure 1, Methods**). Then, we applied the PAM clustering procedure **(Methods)** in order to automatically separate those 60 patients into two clusters[44]. Of note, PAM divides the patients only on their comorbidity patterns and is not informed of the hospitalization data. The results showed that the two clusters generated (named A and B; respective sizes=9 and 51) yielded by the 215 comorbidities have different risk of hospitalization (p=0.004; Mann Whitney U test). This validates that the discovered patterns (the 215 comorbidities) are indeed predictive of the hospitalization rate. Of note, the PAM analysis and Mann-Whitney tests were repeated for models built from different OR and FDR cutoffs and lead to the same order of results (data not shown).

Discussion

Comorbidities have been known to be a key factor of ED admission and hospitalization among COPD patients; however, previous studies were constructed on regression models. This study showed that the network topology may contribute to hospital prediction study using comorbidity feature among COPD patients. Some of the features were consistent with previous studies. Hypertension, adult-onset diabetes, mental illnesses and substance abuse have all been previously associated to increase COPD patients' ED visits and hospitalizations rates[46], and evidence is shown in central hubs of Figure 2 panels I-II.

Meanwhile, falls and pregnant-related comorbidities have not been shown in the previous studies on COPD comorbidities. While falls are the leading cause of ED visits and hospitalizations due to an injury in the elderly population, to our knowledge the increased OR of COPD patients' fall related-hospitalizations had not been previously reported[47,48]. However, since age increases both risk of COPD and that of falls, additional studies are required to identify if the risk of falls in COPD patients is higher than that of a matched control. Furthermore, no studies have been described where pregnancy is a comorbid condition of COPD from our understanding. Since this study was extracted from Medicaid data, we included pregnant population due to the inclusion criteria of Medicaid[49]. Pregnancy Risk Assessment Monitoring System study conducted by the Centers for Disease Control and Prevention (CDC) showed that half of respondents (50.2%) were enrolled in Medicaid at any point from preconception through pregnancy and delivery[49]. Therefore, pregnancy-related comorbidities that show in our analysis may not be related with COPD. Because of this, we must interpret the conclusions from the network analysis carefully.

The strength of this study lies in using claims data, which captures a large number of diseases (n=754) as recorded through medical claims, to construct the model on hospitalization among COPD patients compared to previous studies that used a pre-selected list of about twenty to thirty comorbidities. Consequently, we discovered previously unreported comorbidities associated to hospitalization of COPD patients. Administrative data has become an essential research resource since it is easily accessible, contains heterogeneous information of a large population,

and can be promptly available. Claims data has been used to predict hospitalization and re-hospitalization in other studies[50]. CCI or Elixhauser can also be applied to administrative data; however, since they already select certain number of features, it may limit the possibility of finding new patterns or results.

Furthermore, compared to the traditional descriptive table, the network topology uncovered distinctive patterns of multiple comorbidities related COPD hospitalization and their co-dependency. The network-approach may inform clinicians which comorbidities need to be treated intensively to reduce the comorbidities. Focusing on the hub would decrease the risk of other comorbidities which are connected to it. In the network of metabolism syndrome, diabetes mellitus and heart failure are the hub of the network; making an effort to manage that hub would reduce the risk of diseases of heart, lung or pulmonary which decreases the frequency of medical utilization and economic burden. However, to construct the network, we need to be cautious about the threshold. Depending on the threshold, the network topology would vary and affect the quality of information. Too low of a threshold leads to the building of a hair-ball type of network that prevents the sharing of informative messages. Otherwise, too high of a threshold will threaten the loss of hub and may provide information that is not clinically actionable.

Future studies and Limitation

The generalizability of this study to other COPD patients is limited due to the short period of data collection time. Furthermore, we limited this proof-of-concept study to identifying valid "features" of the classifier (the co-morbidities) with a judicious model-free validation strategy. In future studies, a fully-specified classifier is required for individual patient prediction, inclusive of "features," mathematical model, and computed weighted parameters. We also intend to extend the type of features with ten years of unstructured clinical narratives (radiology reports, discharge summaries and pathology reports) that we coded in UMLS with MedLEE[51] and more relationships in SNOMED[52] to improve the predictive power and compare the accuracy rates to those reported in the literature by alternate approaches. Further studies need to include heterogeneous variable types such as severity of comorbidity, length of stay or demographic information as well as severity of comorbidities over time.

Further, this proof-of-concept study was designed to address the hypothesis that network analyses could identify co-morbidity features predictive of future hospitalizations in clinical datasets, while mitigating the curse of dimensionality that plague other feature discovery methods (such as FDR over multiple comparisons). However, future studies are required to identify to cop are the accuracy of predictions of features discovered by network analyses against those found by classical methods such as statistical regression or vector space discrimination (e.g. Support vector machine).

Conclusion

The proposed network modeling of COPD hospitalized patients unveils several sub-networks of comorbidities. The descriptive information with ICD-9 code only is insufficient to reveal the underpinnings of biologic or pathophysiologic connections of comorbidities. This network topology may show the possibility of revealing the co-dependency of comorbidities, which has been buried in the traditional prevalence study that links to those connections. Further, in high dimension datasets, reducing features by design while controlling for multiplicity of comparisons decreases the statistical power of conventional approaches. Finally, we propose that network analysis of clinical comorbidity and their dependencies provides an unbiased and straightforward predictor development that merits further investigation in order to prevent future hospitalization and ED visits for COPD patients.

Acknowledgement

Patient Outcome Research Institute (PCORI) "PATient Navigator to rEduce Readmissions" (PArTNER) grant (http://www.pcori.org/pfaawards/patient-navigator-to-reduce-readmissions-partner/). JK and YAL are funded in part by The University of Illinois at Chicago Center for Clinical and Translational Science NIH/NCAT UL1TR000050 and 1UL1RR029879 grants and The Office of the Office of the Vice-President for Health Affairs of the University of Illinois Hospital and Health Science Center. Role of the Sponsor: None of the funding sources had a role in the design and conduct of the study; in the collection, analysis, and interpretation of the data; or in the preparation, review, or approval of the manuscript.

References

1. Holguin F, Folch E, Redd SC, Mannino DM. Comorbidity and mortality in COPD-related hospitalizations in the United States, 1979 to 2001. *CHEST Journal.* 2005;128(4).
2. Lin P-J, Shaya FT, Scharf SM. Economic implications of comorbid conditions among Medicaid beneficiaries with COPD. *Respiratory medicine.* 2010;104(5):697-704.
3. Wier LM, Elixhauser A, Pfuntner A, Au DH. Overview of Hospitalizations among Patients with COPD, 2008. 2011.
4. Murphy SL, Xu J, Kochanek KD. Deaths: final data for 2010. *National vital statistics reports.* 2013;61(4):1-118.
5. Miravitlles M, Ferrer M, Pont A, et al. Characteristics of a population of COPD patients identified from a population-based study. Focus on previous diagnosis and never smokers. *Respiratory medicine.* 2005;99(8):985-995.
6. COPD Foundation. Impact of COPD on Health Care Costs. 2012; http://www.copdfoundation.org/pdfs/Impact%20on%20Costs.pdf.
7. Austin PC, Stanbrook MB, Anderson GM, Newman A, Gershon AS. Comparative ability of comorbidity classification methods for administrative data to predict outcomes in patients with chronic obstructive pulmonary disease. *Annals of epidemiology.* 2012;22(12):881-887.
8. Fan VS, Curtis JR, Tu S-P, McDonell MB, Fihn SD. Using quality of life to predict hospitalization and mortality in patients with obstructive lung diseases. *CHEST Journal.* 2002;122(2):429-436.
9. Chatila WM, Thomashow BM, Minai OA, Criner GJ, Make BJ. Comorbidities in chronic obstructive pulmonary disease. *Proceedings of the American Thoracic Society.* 2008;5(4):549-555.
10. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of chronic diseases.* 1987;40(5):373-383.
11. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. *Medical care.* 1998;36(1):8-27.
12. Sharabiani MT, Aylin P, Bottle A. Systematic review of comorbidity indices for administrative data. *Medical care.* 2012;50(12):1109-1118.
13. Barabási A-L, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics.* 2011;12(1):56-68.
14. Hidalgo CA, Blumm N, Barabási A-L, Christakis NA. A dynamic network approach for the study of human phenotypes. *PLoS computational biology.* 2009;5(4):e1000353.
15. Coventry PA, Gemmell I, Todd CJ. Psychosocial risk factors for hospital readmission in COPD patients on early discharge services: a cohort study. *BMC pulmonary medicine.* 2011;11(1):49.
16. Strimmer K. fdrtool: a versatile R package for estimating local and tail area-based false discovery rates. *Bioinformatics.* 2008;24(12):1461-1462.
17. van Rossum G, de Boer J. Interactively testing remote servers using the Python programming language. *CWI Quarterly.* 1991;4(4):283-303.
18. Team RC. R: A language and environment for statistical computing. 2012.
19. Lopes CT, Franz M, Kazi F, Donaldson SL, Morris Q, Bader GD. Cytoscape Web: an interactive web-based network browser. *Bioinformatics.* 2010;26(18):2347-2348.
20. Wang X, Quek HN, Cantor M, Kra P, Schultz A, Lussier Y. Automating terminological networks to link heterogeneous biomedical databases. *Studies in health technology and informatics.* 2003;107(Pt 1):555-559.
21. Pantazatos SP, Li J, Pavlidis P, Lussier YA. Integration of neuroimaging and Microarray Datasets through Mapping and Model-Theoretic semantic Decomposition of Unstructured phenotypes. *Cancer informatics.* 2009;8:75.
22. Cantor MN, Lussier YA. Mining OMIM™'for Insight into Complex Diseases. *Studies in health technology and informatics.* 2004;107(Pt 2):753.
23. Bales ME, Lussier YA, Johnson SB. Topological analysis of large-scale biomedical terminology structures. *Journal of the American Medical Informatics Association.* 2007;14(6):788-797.
24. Li H, Lee Y, Chen JL, Rebman E, Li J, Lussier YA. Complex Disease Networks of Trait--Associated SNPs Unveiled by Information Theory. *Journal of the American Medical Informatics Association.* 2012;19(2):295-305.
25. Chen JL, Hsu A, Yang X, et al. Curation-free biomodules mechanisms in prostate cancer predict recurrent disease. *BMC medical genomics.* 2013;6(Suppl 2):S4.
26. Chen JL, Li J, Kiriluk KJ, et al. Deregulation of a Hox protein regulatory network spanning prostate cancer initiation and progression. *Clinical Cancer Research.* 2012;18(16):4291-4302.
27. Cantor M, Sarkar I, Gelman R, Hartel F, Bodenreider O, Lussier Y. An Evaluation of Hybrid Methods for Matching Biomedical Terminologies: Mapping the Gene Ontology to the UMLS®. *Studies in health technology and informatics.* 2003;95:62.
28. Yang X, Li J, Lee Y, Lussier YA. GO-Module: functional synthesis and improved interpretation of Gene Ontology patterns. *Bioinformatics.* 2011;27(10):1444-1446.
29. Goh C-S, Gianoulis TA, Liu Y, et al. Integration of curated databases to identify genotype-phenotype associations. *BMC genomics.* 2006;7(1):257.

30. Liu Y, Li J, Sam L, Goh C-S, Gerstein M, Lussier YA. An integrative genomic approach to uncover molecular mechanisms of prokaryotic traits. *PLoS computational biology.* 2006;2(11):e159.
31. Yang X, Huang Y, Crowson M, Li J, Maitland ML, Lussier YA. Kinase inhibition-related adverse events predicted from< i> in vitro</i> kinome and clinical trial data. *Journal of biomedical informatics.* 2010;43(3):376-384.
32. Sarkar IN, Cantor MN, Gelman R, Hartel F, Lussier YA. Linking biomedical language information and knowledge resources: GO and UMLS. Paper presented at: Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing2003.
33. Lee Y, Li H, Li J, et al. Network models of genome-wide association studies uncover the topological centrality of protein interactions in complex diseases. *Journal of the American Medical Informatics Association.* 2013:amiajnl-2012-001519.
34. Gardeux V, Achour I, Li J, et al. 'N-of-1-pathways' unveils personal deregulated mechanisms from a single pair of RNA-Seq samples: towards precision medicine. *Journal of the American Medical Informatics Association.* 2014:amiajnl-2013-002519.
35. Sam LT, Mendonça EA, Li J, Blake J, Friedman C, Lussier YA. PhenoGO: an integrated resource for the multiscale mining of clinical and biological data. *Bmc Bioinformatics.* 2009;10(Suppl 2):S8.
36. Lussier Y, Borlawsky T, Rappaport D, Liu Y, Friedman C. PhenoGO: assigning phenotypic context to gene ontology annotations with natural language processing. Paper presented at: Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing2006.
37. Chen J, Sam L, Huang Y, et al. Protein interaction network underpins concordant prognosis among heterogeneous breast cancer signatures. *Journal of biomedical informatics.* 2010;43(3):385-396.
38. Chen JL, Li J, Stadler WM, Lussier YA. Protein-network modeling of prostate cancer gene signatures reveals essential pathways in disease recurrence. *Journal of the American Medical Informatics Association.* 2011;18(4):392-402.
39. Regan K, Wang K, Doughty E, et al. Translating Mendelian and complex inheritance of Alzheimer's disease genes for predicting unique personal genome variants. *Journal of the American Medical Informatics Association.* 2012;19(2):306-316.
40. Gamazon ER, Im H-K, Duan S, et al. Exprtarget: an integrative approach to predicting human microRNA targets. *PLoS One.* 2010;5(10):e13534.
41. Yang X, Lee Y, Fan H, Sun X, Lussier YA. Identification of common microRNA-mRNA regulatory biomodules in human epithelial cancer. *Chinese Science Bulletin.* 2010;55(31):3576-3589.
42. Yang X, Huang Y, Chen JL, Xie J, Sun X, Lussier YA. Mechanism-anchored profiling derived from epigenetic networks predicts outcome in acute lymphoblastic leukemia. *BMC bioinformatics.* 2009;10(Suppl 9):S6.
43. Lee Y, Yang X, Huang Y, et al. Network modeling identifies molecular functions targeted by miR-204 to suppress head and neck tumor metastasis. *PLoS computational biology.* 2010;6(4):e1000730.
44. Kaufman L, Rousseeuw PJ. *Finding groups in data: an introduction to cluster analysis.* Vol 344: John Wiley & Sons; 2009.
45. Reynolds K, Muntner P, Fonseca V. Metabolic syndrome underrated or underdiagnosed? *Diabetes care.* 2005;28(7):1831-1832.
46. Baty F, Putora PM, Isenring B, Blum T, Brutsche M. Comorbidities and burden of COPD: a population based case-control study. *PloS one.* 2013;8(5):e63285.
47. Wu S, Keeler EB, Rubenstein LZ, Maglione MA, Shekelle PG. A cost-effectiveness analysis of a proposed national falls prevention program. *Clinics in geriatric medicine.* 2010;26(4):751-766.
48. Stevens JA, Baldwin GT, Ballesteros MF, Noonan RK, Sleet DA. An older adult falls research agenda from a public health perspective. *Clinics in geriatric medicine.* 2010;26(4):767-779.
49. Prevention CfDCa. Pregnancy Risk Assessment Monitoring System Report on CDC' s Winnable Battles. 2012; http://www.cdc.gov/prams/PRAMSReport.html.
50. He D, Mathews SC, Kalloo AN, Hutfless S. Mining high-dimensional administrative claims data to predict early hospital readmissions. *Journal of the American Medical Informatics Association.* 2014;21(2):272-279.
51. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. *Journal of the American Medical Informatics Association.* 2004;11(5):392-402.
52. Lussier Y, Rothwell D, Côté R. The SNOMED model: a knowledge source for the controlled terminology of the computerized patient record. *Methods of information in medicine.* 1998;37(2):161-164.

Automatic Detection of Dilated Cardiomyopathy in Cardiac Ultrasound Videos

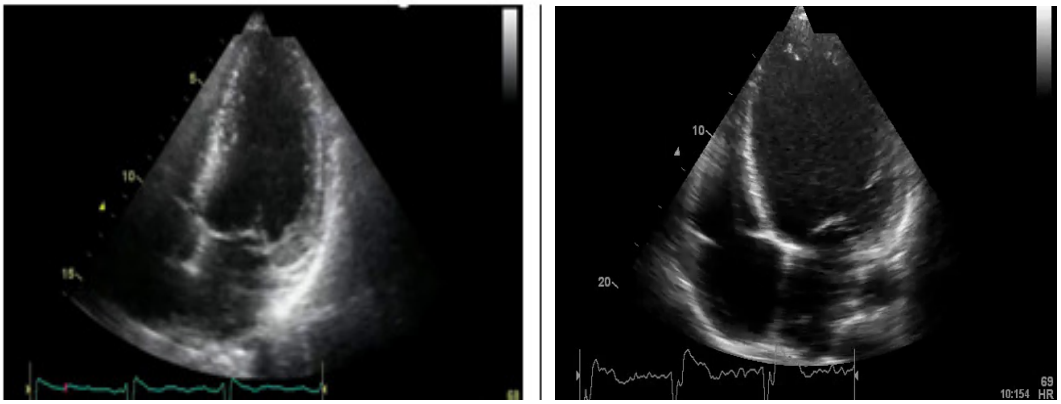
Raziuddin Mahmood[1], Tanveer Syeda-Mahmood, PhD[2],
[1]Kennedy Middle School, [2]IBM Research - Almaden, San Jose, CA

Abstract

In this paper we address the problem of automatic detection of dilated cardiomyopathy from cardiac ultrasound videos. Specifically, we present a new method of robustly locating the left ventricle by using the key idea that the region closest to the apex in a 4-chamber view is the left ventricular region. For this, we locate a region of interest containing the heart in an echocardiogram image using the bounding lines of the viewing sector to locate the apex of the heart. We then select low intensity regions as candidates, and find the low intensity region closest to the apex as the left ventricle. Finally, we refine the boundary by averaging the detection across the heart cycle using the successive frames of the echocardiographic video sequence. By extracting eigenvalues of the shape to represent the spread of the left ventricle in both length and width and augmenting it with pixel area, we form a small set of robust features to discriminate between normal and dilated left ventricles using a support vector machine classifier. Testing of the method of a collection of 654 patient cases from a dataset used to train echocardiographers has revealed the promise of this automated approach to detecting dilated cardiomyopathy in echocardiography video sequences.

Introduction

Dilated cardiomyopathy is a heart disease in which the left ventricle of the heart becomes very large and loses the ability to pump blood to the rest of the body [1]. This condition can lead to heart failure and death. Cardiomyopathy is diagnosed through an ultrasound recording. In this recording, an ultrasonic probe is moved over the heart region. Sound passes through the blood tissues and gets reflected from different parts of the body. These reflections are recorded as signal and converted to an image. The ultrasound device images the heart from different angles and depicts the chambers of the heart in views such as the apical 4-chamber views, 2 chamber views, etc. A typical cardiac ultrasound image in apical four-chamber view appears as shown in Figure 1b.



(a) (b)

Figure 1. Illustration of an ultrasound image depicting the left ventricle. (a) normal left ventricle, (b) dilated left vetricle.

In Figure 1b, the 4 chambers of the heart are seen, with the largest region being left ventricle. The shape and function of the left ventricle are important in characterizing the heart. Damage to the ventricle's shape and structure affects the function of the heart and is often seen in several diseases such as aneurysms, cardiomyopathies and infarctions. The left ventricle is often enlarged and oddly-shaped in case of dilated

cardiomyopathy [1]. This can be seen by comparing the image of Figure 1b with a normal left ventricle shown in Figure 1a which has a more normal bullet-like shape. The enlargement can be noted not only visually in Figure 1b, but also by mapping the pixels back to cm measurements using the calibration scale in the echocardiogram image. It can be seen that the left ventricle length (from apex to mitral valve) is over 90mm which usually indicates dilated cardiomyopathy in both males and females.

The goal of our work was to develop ways to automatically differentiate between normal and dilated left ventricles in echocardiography images. Such methods could aid in the development of computer-based diagnostic tools to aid clinicians in their decision making.

Automatic detection of dilated cardiomyopathy from cardiac ultrasound videos, however, is a difficult problem. In 4-chamber views, although the left ventricle is more clearly visible, the exact boundary of the left ventricle may be difficult to delineate which alters shape measurements. Using simple thresholding techniques that look for low intensity regions as potential candidate regions for left ventricles may not be sufficient as the low intensities near the apex may cause the left ventricle region to be merged with the background. Although the practice guidelines indicate the single measurement of left ventricle length for diagnosing dilated cardiomyopathy [1], errors in automated left ventricle boundary detection may require more measurements be used to provide robustness.

In this paper, we present a new method of robustly locating the left ventricle by using *the key idea that the region closest to the apex in a 4-chamber view is the left ventricular region*. For this, we (a) locate a region of interest (ROI) containing the heart in an echocardiogram image using the bounding lines of the viewing sector to locate the apex of the heart, (b) select low intensity regions as candidates, and (c) find the low intensity region closest to the apex as the left ventricle, and finally (d) refine the boundary by averaging the detection across the heart cycle using the successive frames of the echocardiographic video sequence. We then extract the eigenvalues of the shape to represent the spread of the left ventricle in both length and width and augment it with pixel area to form a small set of robust features. A set of 4-chamber view echocardiogram study videos are used for training a support vector machine classifier [3] using disease labels obtained from their corresponding reports. The machine learns the separation between normal and abnormal classes based on the provided features and their labels. New 4-chamber echocardiogram videos are then processed similarly to isolate shape feature vectors, and then classified into dilated left ventricle class or normal class using the learned support vector machine. The method has been tested on a collection of 654 patient cases from a dataset used to train echocardiographers.

Related work:

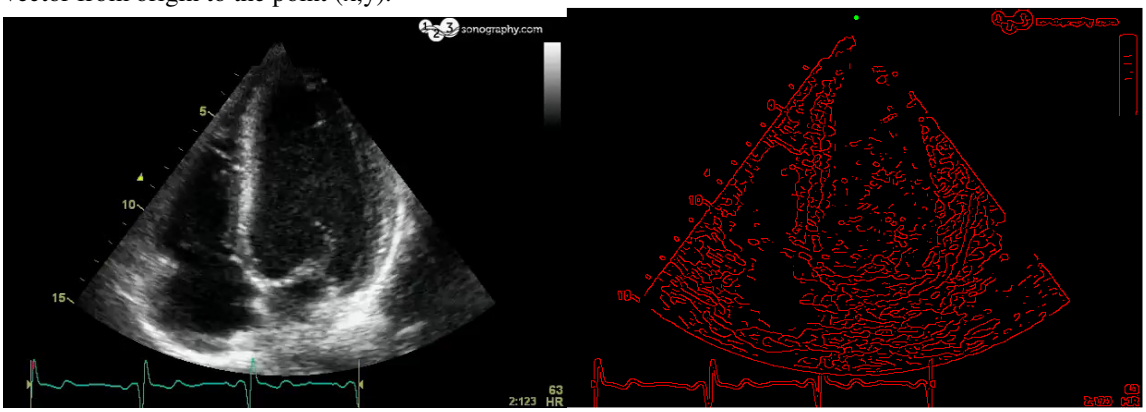
Automatic detection of diseases from echocardiography videos, however, has not been widely addressed. Most of the attention has been paid to valvular diseases from Doppler imaging [13], or measuring hypokinesia (reduced heart motion) and wall thickness in echocardiography videos [10]. Recently, work has been reported on the detection of differences between normal and abnormal left ventricular shapes in echocardiography videos using a modeling approach[12]. Our approach focuses on the detection of dilated cardiomyopathy condition, which to our knowledge has not been investigated earlier by other automated methods.

In medical imaging community, the left ventricular (LV) shape itself has been well-studied primarily for the purpose of segmenting the left ventricle in echocardiography images. A variety of techniques including active shape and appearance models [3, 4, 5], snakes and active contours [6,8], parametric shape descriptors of endocardial contours[8], deformable models and templates [6], and level set techniques are available. Model-based approaches such as active shape models are difficult to learn from a class of shapes as they need manual marker identification as well as prior registration of shapes during model training. Our experimentation also revealed that many region-based approaches over or under-segment the left ventricle, particularly, in diseased cases, resulting in inaccurate boundaries for shape characterization. We also experimented with an active shape model approach to localize LV as described in [10] but found it could locate the left ventricle accurately in only 35% of the cases of 4-chamber views. As a result, we implemented a new bottom-up approach to LV detection. Our approach consists of 4 major processing

stages, namely, (a) ROI identification, (b) LV candidate region generation, (c) LV region selection using apex, (d) LV region refinement using spatio-temporal information. Each of these processing steps are explained below.

Identifying the region of interest and apex:

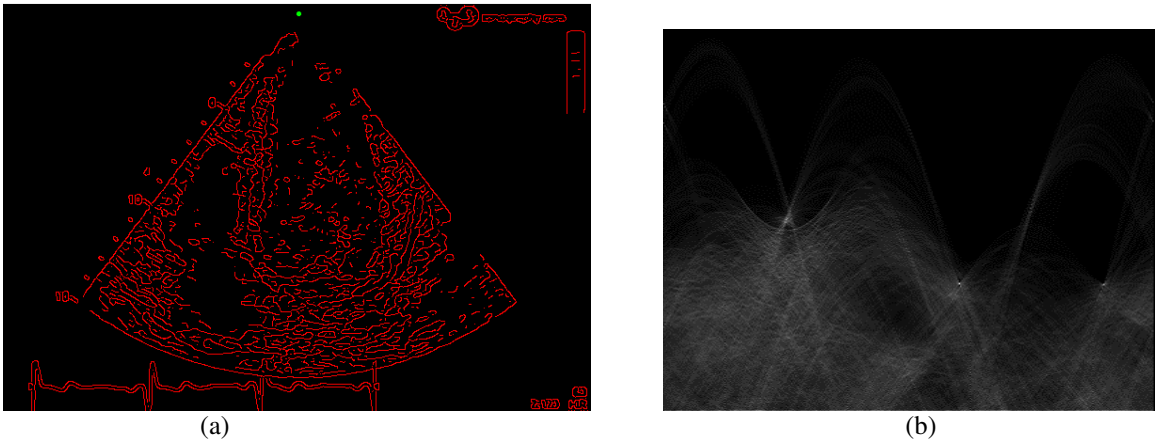
The ultrasound scan sector in the image is usually bounded by dominant lines which can be highlighted using an edge detector, such as the Canny edge detector [14]. Figure 2b shows the edge image generated for the original image of Figure 2a. The popular way to detect strong lines is through the Hough transform [15] which detects the pixels that fall on a line in polar coordinate system by converting lines to points in the Hough space of (r,d) where r is the distance of a point (x,y) from the origin and d is the angle of the vector from origin to the point (x,y).



(a) (b)

Figure 2. (a) Original image. (b) Edge image showing the potential bounding lines of the sector or region of interest. Image source: 123sonography.com

By recording the lines passing through all edge pixels in various orientations, we get the Hough image as shown in Figure 3b. Here the dominant edges seen in the edge image are noted as bright spots in the Hough transform corresponding to the number of pixels that voted for the line, with longer lines getting more votes than short line segments. Of these, the bright spots that are on a horizontal line indicate lines of the same radius and could be potential bounding lines of the sector. Further, if the angle between them as seen by the horizontal separation between the bright spots is within a reasonable angle for a viewing sector (80 to 120 degrees), then they are very likely to be the bounding lines of the sector.



(a) (b)

Figure 3. Illustration of Hough transform. (a) Edge image. (b) Hough transform rendered as an image. The horizontal axis in (b) is the angle, and the vertical axis is the radius in the polar coordinate system.

Our method exploits this observation to analyze the bright spots in the Hough transform image and project them back as bounding lines in the edge image. Verification with the pixels underneath further establishes the bounding lines of the sector. Once the bounding lines are found, the point of intersection can be easily located. Figure 3a shows the point of intersection found in the edge image by a green dot. This is our best estimate of the apex of the heart as well based on the visible content in the sector.

Identifying candidate left ventricle regions:

Since the left ventricle in the 4-chamber views is usually a darker region, we analyze the histogram of the image and separate it into intensity levels to pick relatively low intensity regions. Specifically, we use the multi-level Otsu thresholding method [9] to divide the histogram into 5 intensity levels capturing the 5 ranges of intensities typically seen in echocardiogram images. Of these the second lowest intensity level is used to threshold the original image into two classes. Choosing the second lowest ensures that we also capture cases where due to noise in imaging and phase in the heart cycle, the left ventricular region appears brighter than usual. The region within the bounding lines is then retained in the thresholded image to yield the image shown in Figure 4b. As can be seen, the potential merging of left ventricular region with the background is avoided due to the prior detection of region of interest.

(a) (b)

Figure 4: Illustration of candidate region generation within the bounding sector. (a) Original image. (b) Candidate low intensity regions within the region of interest.

Identifying the left ventricle:

We then use a connected component grouping algorithm to collect all bright pixels in the thresholded image to form candidate regions. Using the observation that the left ventricle is the closest chamber to the apex in a 4-chamber view image on the right, we obtain the distance of the centroid of each region to the point of intersection previously identified during the region of interest localization, and retain the closest rightmost region as our choice for left ventricle.

Depending on the time in the heart cycle where the echocardiogram image is taken, the left ventricular region may be merged with the left atrium (when the mitral valve is open). This can cause the left ventricular region size to be overestimated. In the next step, we integrate information across successive frames in the video to more precisely localize the boundaries of the left ventricle.

Integrating time-varying information:

We process each successive frame of the echocardiography video sequence and extract the left ventricle. The left ventricular shape is best segmented in the end-diastoli position just after the closure of the mitral valve. By tracking the size of the delineated left ventricle through the heart cycle, we pick the end-diastoli

frame as the one with the smallest size region indicated for the left ventricle. When multiple heart cycles are present in the echocardiography sequence, we average the detected size of the left ventricle across all such end-diastoli frames. Finally, we remove small holes within the LV region to form smoothly filled left ventricle region as shown in Figure 5. In this figure, Figure 5a shows a left ventricle detected among the regions of Figure 4b. Figure 5b shows the result of averaging the detection in the end-diastoli frames over the heart cycles found in the echocardiogram video sequence.

(a) (b)

Figure 5. Illustration of LV region detection. (a) Raw region from a single image. (b) average region from end-diastoli frames within the echocardiography video sequence.

Extracting shape features:

Once the LV region is identified, we extract shape features by using the eigenvalues of the 2D shape. Although the practice guidelines for dilated cardiomyopathy recommend measuring the length of the left ventricle from the apex to the mitral value, we adopted the eigenvalues over the exact extremal point-based length measurements in order to provide robustness to boundary localization errors in the left ventricular region. This can also be seen in Figure 5b where the apex region is included in the left ventricle region identification and would alter shape measurements based on the extremal points. Using the ratio of the two largest eigenvalues of the shape, we get a more robust estimation of the spread of the left ventricle in both length and width. Finally, to avoid overestimation of the area in the case of nonconvex shapes of the left ventricle (which can happen in other diseases such as aneurysms), we add a third shape feature based on the pixel area of the left ventricle.

Extracting scale:

The above measurements of shape features must be normalized for zoom effects found in the echocardiography images (when echocardiographers zoom into the regions of interest). The calibration markers found on the region of interest can help identify the scale. The scale difference can be seen in Figures 1a and 1b for the echocardiogram images of two patients from the respective calibration markers shown. In our approach, we detect the number unit markings on the calibration scale using the method described in [12] that is based on an optical character recognition (OCR) algorithm called Tesseract [16]. Using the recognized units, the distance between markers is then converted to pixel coordinates to mm ranges as described in [12]. The resulting shape features are thus normalized and presented for discrimination based on actual physical dimensions rather than the sizes in pixels.

Discriminating between normal and dilated left ventricles:

Given a set of labeled training videos depicting normal and dilated left ventricles, we locate the left ventricle as described above and extract the normalized shape feature vectors. We then find the separation between normal and dilated left ventricle using the Support Vector Machine (SVM) framework. An SVM is a classifier that, when used with two classes, tries to find a boundary in the data space such that the two classes are separated by the maximum possible margin. Given a set of training data which is composed of a set of vectors and their labels, SVM tries to find the parameters of this maximum margin boundary. This

boundary can be thought of as a high dimensional line, which is characterized by a set of weights (α). Since the objective of this work is to discriminate between normal and dilated LV, we build a single SVM model in which the positive examples (+1 labels) are dilated cases and normal cases are the negative examples (-1 labels).

Given a new test echocardiogram video, we proceed as before to extract the left ventricle region and normalized shape features. The trained SVM model is then used to predict the label for the test case as normal or dilated left ventricle indicating dilated cardiomyopathy.

Results:

We evaluated the validity of this approach on an echocardiogram dataset made available from 123sonography.com. This is an educational site for training echocardiographers in the interpretation of echocardiograms. Each video sequence is labeled with the observed condition in the video by their clinical experts. The dataset has over 2000 videos. After automatically analyzing the reports for textual phrases indicating normal or diseased left ventricle, we isolated cases of normal and dilated left ventricles. Figure 6 shows an extract of a report from which the deduction of dilated left ventricles was made and added as a positive example label for the corresponding echocardiography sequence. Raw textual analysis was followed up by manual verification before retaining the labels as ground truth labels. From the 2000 cases, we found about 254 cases of dilated cardiomyopathy and 400 normal cases. We then analyzed all the selected video sequences to locate the left ventricle and selected 124 cases as training data which had accurate detection of the left ventricle and used their shape feature vectors for training data. Of these there were 52 normals and 72 dilated cardiomyopathy cases. The trained support vector machine was then used to classify the rest of the data.

```
4261  Image view: 4 chamber view
4261  Left Ventricle: lv hypertrophy: normal | lv function:
reduced moderate to severe | lv size: dilated
4261  Wall motion / CAD: akinesia apical
4261  Right Ventricle: normal rv
4261  Atria + IAS: la normal | ra normal
4261  Mitral Valve: annular caclification
4261  Prosthesis: tv normal / prosthesis - normal
```

Figure 6. Extract of an echocardiogram report with the first column indicating the number of the corresponding echocardiography study.

Since all data tested already has a ground truth disease label of dilated cardiomyopathy or normal, the classification accuracy was evaluated as follows. Let F= (f1,f2,…fM) denote the dilated cardiomyopathy videos identified by the classifier. Let G= (g1,g2,…gN) denote the dilated cardiomyopathy cases identified in the ground for the same video. Then the classification accuracy per class is defined as $\frac{|F|}{|G|}$. The overall accuracy is averaged over the test samples of the two classes.

Training normals	Training dilated LV	Test normals	Test dilated LV	Total correct detections	Total false detections	Overall accuracy %
52	72	348	182	412	117	77.8

Table 1. Illustration of normal versus dilated LV classification accuracy by SVM.

ROI detection accuracy%	LV region selection accuracy %	LV size estimation accuracy %
92.5	89.3	87.2

Table 2. Illustration of performance of the image processing modules

The results of the evaluation are shown in Table 1. The average classification accuracy is currently at 77.8% across all the test sequences. The overall classification accuracy is affected by (a) accuracy of region of interest (ROI) localization, (b) Left ventricle detection, (c) left ventricle region size and boundary estimation besides SVM classifier accuracy reported in Table 1. The relative contribution of each of these was measured by visually inspecting the images produced in each step of the process for each of the test video sequences. The results are summarized in Table 2. As can be seen, many of these steps have good accuracy but due to the cumulative errors from their application in sequence, the overall accuracy reduces to 77.8%. As the first approach to attempt this problem, we believe these results are promising and encourage further research in this area.

Conclusions:

In this paper, we have addressed, for the first time, the problem of automatically detecting dilated cardiomyopathy from cardiac ultrasound videos. A robust left ventricle detector was proposed and shape features extracted and fed to machine learning framework based on support vector machines to separate normal from dilated left ventricles. The performance of the classification shows the promise of the method towards developing a reliable decision support tool in future.

References

[1] John S. Gottdiener et al., "American Society of Echocardiography Recommendations for Use of Echocardiography in Clinical Trials," in. J Am Soc Echocardiography 2004;17:1086-1119.

[2] 123sonography.com, http://123sonography.com/node/851.

[3] T. Cootes, A. Hill, C. Taylor, and J. Haslam. Use of Active Shape Models for Locating Structures in Medical Imaging. *Image Vision and Computing*, 12:355–366, 1994

[4] 1. J. Bosch, S. Mitchell, B. Lelieveldt, F. Nijland, O. Kamp, M. Sonka, and J. Reiber. Automatic segmentation of echocardiographic Sequences by active appearance motion models,
IEEE Transactions on Medical Imaging, 21:1374–1383, 2002.

[5] D. Linker and V. Chalana. A multiple active contour model for cardiac boundary detection on echocardiographic sequences. *IEEE Transactions on Medical Imaging*, 15:290–298,1996.

[6] N. Paragios et al., "Active shape models and segmentation of left ventricle in echocardiography", in Scale Space and PDE Methods in Computer Vision, LNCS vol. 3459, 2005, pp 131-142.

[7] Sabuncu, M.R. et al. (2010). A Generative Model for Image Segmentation Based on Label Fusion. IEEE Transactions on Medical Imaging, 2010 Oct. 29(10):1714-29.

[8]Leung KY, Bosch JG. "Segmental wall motion classification in echocardiograms using compact shape descriptors," Acad Radiol. 2008 Nov;15(11):1416-24

[9] N. Otsu, "A threshold selection method from gray-scale histogram," IEEE Trans. Syst., Man, Cybern., 9(1):62–66, 1979.

[10] T. Syeda-Mahmood et al., "Characterizing spatio-temporal patterns for disease discrimination in cardiac echo videos," in Proc. MICCAI 2007, pp.261-269.

[11] C.J. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," Data Mining and Knowledge Discovery 2:121–167, 1998.

[12] T. Syeda-Mahmood et al., "Discriminating between normal and abnormal left ventricular shapes in four-chamber view 2D echocardiography," in Proc. IEEE International Symposium on Biomedical Imaging (ISBI), 2014.

[13] T. Syeda-Mahmood et al., "Shape-based similarity retrieval of Doppler images for clinical decision support, " in Proc. CVPR, pp. 855-862, 2010.

[14] Canny, J., A Computational Approach To Edge Detection, IEEE Trans. Pattern Analysis and Machine Intelligence, 8(6):679–698, 1986.

[15] Duda, R. O. and P. E. Hart, "Use of the Hough Transformation to Detect Lines and Curves in Pictures," *Comm. ACM, Vol. 15*, pp. 11–15,1972.

[16] Weisstein, Eric W., "Tesseract", *MathWorld*.

p-medicine: A Medical Informatics Platform for Integrated Large Scale Heterogeneous Patient Data

J. Marés[1], L. Shamardin[1], G. Weiler[2], A. Anguita[3], S. Sfakianakis[4], E. Neri[5], S.J. Zasada[1], N. Graf[6], P.V. Coveney[1*]

[1]University College London, London, United Kingdom; [2]Fraunhofer Institute for Biomedical Engineering, St. Ingbert, Germany; [3]Universidad Politécnica de Madrid, Madrid, Spain; [4]ICS-FORTH, Heraklion, Greece; [5]Custodix, Sint-Martens-Latem, Belgium; [2]University of Saarland, Saarland, Germany;

Abstract

Secure access to patient data is becoming of increasing importance, as medical informatics grows in significance, to both assist with population health studies, and patient specific medicine in support of treatment. However, assembling the many different types of data emanating from the clinic is in itself a difficulty, and doing so across national borders compounds the problem. In this paper we present our solution: an easy to use distributed informatics platform embedding a state of the art data warehouse incorporating a secure pseudonymisation system protecting access to personal healthcare data. Using this system, a whole range of patient derived data, from genomics to imaging to clinical records, can be assembled and linked, and then connected with analytics tools that help us to understand the data. Research performed in this environment will have immediate clinical impact for personalised patient healthcare.

Introduction

Secure access to patient data, coupled to analysis tools running on that data, promises to revolutionise the clinical decision making process for the treatment of a wide range of diseases. The problem of sharing clinical data presents a major hurdle to overcome if patient specific data analytics and computer-based modelling are to be developed for medical research purposes and, ultimately, incorporated into clinical practice.

Because of this, the data sources held by hospitals represent a major resource that is currently not adequately exploited, either by researchers or clinicians. Digitised patient data collected as part of routine clinical practice, and which can be used as input to a wide range of analytics techniques, initially resides in information systems based within the individual hospital where the data is acquired. These data include medical images obtained through techniques such as magnetic resonance imaging (MRI) or computed tomography (CT), biopsy microphotographs, DNA and RNA sequence data, proteomics, metabolomics, and other kinds of medical records.

The data held by clinical data systems can be used in at least two different ways: (1) to compose large, (pseudo)-anonymised datasets from multiple sources, in order to perform inference based machine learning and to structure and support clinical trials; (2) to run workflows in support of clinical decision making processes on individual patients.

To make this personal data available to scientists and clinicians, we have developed a data warehouse as part of the EU FP7 p-medicine project, coupled to a medical informatics platform, into which data sources from multiple hospitals can be aggregated to generate substantially larger data collections upon which more comprehensive data analytics can be performed. The benefit of this informatics platform is clear: not only do our solutions allow data from diverse sources to be linked and integrated, they also provide a common platform that scientists and clinicians can use to initiate analytics workflows based on that data, as they offer standards compliant interfaces and APIs into which many existing and future tools and services can be plugged. The flexibility of the triplestore approach (described below) also allows us to model data in new and innovative ways. These capabilities are key to meeting the needs of the new and rapidly growing field of personalised medicine. In this paper, we describe the components of our medical informatics platform, as well as how external tools can be interfaced to the warehouse to upload, download and analyse data. In our view, the flexibility and generic nature of this platform make it applicable to a whole range of medical informatics problems. It permits immediate biomedical and clinical research to be conducted. Its novelty resides in its ability to seamlessly integrate large scale heterogeneous medical

* Corresponding Author. E-mail: p.v.coveney@ucl.ac.uk.

data in a secure, federated and distributed environment, spanning all levels from local, regional to national and international scales.

Healthcare Information Management Systems

There are many different approaches to building electronic health record and hospital information systems. Systems such as PatientCentre from iSOFT[1] are deployed within an individual hospital to manage patient data, although the GP2GP system[2] in the UK allows GP practices to transfer records amongst themselves. There are emerging standards for integrated electronic health record systems, such as HL7[3], which facilitate data transfer between systems. Online "cloud" based systems such as Microsoft HealthVault[4] allow individual patients to store and manage their health records via an online service.

Clinical data management systems are designed to manage data coming from clinical trials, and thus are often used to federate data from multiple administrative domains. Systems such as the IBM Cognos platform[5] provide business intelligence services to pharmaceutical and life sciences companies conducting clinical trials. Microsoft's Amalga platform[6] brings historically disparate data together and makes it easy to search and gain insight from that data.

However, while these systems are designed to manage and integrate large amounts of data (potentially all of the patients treated in a hospital), they do not generally deal with sharing patient data for research purposes between multiple institutions, including the ones located in different countries. In addition, traditional electronic health care record systems do not go beyond basic data management to provide advanced analysis and decision support capabilities, which rely on high performance computing platforms currently out of the control of the administrators of the data management system.

The caBIG project[7] has developed an informatics platform designed to share basic research, clinical trials imaging data, and biobanking samples between researchers, and allow them to use the data to run analytics techniques. However, caBIG's approach has been viewed as being too broad and technology led, and not designed around the needs of clinical users and researchers[8].

tranSMART[9] is an interesting recent development in the area of clinical research data management, and has found significant use in the pharmaceutical industry. tranSMART is a knowledge management platform that enables scientists to develop and refine research hypotheses by investigating correlations between genotypic and phenotypic data, and assessing their analytical results in the context of published literature and other work. However, tranSMART does have a number of limitations: it is focused around clinical records and molecular data and it does not integrate imaging data, nor does it provide access to high performance computation for data analytics.

i2b2 provides a scalable computational framework to address the bottleneck limiting the translation of genomic findings and hypotheses in model systems relevant to human health. New computational paradigms and methodologies are begin developed and tested as part of the i2b2 project in several disease cases. The server side of the software provides file and data repository facilities, as well as ontology and identity management tools. Plug-ins allow system capabilities to be expanded, with high performance computing capabilities for data analysis for example. Users access the system from client workbench and web portal tools.

System Design Overview

Our platform is organized inside a secure framework in order to ensure the confidentiality of the stored clinical data. Inside this framework there are five main components: the upload tools, the clinical trials manager (ObTiMA, described in more detail below), the data warehouse, the ontology annotator/data translator and the pseudonymisation services. Figure 1 shows a diagram of the system's design.

Upload tools and clinical trials manager components are the ones that interact with the world outside the secure framework but, at the same time, ensure that the connections with the inner components are secure. The data warehouse is the core component that stores all the data and makes it available through a public API that can be accessed only by secure connections. The ontology annotator/data translator is the component that takes complex data uploaded to the data warehouse and adapts it to conform to standard formats and protocols in order to make it easier to use and transforms it to a common vocabulary in order to enable the semantic integration of the data. Finally, the secure framework in which the system is embedded is responsible for authentication, authorization and accounting across the informatics platform. In addition, it ensures that data is pseudonymised when it is pushed into the data warehouse.

The following sections describe in detail the data ingestion workflow, the data warehouse, the pseudonymisation process, the integration of data and ontology markup and the use of the informatics platform.

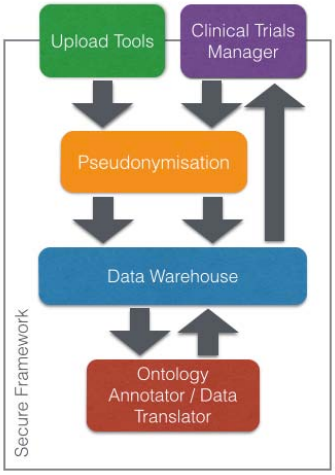


Figure 1. System design overview diagram. All traffic is secured by the authentication and authorization services in the secure framework. In addition, all data pushed into the data warehouse must pass through the pseudonymisation process to assure the privacy of the data stored in the warehouse. Finally, when data to be annotated is uploaded to the data warehouse, it triggers the ontology annotator and the data translator modules.

Data Ingestion Workflow

Health data from patients is collected by the treating physicians and stored within the treating hospitals. This data is exported by uploading it to the data warehouse. Data can also be entered into the platform's Clinical Trial Management System, ObTiMA (described in a following section). From there, data can then also be further transferred to the data warehouse. Researchers and users within the network of trust (the research domain) only have access to the *de facto* anonymous data in the data warehouse. Re-identification is only possible through the Trusted Third Party (TTP) towards the hospital.

The process of uploading new data in the data infrastructure consists of several steps: export of the data from primary sources, transformation of the data to a data format that is supported by the "upload infrastructure", pseudonymisation of the data to remove or obscure the patient identifiers, semantic annotation to attach "meaning" to the uploaded data, and semantic translation to transform the uploaded information to the semantic framework and its set of supported ontologies. The general flow of the data between the main components in the "data push" scenario is shown in Figure 2.

As in the well-known "Extract-Transform-Load" (ETL)[10] process, we can identify three main stages in the overall "upload data" scenario:

(i) Export of data from their original data sources. Our platform supports three different ways of exporting data from their sources. In the first case the data are already available on the user's computer where the user has full access to the files to be updated and the only foreseen complexity is to ensure that the data formats are supported by the upload infrastructure either directly or through a suitable transformation tool. In the second case, the data, in a multitude of formats, are stored within a "system" that allows the user to export them. In the third and most challenging case the data are stored in a Relational Database where the users need to know about the internal details of the source system, its relational schema, and must define in Structured Query Language (SQL) what is to be exported. In the latter case there are certain concerns raised with respect to the security and stability of the primary data providing systems (e.g. Hospital Information Systems) when accessed directly during their mission critical operation in the treatment domain. For these reasons the proposed architecture does not include in its design generic adapters or gateways for retrieving the information stored in the original data sources. Rather, such special data export adapters can be provided on case-by-case basis depending on the operational characteristics of a given data providing system.

(ii) Format of the uploaded data. In general the data format can be domain specific (e.g. DICOM images) and either structured or unstructured. For the raw binary formats such as medical images, the uploaded format can

be identical to the original one after the pseudonymisation process. For structured data, Comma Separated Values (CSV) can be used, a plain text format that stores tabular data. In the case of a relational database where multiple tables are to be exported, the adoption of the CSV format requires that multiple CSV files be created. Although there are certain, more expressive alternatives such as XML (Extensible Markup Language) and RDF, the CSV format provides a "lowest common denominator" for the exchange of clinical data and enjoys the best support within the pseudonymisation infrastructure; moreover, it is "natively" supported by spread sheet applications and all relational database management systems[11].

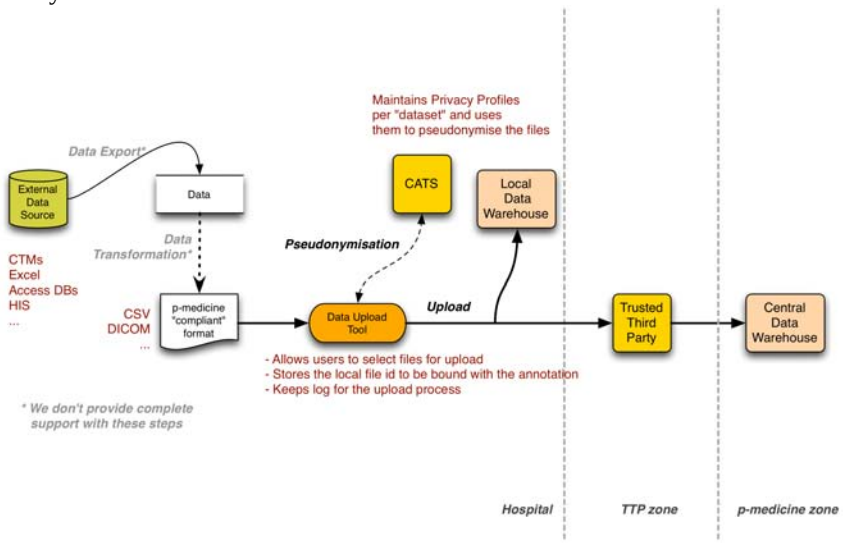


Figure 2. The data upload process. Data is exported in a p-medicine "compliant" format before being pushed into the local data warehouse and the central data warehouse via the trusted third party.

(iii) Data anonymisation and upload. The Data Upload tool is a desktop GUI application for "pushing" the data through the security infrastructure to their final destination, which is the data warehouse. It provides a graphical user interface for the push services, a user-friendly tool hiding all the complexity behind its easy-to-use interface. Its main functionality is to allow a user to load files containing patient data in a variety of formats, including CSV or DICOM, perform a first round of anonymisation through the use of the Custodix Anonymization Tool Services (CATS), see the Pseudonymisation section below), and then upload the data to the data warehouse..

Data Warehouse

The data warehouse (DWH) provides the main information storage system at the core of our platform. It consists of three main components: the triplestore, the filestore and the imagestore. DWH is built to address the central notion of reproducible research. The key concept that is implemented to make research reproducible is the automatic versioning of triplestore contents: any modification of the triplestore creates a new and unique version of the triplestore (a modification of the triplestore is also called a transaction). All responses to the triplestore queries contain explicit information on the triplestore version used to fulfil the request. Any of the triplestore versions can be queried, but cannot be modified at any time.

Semantic data held in the triplestore originate from the "raw" source files, which may be stored in the DWH filestore. There is a well-defined process of extraction of semantic data from some of the common file formats; this data is transformed after extraction using the data annotation and data translation tools, and the semantically translated data is added to the DWH triplestore. Changes to the annotation of the source data are immediately reflected in the triplestore by rolling back the transaction containing the old data originating from this file's extraction and translation chain, and by creating a new transaction with new data.

Image data upload is not supported by the DWH; however, they are accessed via DWH since it acts as a protocol proxy between the DWH user and the associated Picture Archiving and Communication System (PACS) servers. Imaging metadata is automatically transformed into raw semantic form which can be used as input for the data annotation and translation chain (explained in a following section), equipped with matching triplestore update semantics similar to the filestore.

A data warehouse can be deployed centrally in the research domain or locally in the treatment domain. In the central data warehouse (CDW) the data has been pseudonymised twice (the second pseudonymisation performed by the TTP) and is protected by a contractual framework and access controls, meaning the data can be regarded as *de facto* anonymous and can be used for research. In a local data warehouse (LDW) the data is pseudonymised only once, with the pseudonymisation key kept by the hospital. That means re-pseudonymisation of the data in the LDWs and therefore reuse of that data for treatment purposes is possible, whereas for data stored in a CDW this is only possible under very limited circumstances. From the technical point of view, LDWs and CDWs are identical.

Triplestore, Filestore, Imagestore

The DWH triplestore API permits querying the triplestore contents using SPARQL queries. It also allows plain export of the statements matching the given criteria (subject, predicate and/or object match filters). Updates to the triplestore can be made either by direct upload/deletion of specified statements, or by using the SPARQL UPDATE language. Raw data and the final processed triples added to the triplestore are tightly coupled as the user is allowed to update the annotation description at any time and then the old triples generated by this file can be updated to the new format at any time. In addition, to maintain the versioning control, the new triples are put in a new version of the triplestore together with all other triples already existent. Each triple may appear in some triplestore version exactly once, so the concept of modification rollback (transaction rollback) requires a more precise definition.

Filestore manages the storage of any type of file into DWH. Its API allows uploading and downloading a file just by providing the appropriate URL with the file identifier. In addition, when uploading a new file, whether a comma-separated value (CSV) file or an Access database, a mechanism to extract triples from the file is triggered and they are stored in the triplestore.

Finally, the imagestore manages the DICOM (Digital Imaging and Communications in Medicine) image storage while its API allows two different actions: downloading of DICOM image files and accessing the extracted triples pertaining to a DICOM image file. This is achieved by simply providing the appropriate image ID in the request URL.

Pseudonymisation

There is a clear distinction between the health treatment domain and the research domain. Medical data contain sensitive information about many patients' health and wellness. For research purposes, such data must be used in anonymised form whenever possible for legal and ethical reasons. Anonymisation is the best way to protect a patient's privacy. The platform should allow the re-identification of a patient when the research results reveal that a certain therapy would be highly effective for that given patient. Therefore data cannot be fully anonymised.

The secure framework is thus based on *de facto* anonymised data. Pseudonymous data can be regarded as *de facto* anonymous data[12,13] whenever the researchers working with the data do not possess the key linking it back to the patient and there are protection measures in place that prevent the researchers from trying to re-identify patients.

Where authorisation is given to upload data to the data warehouse, it is the clinician who triggers the transmission of the respective medical data to the warehouse. The data is first pseudonymised at the source, i.e. the hospital, and only then uploaded through the Trusted Third Party (TTP) into the data warehouse. The TTP anonymises the data through a second pseudonymisation round. The TTP, which does not dispose of any health data, also serves as a vault containing the link back to the patient if needed. The use of a TTP also assists in linking data from the same patient emanating from multiple sources.

Researchers and users within the network of trust (the research domain) only have access to the anonymised data in the data warehouse. Re-identification is only possible through the TTP and is provided unidirectionally to the source hospital.

In the current version of our health informatics platform, we make use of Custodix Anonymisation Tool Services (CATS), a set of tools and services responsible for the de-identification or anonymisation of patient data files, provided by the Belgian company Custodix, which acts as TTP. CATS anonymises or pseudonymises a data file based on a set of pre-configured transformation rules through so-called privacy profiles. The adequate definition of the transformation functions to be applied to an input file involves a thorough risk assessment. This is largely a manual task but, once defined, CATS can handle data without much effort. Once the transformations have been defined on a generic data model, all that the data uploader needs to do is map the data to that generic model. This two-step approach allows for uniform processing of data in different formats. It is also more convenient for setting up a project and provides a higher assurance level with respect to compliance.

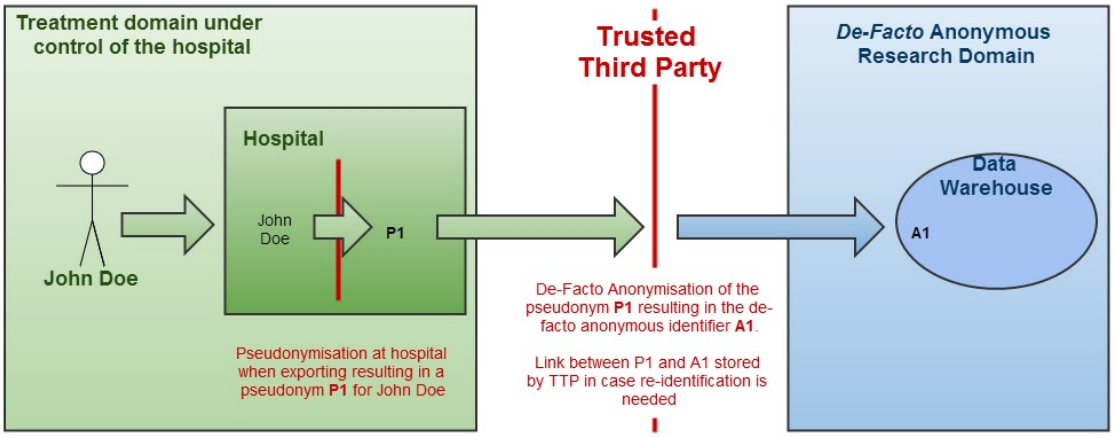


Figure 3. Pseudonymisation overview. Each user receives a pseudonym when data is pushed from the hospital. Then it is sent to the trusted third party (TTP) and it generates an anonymous identifier for this user, which is the one that is stored in the data warehouse. However, the TTP stores the link in case re-identification is needed.

Pseudonymisation functions typically generate a pseudonym based on a patient's identifying information (such as ID, name, date and place of birth). CATS therefore uses the Patient Identity Management System (PIMS) to issue pseudonyms. PIMS tries to assign the same pseudonym to the same patient by checking whether that patient has already been registered in the common PIMS database. A new pseudonym will only be assigned when there is no such patient registered so far. If the patient had already been registered the existing pseudonym will be attributed to him/her. Thus PIMS avoids the creation of different pseudonyms for the same patient (synonyms) as well as the creation of the same pseudonyms for different patients (homonyms).

Ideally, though, identifying information should never leave the source. Therefore CATS encrypts individual identifying attributes (name, data of birth, address, etc.) before sending them to PIMS. Due to the nature of cryptographic algorithms, very similar attributes (e.g. typographical errors) will be transformed to different encrypted values. Encryption does not maintain the similarity between records. For this reason, fault-tolerant matching (implemented by the matching engine in PIMS) on individual attributes is not possible. It is important to keep in mind that there is still a risk of re-identification when using encrypted attributes. Through statistical or frequency analysis techniques, re-identification of (parts of) encrypted attributes can still be achieved. To tackle the problem of matching encrypted records PIMS uses Q-grams[14] and bloom filters[15].

A Q-gram is a word with length Q that is a substring of a given word. Q-grams are used in fault-tolerant matching of words. The biggest disadvantage of this algorithm is the massive amount of subsets that are generated and encrypted at the source. Each encrypted subset must be sent to the matching service that will match with all previously received subsets. This requires substantial resources (network, storage, memory, etc.). A bloom filter is a compact data structure that allows checking for whether a given element is part of a collection, without the need to store the complete collection. Due to the compact representation false positives are possible. A bloom filter could indicate that an element is part of a collection but in reality it is not. On the other hand, false negatives can never occur. Bloom filters can be applied to investigate whether a given substring is part of a given word, without showing the complete word. To overcome the disadvantages of the Q-gram technique, we use bloom filters to match encrypted data. The algorithm itself is similar, but the collection of Q-grams for a given attribute is encrypted in a bloom filter instead of a collection of subsets.

Integration of Data and Ontology Markup

To apply meaning to the data in the data warehouse, and perform reasoning across datasets, all uploaded data is annotated with an ontology. To do this, we have coupled a data annotation tool, called Ontology Annotator (OA) with our data warehouse. The OA is a web-based tool for annotating the databases prior their integration in the DWH. The annotations generated with the OA consist of the semantic alignment of databases with HDOT (Health Data Ontology Trunk). This tool is aimed at end-users–scientists and clinicians as well as database administrators–who must have some comprehension of basic database concepts, but do not necessarily have expertise in the RDF paradigm. The OA receives as input the RDF schema of a database and provides a graphical interface that represents

both the schema and the HDOT ontology elements in a graph-based representation. The output is an XML-based serialization of the semantic equivalences of elements of the database with elements of HDOT, as defined by the user. The resulting annotation is submitted to the DWH, and associated to the corresponding database. This information is later used to translate the database to an HDOT-compliant form at the entity and attribute levels.

The OA relies on a view-to-view alignment approach, as opposed to the classical element-to-element approach. This means that database annotations store pairs of semantically equivalent views, instead if single elements. While there can be found works that already apply this approach[16,17], these are restricted to tabular format, and ours is the first one that employs it with RDF sources, as described elsewhere[18]. The atomic elements mapped in our alignment format are RDF views, instead of classes or properties. This results in a novel capability to solve cases of semantic heterogeneity which cannot be handled by other approaches. As a consequence, a database annotation is formed by a set of pairs of RDF views, one belonging to the annotated database, and one to HDOT, which we refer to as *entries*. Figure 4 shows the structure of the annotations in the semantic layer of our platform.

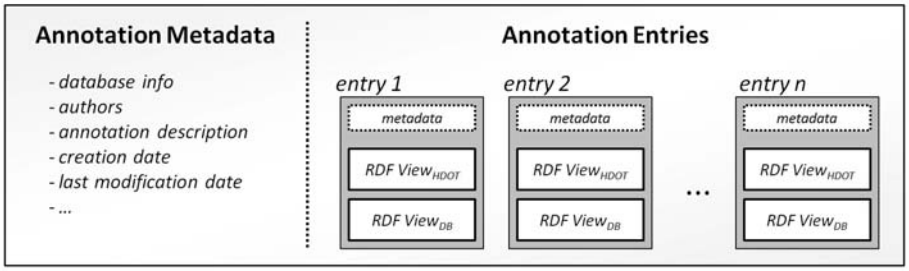


Figure 4. Structure of the annotations defined for each database. Semantic equivalences are defined with the set of entries, each containing an equivalence of a database view and an HDOT view.

This approach has been successfully applied in experiments with real data, where two heterogeneous datasets (one from the SIOP clinical trial, and one with miRNA data from some of the patients if this trial) where homogenized and merged. The former dataset contained clinical information of patients, with attributes such as age, the presence or absence of relapse, and patient identifier. The latter dataset was formed by a CSV file with 848 columns for each described miRNA molecule expression, and one column for the patient identifier. This was automatically translated to RDF by the DW. The two datasets schemas were annotated and properly aligned with the HDOT ontology. In the SIOP database, the RDF view for accessing the patients' identifiers (formed by two RDF paths, <"*patient_identifier*" → "*patient_identifier_patient_id*" → "*patient*"> and <"*patient_identifier*" → "*patient_identifier_value*" → "*string*">) was aligned with an HDOT view with similar semantics (formed by a single RDF path, <"*patient*" → "*denoted by*" → "*patient identifier*" → "*has value*" → "*string*">). In the miRNA dataset, the RDF view denoting patients' identifiers ("*Row*" → "*PatientID*" → "*string*") was aligned with the same HDOT view, thus achieving the effective merge of the two datasets in the DW triplestore.

The developed approach for annotating RDF databases is extremely generic, and allows translating any possible data model to HDOT, as long as it is modeled in RDF. The limit is actually imposed by the availability of appropriate concepts and relations in the HDOT ontology, which can be easily extended if required.

View-based annotations produced by the Ontology Annotator are used as input by the data translation process to generate HDOT-based data, achieving the desired semantic integration of the information. The output of this process is a set of RDF triples which are loaded into the DWH triplestore. These triples include provenance information and a timestamp, to allow tracking the merged data in the DWH. Translation of each annotated database is performed independently from each other. The failure in the translation of one database does not affect others, as the only consequence is that the triplestore does not receive triples from that database (upon this event, the user that created the annotation is duly notified). Figure 5 depicts the translation process performed by the data translator.

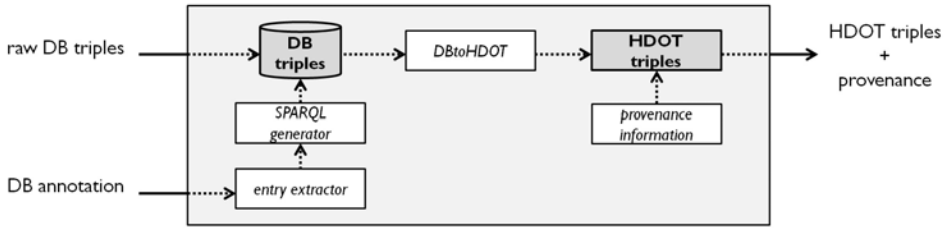


Figure 5. The data translation process uses database annotations as input. Resulting data conform to a common vocabulary provided by a cancer-related ontology. Provenance information is added to the transformed data.

Use of the Data Warehouse

The data warehouse outlined in this paper is designed to provide a flexible, robust, scalable, federated, distributed informatics platform within which to store heterogeneous medical data, ranging from clinical records, to medical images, to genomic data, in a linked, patient oriented format. In addition it has an easy use and deployment. In order to access this data, the warehouse provides a RESTful API to access, query and manipulate data. The API means that the warehouse can act as a back end to a whole range of data analytics tools, from generic statistical packages such as R, to bespoke processing tools (e.g. MINFI, LUMI, LIMMA, BWA, PICARD, SAMTOOLS, GALAXY, etc). It also makes it possible to quickly and easily tailor specific interfaces, such as web portals or mobile apps, to different groups of data warehouse users. This means that we can respond rapidly to the requirements of users to couple in the interface and analytics tools required by project researchers. By way of an example, we discuss how we have coupled our data warehouse to a clinical trials management platform.

The Ontology-based Trial Management Application[19] (ObTiMA) system is intended to support clinicians in both designing and conducting clinical trials in a user-friendly way. It has the ability to integrate an ontology into the trial building process, which makes it an extremely attractive platform for managing clinical trials. The HDOT ontology has been integrated into ObTiMA and a trial chairman is enabled to design a Case Report Form (CRF) compliant with the HDOT ontology. Therefore, data that is collected during a trial can be easily mapped to terms in the HDOT ontology. In order to integrate ObTiMA with our data warehouse, two modules have been developed for ObTiMA, which enable users to exploit the features of the platform and integrate ObTiMA with the DWH, namely the sync and the push services. Sync services enable the reuse of data stored in hospital information systems in running multicentric trials in ObTiMA. They facilitate clinical research by enabling a single entry of data for both research and healthcare avoiding redundant data entry and maximize the use of information from healthcare for research purposes. Push services have been developed in order that data collected via ontology-based CRFs in ObTiMA can be pushed into the DWH and provided in a format compliant with HDOT ontology, fully automatically without any manual preprocessing or annotation steps.

Sync Services

The sync services enable the reuse of data stored in hospital information systems in running multicentric trials in ObTiMA. In such trials several hospitals, with different heterogeneous hospital information systems (HISs), are involved, where the patient data is stored that can be reused. In such settings many authors have endorsed reusing electronic health record data in clinical trial management systems to enhance clinical trial processes and especially to avoid redundant data entry into these systems[20,21]. There are several projects that aim to link electronic health record (EHR) and clinical trial management systems (CTMS)[22,23], focusing especially on approaches to avoid redundant data entry. One of the main obstacles is the lack of semantic interoperability between different HIS systems and between HIS systems and trial management systems, partly resulting from a lack of harmonized semantic standards in the areas of health care and clinical research. The ObTiMA sync services solve these problems by utilizing the platform's semantic layer and retrieving data from the DWH. This approach has the advantage that it does not require for each new trial a full manual mapping of the data models in the EHR onto the CRFs of the trial as required in current approaches. Furthermore, it does not put restrictions on the hospital information systems used, such as compliance to standard data sets that are in general not fulfilled by current systems.

The relevant data flow for the sync services is shown in Figure 6 (left). In order to enable the reuse of data from the HIS in each hospital a local data warehouse (LDW) needs to be installed into which the data from the HIS is pushed. During this step the data is pseudonymised and semantically integrated. The sync services retrieve data from the LDWs into ObTiMA. Since the data is already integrated compliant with HDOT, the data can be mapped semi-automatically into the ontology-based CRFs of the clinical trial. Although the data in the LDW is pseudonymised, it is ensured by the Patient Identity Management Services (PIMS) of the platform's security infrastructure that appropriate data for patients in ObTiMA can be found in the LDW. The retrieved HIS data with the information as to which CRF item they are mapped into and information about its origin is shown to the clinician for confirmation. Only data originating from a hospital information system or electronic health record will be shown. The user then needs to review and confirm the data. The user can then either allow the data to be stored in the associated CRF or delete the data. Data that has been accepted are stored in the associated CRFs in the same way as if the user would have entered it manually.

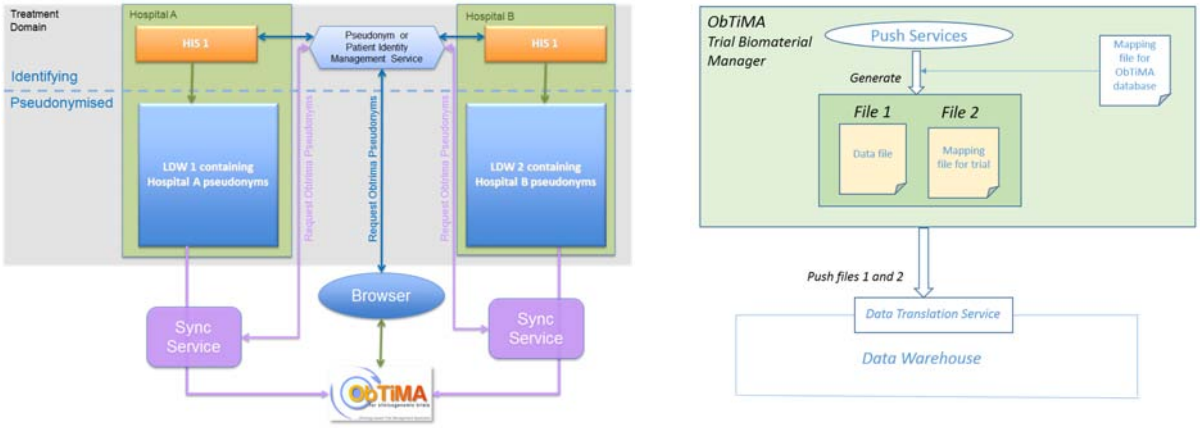


Figure 6. Data flow for sync services (left) and push services (right). The sync services retrieve data from the LDW installed in the hospital. The Patient Identity Management Service ensures that pseudonyms can be mapped to patients in ObTiMA. The push services are able to push trial data from ObTiMA either to the LDW or DWH. To translate trial data to a format compliant with HDOT the Data Translation Services are used.

Push Services

ObTiMA is seamlessly integrated with the DWH via push services whereby the data collected in ontology-based CRFs is ingested into the DWH and provided in a format compliant with the HDOT ontology without any manual pre-processing or annotation step. Pushing the data can be easily triggered by a trial chairman from the ObTiMA user interface. The process of pushing data from ObTiMA into the DWH is depicted in Figure 6 (right). In order to translate the data from the ObTiMA database to a format compliant with HDOT that can be stored in the DWH the Data Translation Services of the platform's semantic layer are used. These services require two input files: a "data file" and a "mapping file" for the selected trial. The "data file" contains the trial data that will be pushed in the form of RDF triples. The "mapping file" contains the ontology-annotations necessary to translate the trial data into an HDOT compliant format. The two files are automatically generated and sent to the ontology annotator and data translator tools and stored in the DWH.

Conclusion

The digitization of patient health care records, and the diversity of data sources comprising these, make imperative the development of easy-to-use, standardised health informatics platforms. The system we have presented in this paper is designed to do just that, linking pseudonymised patient data from multiple clinical sources, on which analytics and modelling tools may be applied. The flexible, distributed nature of our system makes it highly robust and scalable, and the use of pseudonymisation means that, unlike many similar platforms, should results from the analytic processess we apply to our data be found to have an impact on an individual patient, we can, via the trusted third party, feed back those results to the clinicians treating the patient.

In contrast to the similar systems presented at the beginning of this paper such as tranSMART and i2b2, our system is characterized by an advanced data warehouse, which is capable of automatically processing a variety of uploaded file formats to extract relevant data and store it in our data triples. Unlike other systems, our platform is designed from the ground up to integrate heterogeneous data types, including imaging, genomic and clinical records data. Our system also provides unique capabilities to trace the provenance of research data and to roll back commits on the data held in our warehouse. Having a central knowledge base with such heterogeneous types of data makes the data warehouse a novel and powerful way to do research. This ability to combine data of widely varying types and perform analyses on them facilitates the opportunity to make entirely new medical discoveries.

Acknowledgements

The work reported here has been funded in part by the EU FP7 p-medicine (no FP7-ICT-2009-270089) project.

References

1. iSOFT PatientCentre. Available from: http://www.isofthealth.com/en- GB/Solutions/UKCentre.aspx.
2. GP2GP. Available from: http://www.connectingforhealth.nhs.uk/systemsandservices/ gpsupport/gp2gp
3. R. Dolin, L. Alschuler, C. Beebe, P. Biron, S. Boyer, D. Essin, E. Kimber, T. Lincoln, J. Mattison. The HL7 clinical document architecture. Journal of the American Medical Informatics Association Vol. 8, Num. 6 (2001) pp. 552-569
4. Microsoft HealthVault. Available from: http://www.healthvault.com/personal/index.aspx
5. IBM Cognos. Available from: http://www-01.ibm.com/software/data/cognos/clinical-trial-management-software.html
6. Microsoft Amalga. Available from: http://www.microsoft.com/en- us/microsofthealth/products/microsoft-amalga.aspx
7. D. Fenstermacher, C. Street, T. McSherry, V. Nayak, C. Overby, M. Feldman. The cancer biomedical informatics grid (caBIG TM). Engineering in Medicine and Biology Society. IEEE-EMBS (2005), pp. 743-746
8. A. Califano, A.M. Chinnaiyan, G.M. Duyk, S.S. Gambhir, T. Hubbard, D.J. Lipman, L.D. Stein, J.Y. Wang, O.T. Bartlett, C.L. Harris. An assessment of the impact of the NCI Cancer Biomedical Informatics GRID (caBIG) (2011). Available from: http://deainfo.nci.nih.gov/advisory/bsa/bsa0311/caBIGfinalReport.pdf
9. B. Athey, M. Braxenthaler, M. Haas, Y. Guo. tranSMART: An Open Source and Community-Driven Informatics and Data Sharing Platform for Clinical and Translational Research. AMIA Summit on Translational Science, Vol. 2013, (2013) pp. 6-8
10. P. Vassiliadis, A. Simitsis. Extraction Transformation and Loading. In Encyclopedia of Database Systems, Springer (2009) pp. 1095-1101
11. CSV on the Web Working Group. Available from: http://www.w3.org/2013/csvw/
12. N. Forgó. Ethical and Legal Requirements for Transnational Genetic Research. Hart Publishing (2010)
13. P.V. Coveney, V. Diaz, P. Hunter, M. Viceconti. Computational Biomedicine. Oxford University Press, (2014)
14. T. Churches, P. Christen. Blind Data Linkage Using n-gram Similarity Comparisons. In Advances in Knowledge Discovery and Data Mining, LNCS 3056, Springer (2004) pp. 121-126
15. R. Schnell, T. Bachteler, J. Reiher. Privacy-preserving record linkage using Bloom filters. In BMC Medical Informatics and Decision Making, Vol. 9, Num. 1, Biomedical Central (2009), pp. 1-11
16. C. Knoblock, P. Szekely, J. Ambite, A. Goel, S. Gupta, K. Lerman, M. Muslea, M. Taheriyan, P. Mallick. Semi-automatically Mapping Structured Sources into the Semantic Web. In The Semantic Web: Research and Applications, LNCS 7295, Springer (2012), pp. 375-390
17. R. Parundekar, C. Knoblock, J. Ambite. Discovering Concept Coverings in Ontologies of Linked Data Sources. In The Semantic Web: ISWD 2012, LNCS 7649, Springer (2012), pp. 427-443
18. A. Anguita, M. García-Remesal, D. de la Iglesia, N. Graf, V. Maojo. Toward a View-oriented Approach for Aligning RDF-based Biomedical Repositories. Methods of Information in Medicine, Vol. 53, Num. 4 (2014)
19. Ontology-based Trial Management Application. Available from: http://obtima.org
20. J. Powell, I. Buchan. Electronic Health Records should support clinical research. Journal of Medical Internet Research, Vol. 7, Num. 1, (2005)
21. H. Stenzhorn, G. Weiler, M. Brochhausen, F. Schera, V. Kritsotakis, M. Tsiknakis, S. Kiefer and N. Graf. The ObTiMA System - Ontology-based Managing of Clinical Trials. In Proceedings of the 13th World Congress on Health (Medical) Informatics (Medinfo 2010). Studies in Health Technology and Informatics Series, Vol. 160 (2010) pp. 1090-1094
22. CDISC Healthcare Link Initiative. Available from: http://www.cdisc.org/healthcare-link
23. R. Kush, L. Alschuler, R. Ruggeri, S. Cassells, N. Gupta, L. Bain, K.Claise, M. Shah, M. Nahm. Implementing Single Source: The STARBRITE Proof-of-Concept Study. Journal of the American Medical Association, Vol. 14, Num. 5 (2007), pp. 662-673

U-path: An undirected path-based measure of semantic similarity

Bridget T. McInnes, PhD[1], Ted Pedersen, PhD[2], Ying Liu, PhD[3], Genevieve B. Melton, MD[4] and Serguei V. Pakhomov, PhD[4]

[1]**Virginia Commonwealth University, Richmond, VA**
[2]**University of Minnesota, Duluth, MN**
[3]**The Advisory Board Company, San Francisco, CA**
[4]**University of Minnesota, Minneapolis, MN**

Abstract

In this paper, we present the results of a method using undirected paths to determine the degree of semantic similarity between two concepts in a dense taxonomy with multiple inheritance. The overall objective of this work was to explore methods that take advantage of dense multi-hierarchical taxonomies that are more $graph\text{-}like$ than $tree\text{-}like$ by incorporating the proximity of concepts with respect to each other within the entire $is\text{-}a$ hierarchy. Our hypothesis is that the proximity of the concepts regardless of how they are connected is an indicator to the degree of their similarity. We evaluate our method using the *Systematized Nomenclature of Medicine Clinical Terms* (SNOMED CT), and four reference standards that have been manually tagged by human annotators. The overall results of our experiments show, in SNOMED CT, the location of the concepts with respect to each other does indicate the degree to which they are similar.

1 Introduction

The automated discovery of groups of semantically similar concepts and terms is critical to improving the retrieval[1] and clustering[2] of biomedical and clinical documents, and the development of biomedical terminologies and ontologies[3]. Additionally, semantic similarity measures could be used indirectly in applications such as finding articles with similar content in PubMed[4], and clustering symptoms and disorders found in the text of clinical reports for post-marketing medication safety surveillance[5 6]. Similarity measures quantify the degree to which two concepts are similar based on their taxonomical proximity through the $type\text{-}of$ (or $is\text{-}a$) relationships that exist between them. This is often referred to as a hyponym relationship where one term's ancestral pedigree is included within that of another term. The path passing through a common descendant would link the two concepts. The $undirected\ path$ ($u\text{-}path$) measure is a method to obtain the degree of semantic similarity between two concepts in lexical resources with significant multiple inheritance and are more *graph-like* than *tree-like*. This measure quantifies this degree based on the reciprocal of the shortest path between two concepts regardless of the direction of graph traversal.

Other similarity measures use the shortest path in their calculation but require that the concepts be connected through their least common subsumer (LCS) such as the conceptual distance measure proposed by Rada, et. al[1] and subsequently implemented by Caviedes and Cimino[7], the measure proposed by Leacock and Chodorow[8], and the path measure, which we refer to as $lcs\text{-}path$, as implemented by Pedersen, et. al[9]. The $lcs\text{-}path$ measure and the measure proposed by Leacock and Chodorow[8] were initially developed with WordNet[10] as their lexical resource. Concepts in WordNet are primarily organized in a acyclic hierarchy, free from multiple inheritance, therefore the shortest path between any two concepts would contain the LCS. This is not the case though for other taxonomies such as the *Systematized Nomenclature of Medicine Clinical Terms*[11] (SNOMED CT), which is a dense multi-hierarchical taxonomy of clinical terms.

The $u\text{-}path$ measures relaxes the requirement that the path between the two concepts must go through the LCS. Figure 1 shows the individual shortest path lengths used by the $lcs\text{-}path$ and $u\text{-}path$ measures to calculate the degree of similarity between Neuropathy and Paralysis in SNOMED CT. Neuropathy and Paralysis are both disorders that may involve the peripheral nervous system. In this example, the length of the shortest path is four but increases to ten when requiring the path to be connected through the concepts' LCS.

The overall objective of this work is to begin exploring methods that take advantage of dense multi-hierarchical taxonomies. Our hypothesis is that the proximity of the concepts, regardless of how they are connected, is an indicator of

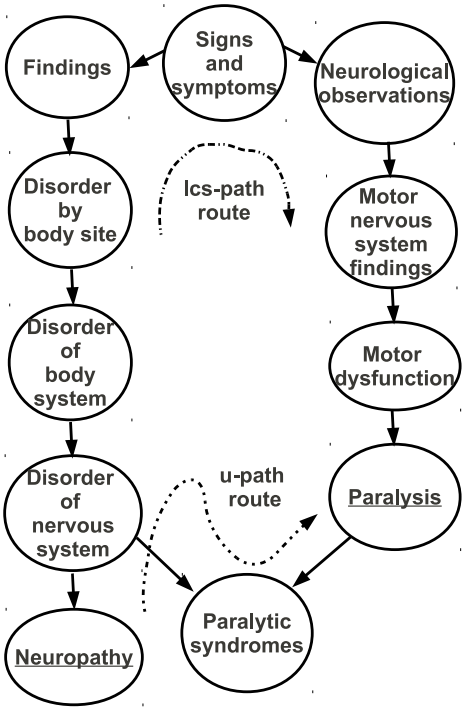


Figure 1: Example of the $lcs\text{-}path$ and $u\text{-}path$ measures

the degree of their similarity. Towards this end, we evaluate $u\text{-}path$ using SNOMED CT and four reference standards that have been manually tagged by human annotators. The results show, in SNOMED CT, the location of the concepts with respect to each other regardless of their LCS can be used to indicate the degree to which they are similar.

2 Semantic Similarity Measures

Existing semantic similarity measures can be categorized into two groups: path-based and information content (IC)-based. Path-based measures rely solely on the shortest path information, whereas IC-based measures incorporate the probability of the concept occurring in a corpus of text.

Path-based Measures

Rada, et al.[1] introduces the conceptual distance measure ($c\text{-}dist$), which is calculated as the length of the shortest path between two concepts that connects the concepts through their least common subsumer (LCS). The LCS is the most specific ancestor shared by two concepts. The length is calculated by counting the number of nodes between the two concepts. The $lcs\text{-}path$ measure is a modification of this and is calculated as the reciprocal of the length of the shortest path as shown for concepts c_1 and c_2 in Equation 1.

$$\text{sim}_{path}(c_1, c_2) = \frac{1}{\text{minpath}(c_1, c_2)} \tag{1}$$

Wu and Palmer[12] extend this measure by incorporating the depth of the LCS. In a multi- hierarchical taxonomy, we define the depth to be the minimum path between the concept and the root. In this measure, the similarity is twice the depth of the two concepts LCS divided by the sum of the depths of the individual concepts as defined in Equation 2.

$$\text{sim}_{wup}(c_1, c_2) = \frac{2 * \text{depth}(\text{lcs}(c_1, c_2))}{\text{depth}(c_1) + \text{depth}(c_2)} \tag{2}$$

Leacock and Chodorow[8] extend the path measure by incorporating the depth of the taxonomy. Here, the similarity is the negative log of the shortest path between two concepts divided by twice the total depth of the taxonomy (D) as defined in Equation 3.

$$\text{sim}_{lch}(c_1, c_2) = -\log \frac{\text{minpath}(c_1, c_2)}{2 * D} \tag{3}$$

Nguyen and Al-Mubaid[13] incorporate both the depth and LCS in their measure. In this measure, the similarity is the log of two plus the product of the shortest distance between the two concepts minus one and the depth of the taxonomy (D) minus the depth of the concepts LCS (d). Its range depends on the depth of the taxonomy.

$$\text{sim}_{nam}(c_1, c_2) = \log(2 + (\text{minpath}(c_1, c_2) - 1) * (\text{D} - \text{d}))) \tag{4}$$

Batet, et al.[14] introduce a measure that takes in account the common concepts shared (referred to as shared superconcepts) between the two concepts (c_i and c_j) and their LCS ($lcs(c_i, c_j)$. A concept's (c_i) set of superconcepts ($T(c_i)$) consist of all of the concepts found in all of the shortest paths between c_i and the LCS. In this measure the log ratio of the shared superconcepts as defined in Equation 5 where $T(c_i) = \{c_j \in C | c_j \text{is a superconcept of} c_i\}$.

$$\text{sim}_{batet}(c_1, c_2) = -\log_2 \frac{|T(c_1) \cup T(c_2)| - |T(c_1) \cap T(c_2)|}{|T(c_1) \cup T(c_2)|} \tag{5}$$

Information Content-based Measures

Information content (IC) measures the specificity of a concept in a hierarchy. The fundamental assumption with the IC measures is that the more frequent a concept is, the less specific it is. Therefore, a concept with a high IC value is more specific to a topic than one with a low IC value. IC is formally defined as the negative log of the probability of a concept (c^*) as shown in Equation 6.

$$IC(c^*) = -1 * log(P(c^*)) \tag{6}$$

The probability of a concept is determined by summing the probability of the concept ($P(c)$) occurring in some text plus the probability its descendants ($P(d)$) occurring in some text as shown in Equation 7

$$P(c*) = P(c) + \sum_{d \exists descendant(c)} P(d) \tag{7}$$

The initial probability of a concept ($P(c)$) and its descendants ($P(d)$) is obtained by dividing the number of times a concept is seen in the corpus (freq(d)) by the total number of concepts (N) as seen in Equation 8.

$$P(d) = freq(d)/N \tag{8}$$

Resnik[15] modified IC to be used as a similarity measure. He defined the IC of two concepts (c_1 and c_2) to be the IC of their least common subsumer (LCS) as seen in Equation 9. The LCS is the most specific concept two concepts share as an ancestor; if two concepts have more than one LCS, we use the most specific one.

$$\text{sim}_{res} = \text{IC}(\text{lcs}(c_1, c_2) = -\log(P(\text{lcs}(c_1, c_2))) \tag{9}$$

Jiang and Conrath[16] and Lin[17] extended Resnik's IC-based measure by incorporating the IC of the individual concepts. Lin defined the similarity between two concepts by taking the quotient between twice the IC of the concepts' LCS and the sum of the IC of the two concepts as seen in Equation 10.

$$\text{sim}_{lin} = \frac{2 * \text{IC}(\text{lcs}(c_1, c_2))}{\text{IC}(c_1) + \text{IC}(c_2)} \tag{10}$$

Jiang and Conrath defined the distance between two concepts to be the sum of the IC of the two concepts minus twice the IC of the concepts' LCS. We modify this measure to return a similarity score by taking the reciprocal of the distance as seen in Equation 11.

$$\text{sim}_{jcn} = \frac{1}{\text{IC}(c_1) + \text{IC}(c_2) - 2 * \text{IC}(\text{lcs}(c_1, c_2))} \tag{11}$$

3 Method

As discussed in Section 2, $lcs\text{-}path$ is the reciprocal of the length of the shortest path between two concepts in a hierarchy in which the shortest path is calculated by first finding the LCS of the two concepts and then aggregating the distance of all paths that connect the concepts through the LCS. In the $u\text{-}path$ measure, the paths are not required to contain the LCS and may meander through the hierarchy. We view $u\text{-}path$ as a measure of similarity because it is based strictly on path information found in $is\text{-}a$ relations, but we have relaxed the requirement that the path between the two concepts must go through the LCS.

The $u\text{-}path$ measure is related to the measure proposed by Hirst and St. Onge[18] (hso) that, like $u\text{-}path$, quantifies the strength of similarity between two concepts based on their closeness within a hierarchy. The measure assumes that two concepts are semantically close in a taxonomy if they are connected by a path that is neither too long nor too meandering and the relations between them are either $is\text{-}a$ or $has\text{-}a$ relations. The $u\text{-}path$ measure has a similar assumption but does not put restrictions on the length or meandering of the path, only the type of relation used.

In a subsequent studies using WordNet, Budanitsky and Hirst[19] report that hso obtains similar results to the path-based measure proposed by Leacock and Chodorow[8], which is similar to the results reported by Patwardhan, et. al[20]. We believe this is because hso, like $u\text{-}path$, relies on the concepts being densely connected. Unfortunately, the disadvantage of using hso on a densely connected graph is that it requires keeping track of the length and direction changes of all of the paths between the two concepts. This becomes difficult and in some cases infeasible when using dense multi-hierarchical structures, like SNOMED CT. The $u\text{-}path$ measure presented here, does not have this limitation.

4 SNOMED CT

The Systematized Nomenclature of Medicine–Clinical Terms[11] (SNOMED CT) is a comprehensive clinical terminology created for the electronic representation of clinical health information and is one of the terminology sources in the Unified Medical Language System (UMLS) Metathesaurus. The 2010AB version of the Metathesaurus contains over 1.7 million biomedical and clinical concepts from over 100 different terminology sources that have been semi-automatically integrated into a single resource. The terminology sources in the Metathesaurus can be treated independently or in combination with other sources. Currently, SNOMED CT is the largest hierarchical terminological source in the Metathesaurus.

The Metathesaurus contains a variety of different links between concepts specifying their relationship. The two hierarchical relations, used in this study, are parent/child (PAR/CHD) and broader/narrower (RB/RN) relations. A PAR/CHD is a hierarchical relation between two concepts that has been explicitly defined by the source. An RB/RN relation is a hierarchical relation that does not explicitly come from a source but is created by the UMLS editors during the integration process. In the case of SNOMED CT, the PAR/CHD relations are strictly $is\text{-}a$ relations and the RB/RN relations contain $part\text{-}of$ and $was\text{-}a$ relations. We use the PAR/CHD relations for our experiments.

In version 2010AB, SNOMED CT contains 280,695 concepts with a PAR/CHD relation; 198,241 leaf nodes and 82,454 non-leaf nodes. The depth of the taxonomy is 34, the average depth of a non-leaf node is 9.2 and the average depth of a leaf node is 11.8. The average branching factor of a non-leaf node is 5.1 and on average each node (concept) has 51 distinct paths to the root.

5 Reference Standards

We use four reference standards [a] to evaluate the $u\text{-}path$ measure: the UMNSRS tagged for similarity, the UMNSRS tagged for relatedness, the MayoSRS tagged for relatedness and the MiniMayoSRS tagged for relatedness. We include reference standards tagged for relatedness in our evaluation in order to conduct a comparison with previous work and due to the scarcity of datasets tagged strictly for similarity. Relatedness measures quantify the relationship between two concepts that are not necessarily in a strict $is\text{-}a$ or hyponym relationship; it is domain-dependent and grounded in human perception which takes into account that two concepts may be related in other ways. For example *up* is the opposite of *down*, an *elbow* is part-of an *arm*, and a *scalpel* cuts *tissue*. In this section, we describe the reference standards and then briefly discuss some of their differences.

MayoSRS: MayoSRS, developed by Pakhomov, et al.[21], consists of 101 clinical term pairs whose relatedness was determined by nine medical coders and three physicians from the Mayo Clinic. The relatedness of each term pair was assessed based on a four point scale: (4.0) practically synonymous, (3.0) related, (2.0) marginally related and (1.0) unrelated. We evaluate our method on the mean score of the physicians and medical coders as provided by Pakhomov, et al.[21].

MiniMayoSRS: MiniMayoSRS is a subset of the MayoSRS and consists of 30 term pairs on which a higher inter-annotator agreement was achieved. The average correlation between physicians is 0.68. The average correlation between medical coders is 0.78. We evaluate our method on the mean of the physician scores and the mean of the coders' scores in this subset in the same manner as reported by Pedersen, et al.[22].

UMNSRS: UMNSRS, developed by Pakhomov, et al.[23], consists of 725 clinical term pairs whose semantic similarity and relatedness was determined independently by four medical residents from the University of Minnesota Medical School. The similarity and relatedness of each term pair was annotated based on a continuous scale by having the resident touch a bar on a touch sensitive computer screen to indicate the degree of similarity or relatedness. The Intra-class Correlation Coefficient (ICC) for the reference standard tagged for similarity was 0.47, and 0.50 for relatedness. Therefore, as suggested by Pakhomov and colleagues, we use a subset of the ratings consisting of 401 pairs for the similarity set and 430 pairs for the relatedness set which each have an ICC of 0.73.

Comparison of Reference Standards

There are primarily two main differences between the UMNSRS, MayoSRS and MiniMayoSRS reference standards. The first difference is that the scores assigned to the UMNSRS term pairs by the human annotators are on a continuous scale, where the scores assigned to the MayoSRS and MiniMayoSRS are on a four point scale.

The second difference is the range of semantic groupings of the term pairs. A semantic group is a coarse-grained grouping of the semantic types in the UMLS developed by McCray, et al.[24] to provide a coarse-grained distinction between UMLS concepts based on their semantic validity, parsimony, completeness, exclusivity, naturalness, and utility. Fifteen such semantic groups have been currently defined for the UMLS Metathesaurus concepts [b].

Over half of the term pairs consist of Disorder-Disorder term pairs for each of the reference standards, although MayoSRS and MiniMayoSRS contain the largest percentage. The second largest percentage of term pairs in the UMNSRS reference standards are Disorder-Chemical & Drug term pairs which occur only a few times in MayoSRS and MiniMayoSRS. Examples of these type of term pairs such as Obesity (C0028754) and Orlistat (C0076275) which the annotators found more similar/related than the term pairs Glipizide (C0017642) and Haemophilia (C0684275). The third largest percentage of term pairs in the UMNSRS reference standards are Chemical & Drug term pairs which do not occur in MayoSRS or MiniMayoSRS. The MayoSRS and MiniMayoSRS reference standards contain the most

[a] http://rxinformatics.umn.edu/SemanticRelatednessResources.html
[b] http://semanticnetwork.nlm.nih.gov/SemGroups/

diverse term pair groupings, eleven and eight respectively, where the UMNSRS reference standards only contain the three. Table 1 shows a breakdown of the semantic groups for each of the reference standards.

Table 1: Semantic Groupings of Term Pairs in the Reference Standards

Semantic Group		MayoSRS	MiniMayoSRS	UMNSRS	
Term 1	Term 2	relatedness	relatedness	similarity	relatedness
Activities & Behaviors	Phenomena	1			
Anatomy	Anatomy	1	1		
Chemical & Drug	**Chemical & Drug**			**77**	**82**
Chemical & Drug	Devices	1			
Chemical & Drug	Procedures	1	1		
Disorder	Anatomy	4	2		
Disorder	**Chemical & Drug**	**10**	**1**	**113**	**126**
Disorder	Concepts & Ideas	3	1		
Disorder	**Disorder**	**66**	**21**	**211**	**222**
Disorder	Devices		1		
Disorder	Physiology	5			
Disorder	Procedures	7	1		
Physiology	Physiology	1			
Total		101	30	401	430

6 Experimental Framework

We conducted our experiments using the freely available open source software package UMLS::Similarity [25] version 1.13 [c]. This package takes as input two terms or concepts and returns the similarity between any two concepts using the path information in any of the sources available in the UMLS, including SNOMED CT, for each of the measures discussed in Section 2.

For our experiments, the path information was obtained using the PAR/CHD relations between concepts in SNOMED CT from the 2010AB version of the UMLS. This work was initiated in 2010 and for the sake of continuity and comparability of results, the 2010AB version was used throughout.

We calculated the IC of a concept for the IC-based measures using frequency information obtained from the National Library of Medicine's *UMLSonMedline* dataset. This dataset consists of concepts from the 2009AB UMLS and the number of times they occurred in a snapshot of Medline taken on 12/01/2009. The frequency counts were obtained by using the Essie Search Engine which queried Medline with normalized strings from the 2009AB MRCONSO table in the UMLS. The frequency of a CUI was obtained by aggregating the frequency counts of the terms associated with the CUI to provide a rough estimate of its frequency.

7 Results and Discussion

Table 2 shows the Spearman's Rank Correlation coefficients between the human scores and the scores obtained by $u\text{-}path$, $lcs\text{-}path$, and the measures proposed by Wu and Palmer [12] (wup), Nguyen and Al-Mubaid [13] (nam), Resnik [15] (res), Jiang and Conrath [16] (jcn) and Lin [17] (lin) for the four reference standards.

The results show that for MiniMayo and MayoSRS the $u\text{-}path$ measure obtained a higher correlation with human judgments than the other measures, but this was not the case for the UMNSRS reference standards. For both the UMNSRS tagged for similarity and relatedness, $lcs\text{-}path$ or jcn obtained the highest correlation scores.

The results also show that jcn and lin obtain a higher overall correlation on the UMNSRS reference standards tagged for similarity and relatedness than the $u\text{-}path$ measure, but this is not the case for the res measure. The difference between res and the other IC-based measures is that jcn and lin incorporate the IC of the individual concepts in conjunction with the IC of the LCS, where res just uses the IC of the LCS. We believe the lower results of res may

[c] http://search.cpan.org/dist/UMLS-Similarity/

Table 2: Spearman's Rank Correlation Results

Measure	Reference Standard				
	MiniMayoSRS		MayoSRS	UMNSRS	
	coders	physicians		sim.	rel.
u-path	**0.7142**	**0.5527**	**0.3187**	0.4776	0.2725
lcs-path	0.5341	0.3628	0.2324	**0.5182**	0.2903
wup	0.5346	0.4139	0.2344	0.4912	0.2425
nam	0.4228	0.2985	0.1461	0.3252	0.1634
res	0.5150	0.3852	0.2549	0.4737	0.2550
jcn	0.5437	0.4298	0.3145	0.5132	**0.3418**
lin	0.5524	0.4315	0.2948	0.4981	0.2909

indicate that it is the IC of the individual concepts that are providing the additional relevant information rather than the IC of the LCS.

One main difference between the UMNSRS and the MayoSRS and MiniMayoSRS datasets are the semantic groupings of the term pairs. The MayoSRS and MiniMayoSRS contain only Disorders-Symptom pairs while the UMNSRS data contains a mixture of Disorder, Symptom and Drug pairs. Table 3 shows a breakdown of the correlation results based on the term pairs' semantic groups in the UMNSRS tagged for relatedness; as well as the overall correlation score for reference. Table 4 shows the same results on the UMNSRS tagged for similarity; and Table 5 shows the results for the MiniMayoSRS and MayoSRS.

Table 3: Spearman's Rank Correlation of UMNSRS tagged for Relatedness

Semantic Groups	u-path	lcs-path	jcn
Disorder-Disorder	0.4028	0.4290	0.4383
Disorder-Chemical&Drug	-0.1037	-0.1188	0.1369
Chemical&Drug-Chemical&Drug	0.3925	0.3761	0.4356
UMNSRS rel.	0.2725	0.2903	0.3418

For the UMNSRS tagged for relatedness, the results show that $u\text{-}path$ obtains a higher correlation for Chemical & Drug term pairs than $lcs\text{-}path$ but not for Disorder term pairs. The jcn measure obtains a higher correlation than either path-based measures for each of the semantic groups.

The results also show that for $u\text{-}path$, $lcs\text{-}path$ and jcn the correlation results are low for Disorder-Chemical & Drug term pairs. We believe this is because in SNOMED CT there does not exist many $is\text{-}a$ relationships between drugs and disorders which is the relationship all of the similarity measures are exploiting. We hypothesize that relatedness measures, rather than similarity measures, maybe a better measure for these type of relations.

Table 4: Spearman's Rank Correlation of UMNSRS tagged for Similarity

Semantic Groups	u-path	lcs-path	jcn
Disorder-Disorder	0.4521	0.5055	0.4329
Disorder-Chemical&Drug	0.1449	0.1489	0.1783
Chemical&Drug-Chemical&Drug	0.5693	0.5977	0.6885
UMNSRS sim.	0.4776	0.5182	0.5132

For the UMNSRS tagged for similarity, the results show that $lcs\text{-}path$ and jcn obtain a higher correlation with humans than $u\text{-}path$ over each of the different semantic groups except for Disorder pairs in which $u\text{-}path$ obtains a higher correlation than jcn. The low correlation results for Disorder-Chemical & Drug term pairs confirms our previous analysis above that there is a limited number of $is\text{-}a$ relations between drugs and disorders for the similarity measures to exploit.

Table 5 shows the correlation results for the semantic groups in the MiniMayoSRS and MayoSRS that have at least twenty term pairs. The results show that for Disorder term pairs, the $u\text{-}path$ obtains a higher correlation with the human judgments than $lcs\text{-}path$ and jcn for these two reference standards. We believe that this indicates that the

related disorders in these reference standards are co-located in similar areas in the taxonomy and not necessarily in a direct path through the LCS.

Table 5: Spearman's Rank Correlation of MayoSRS and MiniMayoSRS

Semantic Groups for MayoSRS	u-path	lcs-path	jcn
Disorder-Disorder	0.3535	0.2649	0.2579
MayoSRS	0.3187	0.2324	0.3142
Semantic Group for MiniMayoSRS	**u-path**	**lcs-path**	**jcn**
Disorder-Disorder (physicians)	0.4223	0.2580	0.2733
Disorder-Disorder (coders)	0.5933	0.3958	0.3353
MiniMayoSRS (physicians)	0.5527	0.3628	0.4298
MiniMayoSRS (coders)	0.7142	0.5241	0.5437

To further analyze the difference between the $u\text{-}path$ and $lcs\text{-}path$, Table 6 shows the number of term pairs in each of the reference standards, and the corresponding percentage whose shortest path did not go through the LCS. The percentage of non-LCS paths in the reference standards indicate that SNOMED CT is indeed a dense multi-hierarchical structure in which the shortest path between concepts does not always contain the LCS. The overall higher percentage of non-LCS paths in the MiniMayoSRS and MayoSRS may explain why $u\text{-}path$ obtained a higher correlation with human judgments than $lcs\text{-}path$. Although, this is not seen when analyzing the semantic grouping results in the UMNSRS reference standards, where the Chemical & Drug pairs contain the lowest number of non-LCS paths and $u\text{-}path$ obtains a higher correlation than $lcs\text{-}path$. However, we believe that given the relatively large percentage of concept pairs that have a path between them shorter than the LCS path indicates that $u\text{-}path$ has certain possibilities and merits further exploration.

Table 6: Non-LCS Shortest Paths

reference standard	semantic groups	# term pairs	# non LCS paths	% of non LCS paths
MiniMayoSRS	all	29	17	0.59
MayoSRS	all	101	57	0.50
UMNSRS (relatedness)	all	430	184	0.43
	Disorder pairs	223	110	0.49
	Disorder-Chemical&Drug	125	63	0.50
	Chemical&Drug pairs	82	11	0.13
UMNSRS (similarity)	all	401	183	0.46
	Disorder pairs	212	106	0.50
	Disorder-Chemical&Drug	112	63	0.56
	Chemical&Drug pairs	77	14	0.18

8 Conclusions and Future Work

In this paper, we present the results of a method that quantifies the degree of similarity between concepts in a dense taxonomy with multiple inheritance using undirected path information. We show that it obtains a higher correlation than other path-based measures on two reference standards. We also analyze the semantic groupings of the term pairs showing that different measures perform better on different groupings. In the future, we plan to explore more fully the impact of the semantic groups of term pairs on the similarity measures.

The overall objective of this work was to explore a method that takes advantage of dense multi-hierarchical taxonomies that are more $graph\text{-}like$ than $tree\text{-}like$. Our hypothesis was that the proximity of the terms regardless of how they are connected is an indicator to the degree of their similarity. The overall results show, in SNOMED CT, the location of the concepts with respect to each other indicates the degree to which they are similar. In the future, we plan to explore more complex measures that take advantage of this information such as graph-based centrality metrics and the measure proposed by Batet et al. [14 26].

The results showed that $u\text{-}path$ obtained higher correlation results on the reference standards tagged for relatedness.

The multiple inheritance within the taxonomy implies the potential existence of ancestral pedigrees that represent different, although related, dimensions of a concept. Therefore, a path passing through a common descendant would then link two concepts that belong to two dimensions allowing for the connection between the two concepts to fall outside the traditional strict $is\text{-}a$ relations. In the future, we plan to explore this further comparing the results to relatedness measures such as those proposed by Lesk[27]; Patwardhan[28]; Dagan, et al.[29]; Workman, et al.[30]; and Pivovarov and Elhadad[31].

We also plan to explore weighting the undirected path based on its turns within the path. Hirst and St. Onge[18] limited the degree in which the path is allowed to meander based on the types of turns the path is taking. The underlying thought behind this is that the deeper the undirected path passes through a common descendant the less similar the two concepts would be, although the relatedness between the two concepts would still be maintained.

9 Acknowledgments

This work was supported in part by the National Institute of Health, National Library of Medicine (NLM) Grant #R01LM009623-01. We would like to thank Russel Loane, Jim Mork and Lan Aronson from NLM for providing the UMLSonMedline dataset.

References

1. R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1):17–30, 1989.

2. Y. Lin, W. Li, K. Chen, and Y. Liu. A document clustering and ranking system for exploring MEDLINE citations. *Journal of the American Medical Informatics Association*, 14(5):651–661, 2007.

3. O. Bodenreider and A. Burgun. Aligning knowledge sources in the UMLS: methods, quantitative results, and applications. In *Proc. of the 11th World Congress on Medical Informatics*, pages 327–331, 2004.

4. Jimmy Lin and W John Wilbur. Pubmed related articles: a probabilistic topic-based model for content similarity. *BMC bioinformatics*, 8(1):423, 2007.

5. R.K. Pearson, M. Hauben, D.I. Goldsmith, A.L. Gould, D. Madigan, D.J. OHara, S.J. Reisinger, and A.M. Hochberg. Influence of the meddra hierarchy on pharmacovigilance data mining results. *Intl. journal of medical informatics*, 78(12):e97–e103, 2009.

6. R.W. Bill, Y. Liu, B.T. McInnes, G.B. Melton, T. Pedersen, and S. Pakhomov. Evaluating semantic relatedness and similarity measures with standardized meddra queries. In *AMIA Annual Symposium Proc.*, volume 2012, page 43. American Medical Informatics Association, 2012.

7. J.E. Caviedes and J.J. Cimino. Towards the development of a conceptual distance metric for the umls. *Journal of Biomedical Informatics*, 37(2):77–85, 2004.

8. C. Leacock and M. Chodorow. Combining local context and WordNet similarity for word sense identification. *WordNet: An electronic lexical database*, 49(2):265–283, 1998.

9. T. Pedersen, S. Patwardhan, and J. Michelizzi. WordNet::Similarity - Measuring the Relatedness of Concepts. In *The Annual Meeting of the Human Language Techonology and North American Association of Computational Linguistics: Demonstration Papers*, pages 38–41, Boston, Massachusetts, USA, May 2004.

10. G.A. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):41, 1995.

11. K.A. Spackman, K.E. Campbell, and R.A. Côté. Snomed rt: a reference terminology for health care. In *Proc. of the AMIA annual fall symposium*, page 640. American Medical Informatics Association, 1997.

12. Z. Wu and M. Palmer. Verbs semantics and lexical selection. In *Proc. of the 32nd Meeting of Association of Computational Linguistics*, pages 133–138, Las Cruces, NM, June 1994.

13. H.A. Nguyen and H. Al-Mubaid. New ontology-based semantic similarity measure for the biomedical domain. In *Proc. of the IEEE Intl. Conference on Granular Computing*, pages 623–628, Atlanta, GA, 2006.

14. M. Batet, D. Sánchez, and A. Valls. An ontology-based measure to compute semantic similarity in biomedicine. *Journal of biomedical informatics*, 44(1):118–125, 2011.

15. P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proc. of the 14th Intl. Joint Conference on Artificial Intelligence*, pages 448–453, Montreal, Canada, August 1995.

16. J. Jiang and D. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. on Intl. Conference on Research in Computational Linguistics*, pages pp. 19–33, 1997.

17. D. Lin. An information-theoretic definition of similarity. In *Intl Conf ML Proc.*, pages 296–304, 1998.

18. G. Hirst and D. St-Onge. Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet: An Electronic Lexical Database*, pages 305–332, 1998.

19. A. Budanitsky and G. Hirst. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47, 2006.

20. S. Patwardhan, S. Banerjee, and T. Pedersen. Using measures of semantic relatedness for word sense disambiguation. *Computational Linguistics and Intelligent Text Processing*, pages 241–257, 2010.

21. S.V.S. Pakhomov, T. Pedersen, B. McInnes, G.B. Melton, A. Ruggieri, and C.G. Chute. Towards a Framework for Developing Semantic Relatedness Reference Standards. *Journal of Biomedical Informatics*, October 2010.

22. T. Pedersen, S.V.S. Pakhomov, S. Patwardhan, and C.G. Chute. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics*, 40(3):288–299, 2007.

23. S. Pakhomov, B.T. McInnes, T. Adam, Y. Liu, T. Pedersen, and G.B. Melton. Semantic similarity and relatedness between clinical terms: An experimental study. In *Proc. of the American Medical Informatics Association (AMIA) Symposium*, pages 572–576, Washington, DC, November 2010.

24. A.T. McCray, A. Burgun, and O. Bodenreider. Aggregating umls semantic types for reducing conceptual complexity. *Studies in health technology and informatics*, pages 216–220, 2001.

25. B.T. McInnes, T. Pedersen, and S.V. Pakhomov. UMLS-Interface and UMLS-Similarity : Open Source Software for Measuring Paths and Semantic Similarity. In *Proc. of the American Medical Informatics Association Symposium*, 2009.

26. M. Batet, D. Sánchez, A. Valls, and K. Gibert. Semantic similarity estimation from multiple ontologies. *Applied intelligence*, 38(1):29–44, 2013.

27. M. Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proc. of the 5th Annual Intl. Conference on Systems Documentation*, pages 24–26, 1986.

28. S. Patwardhan and T. Pedersen. Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts. In *Proc. of the EACL 2006 Workshop Making Sense of Sense - Bringing Computational Linguistics and Psycholinguistics Together*, pages 1–8, 2006.

29. I. Dagan, L. Lee, and F Pereira. Similarity-based models of word cooccurrence probabilities. *Machine Learning*, 34(1-3):43–69, 1999.

30. E.T. Workman, G. Rosemblat, M. Fiszman, and T.C. Rindflesch. A literature-based assessment of concept pairs as a measure of semantic relatedness. In *AMIA Annual Symposium Proc.*, volume 2013, page 1512. American Medical Informatics Association, 2013.

31. R. Pivovarov and N. Elhadad. A hybrid knowledge-based and data-driven approach to identifying semantically similar concepts. *Journal of biomedical informatics*, 45(3):471–481, 2012.

First-Order Logic Theory for Manipulating Clinical Practice Guidelines Applied to Comorbid Patients: A Case Study

Martin Michalowski, PhD[1], Szymon Wilk, PhD[2], Xing Tan, PhD[3], Wojtek Michalowski, PhD[3]

[1]Adventium Labs, Minneapolis, MN; [2]Poznan University of Technology, Poznan, Poland; [3]University of Ottawa, Ottawa, Canada

Abstract

Clinical practice guidelines (CPGs) implement evidence-based medicine designed to help generate a therapy for a patient suffering from a single disease. When applied to a comorbid patient, the concurrent combination of treatment steps from multiple CPGs is susceptible to adverse interactions in the resulting combined therapy (i.e., a therapy established according to all considered CPGs). This inability to concurrently apply CPGs has been shown to be one of the key shortcomings of CPG uptake in a clinical setting[1]. Several research efforts are underway to address this issue such as the K4CARE[2] and GuideLine INteraction Detection Assistant (GLINDA)[3] projects and our previous research on applying constraint logic programming to developing a consistent combined therapy for a comorbid patient[4]. However, there is no generalized framework for mitigation that effectively captures general characteristics of the problem while handling nuances such as time and ordering requirements imposed by specific CPGs. In this paper we propose a first-order logic-based (FOL) approach for developing a generalized framework of mitigation. This approach uses a meta-algorithm and entailment properties to mitigate (i.e., identify and address) adverse interactions introduced by concurrently applied CPGs. We use an illustrative case study of a patient suffering from type 2 diabetes being treated for an onset of severe rheumatoid arthritis to show the expressiveness and robustness of our proposed FOL-based approach, and we discuss its appropriateness as the basis for the generalized theory.

Introduction

Clinical practice guidelines (CPGs), as knowledge-based tools for disease-specific patient management[5], encapsulate evidence-based practices for devising the most appropriate treatment for patients with regard to relevant patient information and possible diagnoses. CPGs are normally created by a panel of experts and in a number of instances are computerized. However, CPGs are not designed for use on patients with comorbid diseases and who require a combined therapy (i.e., a therapy established according to all simultaneously applied CPGs). This problem was identified as one of the major shortcoming of CPG uptake in practice and as such there exists a need for research to address it[1]. Given a comorbid patient's available medical information, it is likely that the application of multiple disease-specific CPGs results in direct adverse interactions between individual medical actions that manifest as contradictory recommendations (e.g., "administer NSAID" in one CPG and "do not administer NSAID" in the other). The concurrent application of disease-specific CPGs for such patients can also result in undesired consequences due to drug-drug or drug-disease interactions, a term we refer to as indirect adverse interactions[4]. More broadly, the synthesis of two or more guidelines for treating patients with comorbidities is a challenging problem involving sophisticated design of processes for the identification and elimination of potential redundancies, contradictions, and discordances[6]. Therefore, combining CPGs in order to cross-check CPG-based medical recommendations requires the introduction of new combinatorial, logical, or semantic approaches[7].

Our recent research[4,8,9] responds to this need by introducing and formally defining a logical model of CPGs and developing a mitigation algorithm that operates on these models in order to identify and address adverse interactions. In the mitigation algorithm, clinical knowledge is encoded as interaction and revision operators within the constraint logic programming (CLP) paradigm. The operators characterize adverse interactions and describe revisions to logical models required to address encountered interactions. CLP allows one to efficiently solve these models where a solution represents a consistent combined therapy free of any adverse interactions.

More specifically, we apply our CLP-based approach to a situation when at least two CPGs are applied to a comorbid patient in order to obtain a combined therapy – a combination of at least two individual therapies derived from disease-specific CPGs. With available patient information, a combined therapy might not be consistent if there are adverse (in)direct interactions between the diseases, between medications that are applied to the patient as suggested by the disease-specific CPGs, or between diseases and prescribed medications. In these cases, CPGs and

associated therapies need to be revised using secondary clinical knowledge. This knowledge is not encoded in the guidelines themselves but comes from domain experts, textbooks, or repositories of clinical evidence.

While the CPG mitigation paradigm proposed in our earlier research is powerful enough to handle a number of different clinical scenarios associated with the management of comorbid conditions, it also has several limitations. The critical limitation is the lack of support for using temporal and associated precedence relationships between CPG actions. It is well documented that medical actions (tests or administration of medication) often follow time-dependent sequences that need to be preserved, even when new or modified actions are introduced. Our desire to capture the time dependencies commonly found in CPGs motivates the research presented in this paper that extends the CLP-based approach.

Our current research focuses on extending the CLP-based approach to handle the temporal relationships and more broadly it aims at developing a general framework for the concurrent application of CPGs. Towards this end, we present our work on representing CPGs as first-order logic (FOL) theories. FOL allows us to define CPGs as logical theories, take patient information and mitigation strategies that incorporate time and precedence relations into account, and include the modeling capabilities developed in our earlier CLP-based work. As a matter of fact, precedence relationships are only FOL definable. Further, using FOL to represent and manipulate the CPGs allows us to introduce semantics to guide the interpretation of the concepts and relieve the user of "manual" interpretations of the mitigation results – an additional limitation of our CLP-based work.

Our current methodological foundation is adopted largely from our previous work[4,8,9] and expanded as required. The conceptual framework describing this extension, using a case study based on the concurrent management of type 2 diabetes and rheumatoid arthritis, is the focus of this paper. We start by providing a brief overview of FOL and theorem proving – the methods used in our research. We then describe the methodological foundation of our proposal to model the mitigation problem as a FOL theory. Next we present a case study to ground the theory in a clinical example, and we conclude with a discussion of our contributions and potential areas for future work.

Methodology

In order to better illustrate the proposed FOL-based mitigation of guidelines, we start with brief introductions of the basic concepts and notation of FOL and theorem proving. We then present the methodological foundations of our approach.

First-Order Logic and Theorem Proving

First-order logic (FOL) is a formal system in which formulas of a formal language may be interpreted to represent propositions (particular sentences, such as "administer NSAID"). FOL distinguishes itself from propositional logic by providing additional expressive power through the use of quantified variables (for example $x > 1$ meaning all values of x must be greater than 1). With FOL, properties of objects in the domain and relationships between objects can be specified by introducing predicates and a set of inferences rules and axioms allow the derivation of theories from this language. In FOL, a theory is a collection of sentences describing some domain (i.e. CPGs). We further assume that it is possible to reason over a theory, and the result of this reasoning is referred to as a model. Thus, given a theory D and a sentence φ in the language of D, we say that φ is consistent with D if there exists a model of the theory that satisfies φ. The sentence φ can be deduced (or implied) from D if the sentence is satisfied by all models for D. Formally, we write

$$D \vDash \varphi$$

Therefore, the notion of theorem proving is a procedure to check whether a theory is satisfiable (meaning if it is possible to derive models from this theory). One way to check for satisfiability is by formulating an entailment problem (understood as a logical consequence of the sentences where one follows from the other) over D. Moving from our previous work[4,8,9] to a FOL-based approach substantially increases our ability to represent and reason over clinical knowledge by representing CPGs, patient information, and basic properties of mitigation as FOL-based theories. As part of this reasoning, we generate an entailment problem over the generated theory to check for the satisfiability of the theory. A so-called grounded instance of a satisfiable theory represents a consistent therapy.

FOL Theory for CPG Mitigation

Before introducing the main concepts and components of FOL theories employed to mitigate adverse interactions between CPGs, we briefly recall the notion of an *actionable graph* $(AG)^4$ which we use as a representation of an individual CPG. An AG expresses a single CPG in form of a directed graph composed of three types of nodes –

context, *action*, and *decision*, and arcs that correspond to transitions between nodes. A *context* node defines an entry point and indicates the disease associated with the CPG, an *action* node indicates a medical action that needs to be executed, and finally a *decision* node indicates a selection from several alternative choices and it allows for conditional branching. An AG may contain loops however, in this paper we assume an AG contains no loops, and is thus an acyclic graph. The formal definition of an AG, based on the SDA* notation, and several examples are provided in our previously published work[4].

The three key components of using FOL for CPG mitigation are: (1) a vocabulary used to construct the theory, (2) a set of "sub-theories" to describe various components of the mitigation problem, and (3) a set of operators that define the secondary knowledge needed to identify and mitigate adverse interactions in the theory. We describe each in turn below.

The vocabulary is made up of constants (denoted with lower case letters, e.g., *x*, *a*), variables (denoted with upper case letters, e.g. *X*, *Z*) and predicates. Table 1 lists core predicates that are central to the case study described later (the vocabulary includes other predicates as well, but we have omitted them for the sake of simplicity). We note there is no predicate corresponding to a context node, as information embedded in this node is provided by the predicate *disease*(*d*).

Table 1. Defined predicates

Predicate	**Explanation**
node(*x*)	*x* is a node in AG
disease(*d*)	CPG is associated with disease *d*
action(*x*)	*x* is an action node in AG
diagnosed(*d*)	disease *d* is diagnosed for the given patient
decision(*x*)	*x* is a decision node in AG
executed(*x*)	task node *x* is or has been executed
value(*x*, *v*)	value *v* is associated with decision node *x*
dosage(*x*, *n*)	task node *x* is characterized by dosage *n*
directPrec(*x*, *y*)	node *x* directly precedes node *y* (in AG there is an arc from *x* to *y*)
prec(*x*, *y*)	node *x* precedes node *y* (in AG there is a path from *x* to *y*)

A FOL theory is a collection of logical sentences constructed using the defined vocabulary. In our research we create the following theories that represent relevant components of the mitigation problem:

- D_{common} – a common theory that axiomatizes the universal characteristics of CPGs as part of a FOL-based representation for mitigation. It introduces axioms that ensure important properties of precedence and relations between various types of nodes and their characteristics. More specifically, introduced axioms ensure anti-symmetry and transitivity of precedence, uniqueness of node types (a node cannot be both an action and decision node, only an action node may be associated with medication dosage, etc.) and similar properties,
- D^{d}_{cpg} – the theory that represents an actionable graph AG (and thus the underlying CPG) for disease *d*. It encapsulates treatments for this disease, enlists all paths in the AG in the form of disjunctions of conjunctions (where each conjunction corresponds to a path), gives information about direct precedence between nodes, and finally provides information on dosages associated with selected action nodes,
- D_{pi} – the theory that represents available patient information, where each data item is given as a separate logical sentence,
- D_{mit} – the theory (initially empty) that stores new sentences introduced during the mitigation process.

These theories are used as building blocks to construct a combined theory that describes a particular mitigation instance personalized to a patient encounter. Formally, the combined theory D_{comb} is defined as the union of the theories described above:

$$D_{comb} = D_{common} \cup D_{cpg}^{d_1} \cup D_{cpg}^{d_2} \cup \cdots \cup D_{cpg}^{d_m} \cup D_{pi} \cup D_{mit},$$

where $d_1, d_2, \cdots, d_m$ are the diseases a comorbid patient suffers from and for which the patient is concurrently managed according to the associated CPGs. To further simplify the presentation in this paper yet without the loss of generality we limit the number of concurrently applied CPGs to two, thus we define the combined theory as follows:

$$D_{comb} = D_{common} \cup D_{cpg}^{d_1} \cup D_{cpg}^{d_2} \cup D_{pi} \cup D_{mit}$$

FOL-based Mitigation of Adverse Interactions

The process of mitigating (identifying and addressing) adverse interactions consists of two main phases. The first phase aims at mitigating direct adverse interactions that manifest as inconsistencies in the combined theory D_{comb}. Their identification is relatively easy as it involves checking the satisfiability of D_{comb}. If the theory is satisfiable, then there are no direct interactions. Otherwise the theory needs to be revised by applying revision operators. These *revision operators* capture expert knowledge that is not encoded in CPGs and they are formally defined later in the text.

The second phase of the mitigation process aims at identifying and addressing indirect adverse interactions. Checking the satisfiability of D_{comb} is not sufficient for identifying indirect interactions and additional expert knowledge is required. This knowledge is encoded in form of *interaction operators*. An interaction operator IO^k is formally defined as

$$IO^k = \langle \alpha^k \rangle,$$

where α^k is a logical sentence that describes a particular interaction. Checking whether a particular IO^k is applicable to D_{comb} is formulated as the entailment problem $D_{comb} \vDash \alpha^k$.

Encountered interactions (both direct and indirect) need to be addressed by applying relevant revision operators to D_{comb}. A revision operator RO^k is formally defined as

$$RO^k = \langle \beta^k, Op^k \rangle,$$

where β^k is a logical sentence that defines the applicability of the operator (the relevance of a revision operator for a particular theory given specific patient information), and Op^k describes the revisions introduced by RO^k. In particular, Op^k is a list of *n* pairs of logical expressions $\langle \varphi_i^k, \phi_i^k \rangle$ $(i = 1 \ldots n)$ that define individual operations carried out as part of applying the revision operator. To ensure that patient information is not modified and general practices of mitigation are not violated as part of the revision process, the operations Op^k are only applicable to a subset of the theories that comprise D_{comb} – namely $D_{cpg}^{d_1}$, $D_{cpg}^{d_2}$ and D_{mit}. Furthermore there are three possible types of operations in Op^k - *removal*, *addition*, and *substitution*. These types are defined as follows, where $\otimes$ represents an empty expression:

- $\langle \varphi_i^k, \otimes \rangle$ – φ_i^k is removed from any sentence that appears in $D_{cpg}^{d_1}$, $D_{cpg}^{d_2}$ or D_{mit},
- $\langle \otimes, \phi_i^k \rangle$ – ϕ_i^k is added as a new sentence to D_{mit},
- $\langle \varphi_i^k, \phi_i^k \rangle$ – φ_i^k is replaced by ϕ_i^k in any sentence that appears in $D_{cpg}^{d_1}$, $D_{cpg}^{d_2}$ or D_{mit}.

Checking the applicability of RO^k is analogous to checking the applicability to IO^k and translates to formulating and solving the entailment problem $D_{comb} \vDash \beta^k$.

In the context of FOL, a consistent combined therapy is a model of D_{comb} that represents an assignment of values to predicates.

Case Study: Management of a Patient with Type 2 Diabetes and an Onset of Severe Rheumatoid Arthritis

To ground the theoretical concepts presented above, we use the following case study. Let's consider a 70 year old male with type 2 diabetes (DB2), who suffers from symptomatic hyperglycemia and has consistently measured over 8.5 on the A1C test. As such, his sugar level is being controlled by a daily insulin dosage of 40 international units.

The patient also suffers from relatively mild rheumatoid arthritis (RA) (comorbid condition) that is being managed using a maintenance dosage of plaquenil. In the two clinical scenarios described below, we will show how the patient's therapy can be revised through the mitigation of adverse interactions using the formalism described above. For the sake of simplicity we only present the relevant interaction and revision operators and omit the details of the FOL-based theories D_{cpg}^{db2} and D_{cpg}^{ra} representing CPGs for type 2 diabetes and a rheumatoid arthritis.

Clinical Scenario 1: Managing the Administration of Glucocorticoids

In the first scenario we assume there is a sudden relapse of RA accompanied by the onset of severe pain and significantly reduced mobility. Typically in such cases, the patient is initiated on glucocorticoid treatment to control the onset of pain. However, considering that the patient is diabetic, the increased sugar level associated with the administration of glucocorticoids needs to be mitigated with an increased maintenance dosage of insulin. In this clinical scenario, the daily dosage is increased to 48 international units and maintained until sugar level stabilizes or the patient no longer requires glucocorticoid therapy.

Supporting this scenario requires the codification of secondary knowledge describing the use of and interactions with glucocorticoids. As stated in medical literature, glucocorticoids increase the blood sugar level and require the dosage of insulin in DB2 patients (if administered) to be increased by 20%. The interaction operator IO^1 represents the drug-drug interaction when a DB2 patient, also being treated for RA, is prescribed with glucocorticoids while taking insulin.

$$IO^1 = \langle \alpha^1 \rangle,$$

where $\alpha^1 = diagnosed(db2) \wedge\ diagnosed(ra)\ \wedge executed(gluco)\ \wedge executed(insulin)$.

The revision operator RO^1 modifies the theory representing this DB2 and RA patient by increasing the dosage of insulin by 20%.

$$RO^1 = \langle \beta^1, Op^1 \rangle,$$

where $\beta^1 = \alpha^1$ and $Op^1 = \langle dosage(insulin, X), dosage(insulin, X * 1.2) \rangle$. We note the definition of the operation in Op^1 employs variable *X*. Variables act as "wildcards" that can be bound to different constants to increase the flexibility of operations, thus in this case it is possible to define a substitution operation that increases the dosage of a medication by a certain ratio.

For this particular scenario we have the following subset of sentences describing the patient's state D_{pi}.

$$D_{pi} = diagnosed(db2).\, diagnosed(ra).\ executed(gluco).\ executed(insulin).$$

The following sentence is part of the theory for DB2 prescribing 40 international units of insulin to the patient.

$$D_{cpg}^{\mathrm{db2}}\ \ni dosage(insulin, 40)$$

To check for the existence of an indirect interaction we formulate the entailment problem $D_{comb} \vDash \alpha^1$. Because α^1 is consistent with D_{comb}, we know an indirect interaction exists. We then formulate the entailment problem $D_{int} \vDash \beta^1$ -- since $\beta^1 = \alpha^1$ we immediately know RO^1 is applicable to D_{comb} and we consequently apply Op^1 to substitute the dosage of insulin with a dosage that is 20% higher. This is done by removing $dosage(insulin, 40)$ from D_{cpg}^{db2} and adding $dosage(insulin, 40 * 1.2)$ to D_{mit}, which resolves to the dosage of insulin being adjusted to 48 international units (*X* is bound to 40). Note the use of variables to support different values for the same expression, rather then creating an expression for each possible value.

Clinical Scenario 2: Managing Immunosuppressive Medication

In this clinical scenario we consider the same patient that was initiated on glucocorticoid therapy supplemented with a calcium antagonist (in place of a NSAID that is not recommended for diabetic conditions). The daily insulin dosage was increased (as explained in Scenario 1) to manage the elevated sugar level, however the prescribed therapy did not work as expected. In order to better control the relapse of rheumatoid arthritis, the next therapeutic option is to put the patient on DMARD combination therapy that normally includes cyclosporine. While all immunosuppressive medications are diabetogenic (with hyperglycemia being a common adverse event), some like azathioprine proved in clinical trials to be better tolerated by type 2 diabetics. Therefore, the revised DMARD combination therapy is prescribed for this patient, and it uses azathioprine as a replacement for cyclosporine.

The interaction operator IO^2 represents the drug-disease interaction that occurs when a DB2 patient, being treated for RA, is prescribed cyclosporine as part of DMARD combination therapy.

$$IO^2 = \langle \alpha^2 \rangle,$$

where $\alpha^2 = diagnosed(db2) \wedge diagnosed(ra) \wedge executed(cyclosporine)$.

The revision operator RO^2 modifies the theory D_{comb} for the DB2 and RA patient by replacing cyclosporine with azathioprine. As we show below, this revision also maintains the precedence of executed tasks when performing the replacement.

$$RO^2 = \langle \beta^2, Op^2 \rangle,$$

where $\beta^2 = \alpha^2$ and Op^2 is the following list of logical expressions.

$$\langle node(cyclosporine), node(azathioprine) \rangle. \langle action(cyclosporine), action(azathioprine) \rangle.$$

$$\langle directPrec(methotrexate, cyclosporine), directPrec(methotrexate, azathioprine) \rangle.$$

$$\langle directPrec(cyclosporine, observ\_ra), directPrec(azathioprine, observ\_ra) \rangle.$$

To check for the existence of an indirect interaction we formulate the entailment problem $D_{comb} \models \alpha^2$ and subsequently $D_{comb} \models \beta^2$ to find the relevant revision operator. Using patient information represented in D_{pi} above, we infer that RO^2 is in fact applicable to this patient encounter and we apply the corresponding logical expressions to the theory.

We note here that we are revising the theory for the patient to replace cyclosporine with azathioprine **and** maintain the order in which tasks are to be executed according to the CPG for RA (not shown here due to space limitations). Specifically, we are ensuring that the task of administering *methotrexate* (to treat RA) is done before the administration of *azathioprine* (as was the case for cyclosporine), and that the patient's onset of RA is observed (*observ_ra*) after administering azathioprine. Similarly to the first clinical scenario, the removal of sentences is done from D_{cpg}^{ra} and sentences are added to D_{mit}.

Discussion and Conclusions

In this paper we described our preliminary research on developing a general theory of mitigation expressed in FOL for concurrently applied CPGs. This research builds upon our foundational work in using the CLP paradigm to handle mitigation and extends it by using a paradigm with greater expressive power (FOL) to handle temporal relationships such as task precedence. We used a case study of a patient suffering from type 2 diabetes while being treated for an onset of severe rheumatoid arthritis to illustrate the added benefit of our new approach. Through two simple scenarios we demonstrated the power of our FOL-based approach by applying an adjustment of medication dosage and through the substitution of tasks while maintaining the task execution order as defined by one of the CPGs.

Our proposed FOL-based approach provides an expressive and robust language in order to represent and apply temporal relationships represented in CPGs while also easily capturing common knowledge applicable to all mitigation scenarios. Furthermore, by separating the overall theory (D_{comb}) into "sub-theories" we are able to provide finer grained control over how to apply revisions to a proposed theory. Specifically, we limit the addition, deletion, and substitution of logical sentences to the theories representing CPGs ($D_{cpg}^{d_i}$) and to the theory representing added mitigation actions (D_{mit}). As such, we ensure that a patient's state is consistent since information describing a patient (D_{pi}) cannot be inadvertently changed. Additionally, the common characteristics of mitigation, applicable to all mitigation instances, are maintained in a single theory (D_{common}) and cannot be altered by any mitigation operations.

As future research, we are exploring ways to make the proposed approach more general and robust. As stated at the beginning, our ultimate goal is to develop a general framework of mitigation and towards this end we are studying various clinical situations involving comorbid patients to extract the full set of properties that hold across all mitigation scenarios. Furthermore, we are working on inductive reasoning techniques to automatically infer precedence relationships as logical operations are applied to the theories representing CPGs. The addition, deletion, and substitution of logical sentences impacts the underlying structure represented by these theories and automating

the maintenance of correct precedence relationships goes a long way to realizing our goal of using FOL-based methods to drive a point-of-care clinical decision support system.

References

1. Peleg M. Computer-interpretable clinical guidelines: a methodological review. J Biomed Inform 2013;46(4):744-63.
2. Isern D, Moreno A, Sánchez D, Hajnal Á, Pedone G, Varga LZ. Agent-based execution of personalised home care treatments. Appl Intell 2011;34(2):155-180.
3. Stanford University. GLINDA: GuideLine INteraction Detection Architecture [Internet]. [cited March 10, 2014]. Available from: http://glinda-project.stanford.edu/.
4. Wilk S, Michalowski W, Michalowski M, Farion K, Hing MM, Mohapatra S. Mitigation of adverse interactions in pairs of clinical practice guidelines using constraint logic programming. J Biomed Inform 2013;46(2):341-53.
5. Rosenfeld RM, Shiffman RN. Clinical practice guideline development manual: a quality-driven approach for translating evidence into action. Otolaryngol Head Neck Surg 2009;140(6 Suppl 1):S1-4.
6. Sittig DF, Wright A, Osheroff JA, Middleton B, Teich JM, Ash JS, et al. Grand challenges in clinical decision support. J Biomed Inform 2008;41(2):387-92.
7. Fox J, Glasspool D, Patkar V, Austin M, Black L, South M, et al. Delivering clinical decision support services: there is nothing as practical as a good theory. J Biomed Inform 2010;43(5):831-43.
8. Michalowski M, Wilk S, Lin D, Michalowski W, Tan X, Mohapatra S. Procedural approach to mitigating concurrently applied clinical practice guidelines. In: Workshop on Expanding the Boundaries of Health Informatics Using Artificial Intelligence (HIAI'13) at the Twenty-Seventh AAAI Conference on Artificial Intelligence; 2013.
9. Michalowski M, Wilk S, Michalowski W, Lin D, Farion K, Mohapatra S. Using constraint logic programming to implement iterative actions and numerical measures during mitigation of concurrently applied clinical practice guidelines. In: Peek N, Morales RM, Peleg M, editors. Artificial Intelligence in Medicine, 14th Conference on Artificial Intelligence in Medicine, AIME 2013, Murcia, Spain, May/June 2013, Proceedings. Springer; 2013. p. 17-22.

An empirically derived taxonomy of errors in SNOMED CT

Jonathan M. Mortensen, MS, Mark A. Musen, MD, PhD, Natalya F. Noy, PhD[†]
Stanford Center for Biomedical Informatics Research
Stanford University, Stanford, CA 94305-5479 U.S.A.

Abstract
Ontologies underpin methods throughout biomedicine and biomedical informatics. However, as ontologies increase in size and complexity, so does the likelihood that they contain errors. Effective methods that identify errors are typically manual and expert-driven; however, automated methods are essential for the size of modern biomedical ontologies. The effect of ontology errors on their application is unclear, creating a challenge in differentiating salient, relevant errors with those that have no discernable effect. As a first step in understanding the challenge of identifying salient, common errors at a large scale, we asked 5 experts to verify a random subset of complex relations in the SNOMED CT CORE Problem List Subset. The experts found 39 errors that followed several common patterns. Initially, the experts disagreed about errors almost entirely, indicating that ontology verification is very difficult and requires many eyes on the task. It is clear that additional empirically-based, application-focused ontology verification method development is necessary. Toward that end, we developed a taxonomy that can serve as a checklist to consult during ontology quality assurance.

Ontologies and Errors
Ontologies enable researchers to describe and reason about data and knowledge in a computable fashion using a common vocabulary. Because of these properties, ontologies are key to many biomedical tasks ranging from decision support and data integration to search and billing.[1] As a sign of the wide-spread use of ontologies, the National Center for Biomedical Ontology's BioPortal contains over 350 ontologies and terminologies describing almost any biomedical subdomain.[2] The U.S. government has required the use of terminologies such as SNOMED CT and LOINC in Electronic Health Records (EHR) as part of meaningful use.[3] Ontologies also play an essential role in biomedical research, helping to combat the data-deluge. Researchers use ontologies to perform pharmacovigilance using unstructured clinical text, to analyze gene modules, and to integrate biosimulation models,.[4–6] In the life sciences, the Gene Ontology is central to many methods (e.g., micro-array enrichment analysis), and has over 10,000 citations.[7] These examples all demonstrate the importance and diverse uses of biomedical ontologies.

Unfortunately, those same biomedical ontologies we rely on contain errors. Researchers have published reports about critical domain-specific errors in two widely-used biomedical ontologies, the National Cancer Institute Thesaurus (NCIt)[8] and SNOMED CT[9–13]. For example, Rector and colleagues manually explored SNOMED CT and identified errors such as "The foot *is-a-part* of the pelvis".[10] As ontologies increase in size, scale, and number, these domain-specific errors become even more difficult to identify or prevent. Current methods used to identify these errors are either human based or computationally based. In human-based methods, domain experts manually browse an ontology searching for errors. This approach is costly and cannot scale to the size of recent biomedical ontologies. For instance, it is likely impossible for a human to check the nearly 400,000 stated concepts in SNOMED CT, let alone the millions of logical entailments. Computational methods rely on inconsistencies within ontology syntax, structure, or semantics to find errors.[14–17] These methods scale more easily but do not detect accurately *domain-specific* errors. Considering the limitations of current methods, additional research is still needed to achieve scalable and accurate ontology verification.

Adding to these challenges of ontology verification, the real-world impact and true extent of errors is unknown. For example, it is unclear what costs are associated with an EHR that incorrectly classifies a patient due to errors in SNOMED CT. By understanding error patterns in ontologies, one can begin to identify the real-world cost of a certain pattern and prevent it. Using these patterns as a reference, engineers can develop best practices that prevent the creation of errors in the first place. In addition, these patterns can direct additional research efforts focused on the automated, scalable methods that we now require. In this work, we perform an empirical error analysis on a subset of the SNOMED CT hierarchy.

[†] Current Affiliation: Google Inc., Mountain View, CA 94043

Specifically, we make the following contributions:

1. A logic-based heuristic to identify subsets of an ontology likely to contain errors.
2. An expert-curated and validated set of relations in a subset of SNOMED CT.
3. A taxonomy describing the common types of errors. This taxonomy can serve as a checklist during verification.

Methods

To characterize the common error patterns in SNOMED CT, we first selected a subset of SNOMED CT to verify using the CORE Problem List Subset, a listing of terms provided by National Library of Medicine that are most frequent in multiple hospitals across the US, and a logic-based heuristic.[18] Next, we asked 5 experts with both ontology and medical expertise to verify the subset of relations independently. With expert votes and comments collected, we then asked experts to update their response after reviewing responses other experts in the study. The updated responses provided a final enumeration of errors from which we developed a taxonomy of common error patterns. We describe each step in detail below.

Logic-Based Heuristic Filtering

To obtain a reasonably sized set of relations for verification, we elected to filter the entire set of relations in the SNOMED CT by frequency and complexity. In doing so, we made two key assumptions: (1) real world term frequency is a proxy for importance, and (2) complex relations are more likely to contain errors. We began with the January 2013 version of the SNOMED CT ontology. We then extracted an ontology module from SNOMED CT using the CORE Problem List Subset as the signature.[10,18,19] The CORE Problem List Subset serves as the proxy for "importance". We used Snorocket, an OWL reasoner designed for SNOMED CT, to compute the concept hierarchy of this module.[20] To further narrow a set of relations likely to have errors, we selected complex SubClass (*is-a*) entailments from the classified module. A complex entailment has the following characteristics:

- non-asserted – not explicitly stated during ontology design but is a logical conclusion of stated axioms (e.g., the relations "Hypertension *is-a* Cardiovascular Disorder" and "Cardiovascular Disorder *is-a* Disorder" entails a non-asserted axiom "Hypertension *is-a* Disorder")
- non-trivial – the justification for that entailment contained at least 2 axioms
  (e.g., the non-asserted axiom "Hypertension *is-a* Disorder" from the above example is also non-trivial because it requires the interplay of 2 axioms)
- direct – there exists no concept in the inferred SubClass hierarchy between the two concepts in the relation, removing basic subsumption entailments that are not non-trivial, but still rather basic (e.g., "Hypertension *is-a* Disorder" is ***in***direct because "Cardiovascular Disorder" lies between "Hypertension" and "Disorder" in the inferred hierarchy.)

This process selects ~35,000 SubClass relations. We focus solely on SubClass relations in this work because they are the most common type of relation in biomedical ontologies.[21] To reduce further the set of relations, we required that both parent and child be explicitly in the CORE Problem List Subset, not just contained in the module, leaving ~1000 relations. From this set, we randomly sampled 200 relations for verification.

Expert Verification and Delphi

We asked 5 domain experts known directly by the authors who are have expertise in both in medicine and ontology, to verify the filtered set of 200 relations. Specifically, we presented each expert with an online survey of randomly ordered relations reformulated as natural language questions (e.g., "Diabetes is a kind of Disorder of the Abdomen. True or False?") (Fig. 1). In addition, we provided concept definitions when available from the UMLS.[22] In previous work, we developed the optimal format to ask in natural language the verification question (e.g., "is a" vs. "is a kind of") and to provide appropriate context through definitions.[21,23,24] Experts marked each relation as "True" or "False" and also justified their decision. After each expert independently completed the survey, we performed one round of the Delphi method on relations where all experts did not reach agreement (i.e., they did not all give same response). In Delphi, experts saw their responses and justifications along with those of other experts in an anonymized fashion.[25] The experts then had an option to update their response in light of other expert comments and votes.

Analysis and error classification

With the expert verification complete, we determined which relations were truly in error by supermajority voting after the Delphi round (i.e., at least 4 of the experts agreed). In addition, we calculated expert agreement using a free-marginal kappa statistic (κ).[26] The free-marginal kappa does not assume a fixed number of true or false relations, making it appropriate for our application. Finally, the authors qualitatively reviewed the errors and identified common characteristics among them in an ad-hoc fashion. From these common characteristics, we developed a taxonomy of errors.

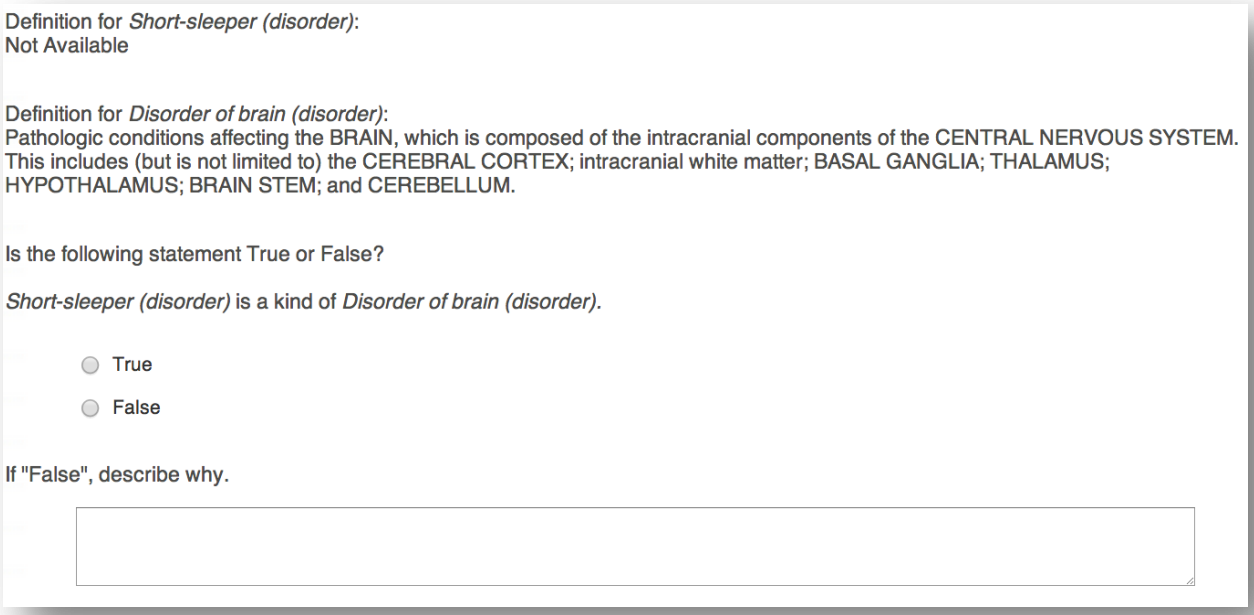


Figure 1. Screenshot of the question form that experts completed to verify relations in SNOMED CT. All 200 questions were randomly ordered per expert.

Results

The experts identified 9 errors in the 200 relations before the Delphi round by supermajority. After the Delphi round, the experts flagged 30 additional relations by supermajority. Table 2 presents all 39 errors. Most errors fall into the categories of anatomical confusion or issues related to cause and effect.

After the experts identified the errors, we developed a taxonomy that specifies the patterns of errors they encountered. Table 1 lists this taxonomy with error descriptions and expert agreement on errors in that category. Table 2 indicates where a specific error falls in this taxonomy. Before Delphi, expert agreement on errors was low (κ=-0.5); agreement increased after Delphi by a marked margin (κ =0.69). For comparison, expert κ was 0.73 before Delphi and 0.9 after Delphi on 148 correct relations (by supermajority). Experts were unable to reach consensus on 13 relations.

Error Type	Description	κ Pre-Delphi	κ Post-Delphi	Error Count
Term Ambiguity	Term representing a concept has an unclear meaning	-0.16	0.73	9
Naming Convention	Convention causes interpretation difficulty (e.g. use of a conjunction within a term)	-0.07	1	3
Modeling	Incorrect domain representation	-0.03	0.69	36
Cause and Effect	Conflation of root disorder and its effects	0.1	0.7	8
Disorder vs. Symptom	Confusion of a disorder and its symptoms	0.07	0.73	3
Disorder vs. Sequelae	Confusion of a disorder and its potential consequences	0.2	0.8	4
Multiple Etiology	Ignoring alternate etiologies of disorder	0.04	1	5
Anatomic Confusion	Improper anatomic location (e.g., regions vs. structure)	-0.1	0.69	28

Table 1. Error taxonomy constructed by observing common characteristics between 39 errors in SNOMED CT. The categories are not disjoint, and thus errors may be members of multiple classes. The κ describes expert agreement before and after Delphi and range between complete disagreement (-1) and complete agreement (1). As reference, on the 148 correct relations, κ was 0.73 Pre-Delphi and 0.9 Post-Delphi. (Error type names shortened for presentation.)

Child	Parent	Term Ambiguity	Naming Convention	Modeling	Cause and Effect	Disorder vs. Symptom	Disorder vs. Sequelae	Multiple Etiology	Anatomic Confusion
Anterior shin splints	Disorder of bone			✓					✓
Short-sleeper	Disorder of brain			✓				✓	
Frontal headache (finding)	Pain in face (finding)			✓					✓
Local infection of wound	Wound			✓	✓		✓		
Anal and rectal polyp	Rectal polyp	✓	✓	✓					✓
Malignant neoplasm of brain	Malignant tumor of head and/or neck			✓					✓
Diabetic autonomic neuropathy associated with type 1 diabetes mellitus	Diabetic peripheral neuropathy			✓					✓
Placental abruption	Bleeding (finding)			✓	✓	✓			
Impairment level: blindness, one eye - low vision other eye	Disorder of eye proper			✓				✓	✓
Gastroenteritis	Disorder of intestine			✓					✓
Microcephalus	Disorder of brain			✓					✓
Thrombotic thrombocytopenic purpura	Disorder of hematopoietic structure	✓							
Fibromyositis	Myositis			✓					✓
Lumbar radiculopathy	Spinal cord disorder			✓					✓
Vascular dementia	Cerebral infarction			✓	✓		✓		
Chronic tophaceous gout	Tophus	✓		✓					✓
Full thickness rotator cuff tear	Arthropathy			✓					✓
Disorder of joint of shoulder region	Arthropathy			✓					✓
Injury of ulnar nerve	Injury of brachial plexus			✓					✓
Basal cell carcinoma of ear	Basal cell carcinoma of face			✓					✓
Bronchiolitis	Bronchitis			✓					✓
Migraine variants	Disorder of brain			✓					✓
Gingivitis	Inflammatory disorder of jaw			✓					✓
Septic shock	Soft tissue infection			✓	✓		✓		
Cellulitis of external ear	Otitis externa			✓					✓
Inguinal pain (finding)	Pain in pelvis (finding)			✓					✓
Disorder of tendon of biceps	Disorder of tendon of shoulder region			✓					✓
Pain of breast (finding)	Chest pain (finding)			✓					✓
Injury of ulnar nerve	Ulnar neuropathy	✓		✓	✓				
Injury of back	Traumatic injury	✓		✓				✓	
Achalasia of esophagus	Disorder of stomach			✓					✓
Pneumonia due to respiratory syncytial virus	Interstitial lung disease	✓		✓					
Sensory hearing loss	Labyrinthine disorder			✓				✓	✓
Degeneration of intervertebral disc	Osteoarthritis			✓					✓
Disorder of sacrum	Disorder of bone	✓		✓					✓
Peptic ulcer without hemorrhage, without perforation AND without obstruction	Gastric ulcer			✓				✓	✓
Diabetic autonomic neuropathy	Peripheral nerve disease			✓					✓
Cyst and pseudocyst of pancreas	Cyst of pancreas	✓	✓						
Calculus of kidney and ureter	Ureteric stone	✓	✓	✓					✓

Table 2. Summary and classification of 39 errors identified in 200 relations from the SNOMED CT CORE Problem List Subset by 5 domain experts. For presentation and readability, fully specified names are shortened and all lack semantic tags (e.g., "(disorder)"). Complete, unabridged list available at request.

Discussion

We identified and classified 39 errors (19.5% of 200 relations) in SNOMED CT into a taxonomy of error patterns. At a high level, the errors were due to ambiguous concept terms, anatomic confusion (the most common), and mix-ups of cause and effect. Such errors include "Septic Shock *is-a* Soft Tissue infection" and "Frontal Headache *is-a* Pain in face". The former error is due to cause (infection) and effect (sepsis) while the latter is due to anatomic confusion (face vs. head/skull). These results are surprising; in summary, they suggest that:

1) the extent of errors in SNOMED CT, particularly in complex relations, is quite large
2) errors are very subtle -- in the first round of verification, the majority of experts missed **20%** (8/39) of the errors
3) it is therefore essential to develop quality assurance with many eyes checking for errors

These results support empirically the results of previous exploratory studies by Rector and colleagues who identified similar errors in SNOMED CT.[10] In contrast to the work by Rector et al., this study used a panel of experts to verify a pre-defined set of relations in multiple stages, allowing us to measure expert disagreement. Furthermore, it confirmed the subtlety and difficultly of ontology verification through an empirical, systematic experiment.

A checklist to focus ontology engineering and verification

This work provides three elements to help future verification and method development. First, it presents a taxonomy of errors. The taxonomy should serve as a checklist during ontology verification (i.e., during verification, an expert should consult the list to ensure that none of these types of errors occur). Recent literature describes the importance of checklists when working with complex domains and complex tasks.[27] Please note that our taxonomy enumerates some of the kinds of errors that can occur, but not the *causes*, of which there are many. The taxonomy does not serve as a solution manual to ontology bugs, but instead helps us find them. Second, this paper presented an expert-curated and validated gold standard set of relations from the CORE Problem List Subset against which researchers can now compare methods. While the standard we developed cannot capture or represent the entirety of relations in biomedical ontologies, it serves as a "first stab" at a gold standard, of which, currently, there are none of any useful size. Finally, we developed a generalized logic-based heuristic to saturate errors in an ontology of interest (i.e., the heuristic selects relations that are likely to be in error).

The difficulty of identifying ontology errors

Our two-step Delphi approach to verification, wherein 5 domain experts initially did not reach agreement, highlights the difficulty of identifying ontology errors. It appears that having more eyes on the problem is essential. However, we believe that the next challenge in identifying errors is finding those that are salient (i.e., that impact a system that uses an ontology). To do so, it is necessary to understand the ontology in the context of the system that uses it. In this experiment, part of the difficulty was reaching consensus not only about errors but also about context. In certain contexts, errors may have no significant impact or instead have a positive impact. For example, the incorrect relation "Local Infection of wound *is-a* Wound" may in fact be helpful in retrieval tasks but not with clinical decision support. The taxonomy we developed can guide researchers as they begin to focus on finding errors in the context of ontology applications in a scalable fashion. For example, one might find that cause and effect errors are particularly important to catch in the context of clinical decision support. With an understanding of context and common error patterns, researchers and developers can develop best practices to avoid and repair salient errors. Finally, these results highlight that ontologies are generally tied to particular contexts of use. We suggest that ontology developers become more explicit in how they represent and communicate the context and appropriateness of an ontology for a given application.

Conclusion

Reinforcing results in the literature, experts identified 39 errors in a set of 200 relations from SNOMED CT. We found this task was rather difficult for domain experts, underlining the need for multiple eyes when error-checking and the need for additional method development. We developed a taxonomy of common error patterns. This taxonomy can focus future development of scalable, automated ontology verification methods, giving researchers a pointer toward salient, common patterns of errors. More importantly, ontology engineers can view the taxonomy as checklist to consult when performing ontology quality assurance. Such efforts are imperative, given the prevalence of biomedical ontologies and our reliance upon them.

Acknowledgments
The authors thank Alan L Rector, Timothy E. Sweeney, Evan P. Minty, and Michael Januszyk for serving as domain experts. This work has been supported in part by Grant GM086587 from the National Institute of General Medical Sciences and by The National Center for Biomedical Ontology, supported by grant HG004028 from the National Human Genome Research Institute, the National Lung, Heart, and Blood Instititute, and the National Institutes of Health Common Fund. JMM is supported by National Library of Medicine Informatics Training Grant LM007033.

References

1. Bodenreider O, Stevens R. Bio-ontologies: current trends and future directions. Brief. Bioinform. 2006;7(3):256–74.

2. Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas CI, Tudorache T, et al. BioPortal: enhanced functionality via new web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Res. 2011;39(Web Server issue):W541–5.

3. Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. N. Engl. J. Med. 2010;363(6):501–4.

4. LePendu P, Iyer S V, Bauer-Mehren A, Harpaz R, Mortensen JM, Podchiyska T, et al. Pharmacovigilance using clinical notes. Clin. Pharmacol. Ther. 2013;93(6):547–55.

5. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet . 2003;34(2):166–76.

6. Hoehndorf R, Dumontier M, Gennari JH, Wimalaratne S, de Bono B, Cook DL, et al. Integrating systems biology models and biomedical ontologies. BMC Syst. Biol. BioMed Central Ltd; 2011;5(1):124.

7. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat. Genet. 2000 May;25(1):25–9

8. Golbeck J, Fragoso G, Hartel F, Hendler J, Oberthaler J, Parsia B. The National Cancer Institute's thesaurus and ontology. J. web Semant. 2003;1(1):75–80.

9. Stearns MQ, Price C, Spackman KA, Wang AY. SNOMED clinical terms: overview of the development process and project status. Proc. AMIA Symp. 2001. p. 662.

10. Rector AL, Brandt S, Schneider T. Getting the foot out of the pelvis: modeling problems affecting use of SNOMED CT hierarchies in practical applications. J. Am. Med. Informatics Assoc. BMJ Publishing Group Ltd; 2011;18(4):432–40.

11. Rector A, Iannone L. Lexically suggest, logically define: Quality assurance of the use of qualifiers and expected results of post-coordination in SNOMED CT. J. Biomed. Inform. 2012;45:199–209.

12. Ceusters W, Smith B, Goldberg L. A terminological and ontological analysis of the NCI Thesaurus. Methods Inf. Med. 2005;44(4):498.

13. Ceusters W. Applying evolutionary terminology auditing to SNOMED CT. AMIA Annu. Symp. Proc. 2010;2010:96–100.

14. Zhu X, Fan JW, Baorto DM, Weng C, Cimino JJ. A review of auditing methods applied to the content of controlled biomedical terminologies. J. Biomed. Inform. 2009;42:413–25.

15. Ochs C, Perl Y, Geller J, Halper M, Gu H, Chen Y, et al. Scalability of abstraction-network-based quality assurance to Large SNOMED Hierarchies. AMIA Annu Symp Proc. 2013. p. 1–10.

16. Geller J, Ochs C, Perl Y, Xu J. New abstraction networks and a new visualization tool in support of auditing the SNOMED CT content. AMIA Annu. Symp. Proc. 2012. p. 237–46.

17. Vrandečić D. Ontology evaluation. Handb. Ontol. Springer Berlin Heidelberg; 2009. p. 293–313.

18. The CORE Problem List Subset of SNOMED CT® [Internet]. Available from: http://www.nlm.nih.gov/research/umls/Snomed/core_subset.html

19. Doran P, Tamma V, Iannone L. Ontology module extraction for ontology reuse: an ontology engineering perspective. Proc. 16th ACM Conf. Inf. Knowl. Manag. 2007. p. 61–70.

20. Lawley MJ, Bousquet C. Fast classification in Protégé: Snorocket as an OWL 2 EL reasoner. Proc. 6th Australas. Ontol. Work. (IAOA'10). Conf. Res. Pract. Inf. Technol. 2010. p. 45–9.

21. Noy NF, Mortensen JM, Alexander PR, Musen MA. Mechanical Turk as an ontology engineer? Using microtasks as a component of an ontology engineering workflow. Web Sci. 2013.

22. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004;32(suppl 1):D267–D270.

23. Mortensen JM, Noy NF, Musen MA, Alexander PR. Crowdsourcing ontology verification. Int. Conf. Biomed. Ontol. 2013.

24. Mortensen JM, Musen MA, Noy NF. Crowdsourcing the verification of relationships in biomedical ontologies. AMIA Annu. Symp. 2013.

25. Linstone HA, Turoff M. The Delphi method : techniques and applications. Reading: Addison-Wesley; 1975.

26. Randolph JJ, Thanks A, Bednarik R, Myller N. Free-marginal multirater kappa (multirater κfree): an alternative to Fleiss' fixed-marginal multirater kappa. Joensuu Learn. Instr. Symp. 2005.

27. Gawande A. The checklist manifesto: how to get things right. Metropolitan Books New York; 2010.

Development and validation of an electronic phenotyping algorithm for chronic kidney disease

Girish N Nadkarni, MD, MPH, CPH[1], Omri Gottesman, MD[1], James G Linneman, MS[3], Herbert Chase, MD[3], Richard L Berg, MS[3], Samira Farouk, MD[1], Rajiv Nadukuru, MSc[1], Vaneet Lotay, MSc[1], Steve Ellis, MS[1], George Hripcsak, MD[2], Peggy Peissig, PhD[3], Chunhua Weng, PhD[2] and Erwin P Bottinger, MD[1]

1. Icahn School Of Medicine at Mount Sinai, New York, NY; 2. Columbia University Medical Center, New York, NY; 3. Marshfield Clinic Research Foundation, Marshfield, WI

Abstract

Twenty-six million Americans are estimated to have chronic kidney disease (CKD) with increased risk for cardiovascular disease and end stage renal disease. CKD is frequently undiagnosed and patients are unaware, hampering intervention. A tool for accurate and timely identification of CKD from electronic medical records (EMR) could improve healthcare quality and identify patients for research. As members of eMERGE (electronic medical records and genomics) Network, we developed an automated phenotyping algorithm that can be deployed to identify rapidly diabetic and/or hypertensive CKD cases and controls in health systems with EMRs It uses diagnostic codes, laboratory results, medication and blood pressure records, and textual information culled from notes. Validation statistics demonstrated positive predictive values of 96% and negative predictive values of 93.3. Similar results were obtained on implementation by two independent eMERGE member institutions. The algorithm dramatically outperformed identification by ICD-9-CM codes with 63% positive and 54% negative predictive values, respectively.

Introduction: Chronic kidney disease (CKD) affects an estimated 10% to 15% of individuals in the United States, Europe and Asia. (1) CKD is a largely asymptomatic ('silent') yet serious condition associated with premature mortality, decreased quality of life, and increased health care expenditure. Approximately two thirds of CKD are attributable to diabetes (40% of CKD cases) and hypertension (28% of cases). (2) CKD is defined in most cases clinically by loss of kidney function as estimated by glomerular filtration rate (eGFR) below a threshold of 60 ml/min/1.73m^2 (normal eGFR range 90 to 120 ml/min/1.73m2) and/or persistent increased urinary albumin excretion lasting more than 90 days. (2)Thus, identification of the vast majority of cases of diabetes- and/or hypertension-attributable CKD rests on appropriate and timely ordering and interpretation of eGFR and/or urinary albumin excretion laboratory results.

Untreated CKD can result in end-stage renal disease (ESRD) and necessitate dialysis or kidney transplantation in 2% of cases. (3) CKD is also a major independent risk factor for cardiovascular disease, all-cause mortality including cardiovascular mortality. (2) Current practice guidelines recommend tight control of blood pressure and/or hyperglycemia in particular in the presence of albuminuria to reduce ESRD and CVD risks in CKD patients. (4) However, CKD is alarmingly under diagnosed in affected primary care patients, even those with diabetes and/or hypertension. For example, among 122,502 adults enrolled in Kidney Early Evaluation KEEP, only 20% of participants with CKD stage 3 and 50% with stage 4-5 were aware of their disease. (5) Alarmingly, 43% or patients with newly diagnosed ESRD had not received specialist nephrology care and of those who did, only 25% had done so for more than one year. (5) As a result of the systematic and widespread failure to establish the diagnosis of CKD with inexpensive routine laboratory tests in primary care, affected yet unaware patient populations may not benefit from preventive measures to reduce major outcomes. The Electronic Medical Records and Genomics (eMERGE) Network is a National Human Genome Research Institute (NHGRI)-funded consortium tasked with developing methods and best-practices for the utilization of the Electronic Medical Record (EMR) as a tool for genomic research. (6) The Network's phenotyping workgroup established best practices to develop and use phenotyping algorithms processing EMR data from disparate sources such as diagnosis and procedure codes, laboratory data, medication use, and imaging studies in order to identify cases and controls with a high degree of accuracy and confidence. The Network's PheKB is a repository for phenotype algorithms. Phenotype algorithms on PheKB are validated at the creating site as well as at least 2 other Network institutions.

Here we describe the development of an automated algorithm as part of the eMERGE phenotyping framework that combines data from various EMR sources to identify diabetic and/or hypertensive patients with CKD. To the best of our knowledge, there have been no previous attempts to combine disparate sources of EMR data to identify CKD cases/controls. Positive and negative predictive values of the algorithm were validated extensively at the primary

development site, Mount Sinai Medical Center, and significantly outperformed common recorded International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) codes associated with CKD. Finally, the CKD algorithm was successfully implemented and validated at two additional eMERGE member institutions. Thus, our novel phenotyping tool can be readily deployed by modern health systems to identify undiagnosed cases together with established cases of CKD associated with diabetes and/or hypertension for population management, quality improvement or research initiatives.

Materials and Methods

Participating Site Overview: Three institutions participated in the validation of this algorithm. Each institution uses an EMR and the key features of this are presented in Table 1. All participants represent the population receiving routine clinical care at the study institutions.

Table 1. Overview of participating institutions' EMR and recruitment models

Institution	EMR Overview	Recruitment Model
Icahn School Of Medicine at Mount Sinai (New York, NY, USA)	Comprehensive vendor-based inpatient and outpatient EMR (Epic Systems, Verona, Wisconsin, USA) since 2003 with 10+ years ICD-9 data	Clinical population based
Marshfield Clinic Research Foundation (Marshfield, Wisconsin, USA)	Comprehensive internally developed EMR since 1985 75% participants have 20+ years medical history	Clinical population based
Columbia University Medical Center (New York, NY, USA)	Comprehensive vendor-based and self-developed inpatient and outpatient EMRs (Allscripts and iNYP) since 1990s with 22+ years of structured and unstructured data	Clinical population based

Algorithm development: We used the 2012 Kidney Disease: Improving Global Outcomes (KDIGO) criteria for defining CKD stage 3 or higher (eGFR<60 ml/min/1.73m^2 for duration ≥3 months, based on documentation or inference. The approach utilized common EMR data including diagnostic codes, medications and laboratory results. All sites utilized the International Classification of Diseases, 9th revision, clinical modification (ICD9-CM) diagnostic codes. Table 2 shows the ICD9 diagnosis and procedure codes applied in the algorithm along with the stages in the algorithm where they were used. ICD9 codes for diabetic/hypertensive kidney disease (A1, B2); kidney transplant (B1); dialysis (B3) were used for inclusion while codes for acute kidney failure (B5) and other causes of kidney disease including HIV, immunologic and developmental causes (B9) were used for exclusion criteria. We also implemented text searches to identify terms in physician observation reports for exclusion criteria. These text searches were picked by GNN and EPB using professional expertise as nephrologists and are listed at the end of Table 2. We chose to utilize a rule-based methodology because this approach has been deployed at eMERGE member institutions collaborating on phenotype development and has proven to be replicable across institutions and to produce strong predictive values.

Table 2. ICD-9 Diagnosis, procedure codes and text searches applied in algorithm

A1: 585.xx	Chronic kidney disease
B1: 55.6x	Transplant of kidney (procedure)
B1: V42.0	Organ or tissue replaced by kidney transplant
B2: 250.4x	Diabetes with renal manifestations
B2: 403.xx	Hypertensive chronic kidney disease
B2: 404.xx	Hypertensive heart and chronic kidney disease
B3: V45.1	Renal dialysis status
B3: V56.xx	Encounter for dialysis and dialysis catheter care
B3: 996.73	Complications of renal dialysis
B3: 38.95	Venous catheter for renal dialysis
B5: 584.xx	Acute kidney failure
B9:042.xx-044.xx	Human immunodeficiency virus (HIV) infection
B9: 282.6	Sickle cell disease
B9: 581.xx	Nephrotic syndrome
B9: 582.xx	Chronic glomerulonephritis
B9: 583.xx	Nephritic and nephropathy
B9: 446.xx	Polyarteritis nodosa and allied conditions
B9: 447.6	Vasculitis
B9: 753.xx	Renal agenesis and dysgenesis
Text search terms used in exclusion criteria	HIVAN/HIV associated nephropathy; Congenital [within 2 words of) kidney(s)]; APKD [adult polycystic kidney disease]; Sickle Cell Disease; IgA Nephropathy; Nephrotic Syndrome; Nephritic Syndrome; Glomerulonephritis; Glomerulosclerosis; Lupus Nephritis; Wegener's granulomatosis; Goodpasture's syndrome

With respect to laboratory results, since various formulae are used to estimate the GFR leading to non-uniformity of eGFR results reported by clinical laboratories to the EMR. We overcome the apparent non-uniformity of eGFR lab results calculation and reporting in EHR by recalculating eGFR de novo with the CKD-EPI formula. (7) The CKD-EPI creatinine equation is based on the same four variables as the MDRD Study equation, but uses a 2-slope "spline" to model the relationship between estimated GFR and serum creatinine, and a different relationship for age, sex and race. The equation was reported to perform better and with less bias than the Modification of Diet in Renal Disease (MDRD) Study equation. We did not allow patients that had serum creatinine tests on consecutive days or within the same day as we assumed these were inpatient encounters. This was done to exclude acute kidney injury (AKI) events from the laboratory tests used for the algorithm. For patients that had duplicate serum creatinine tests meaning they had two tests at the same date/time but different values we kept the test with the maximum value in the algorithm. The patients fulfilling the criteria were then filtered through eMERGE Network's type 2 diabetes (8) and a hypertension algorithm that was developed at Icahn School of Medicine at Mount Sinai. These validated algorithms are shown in Figure 1 and 2. The patients then were classified into diabetic CKD (DCKD), hypertensive CKD (HCKD) or diabetic/hypertensive CKD (DHCKD) cases. The complete algorithm for cases is shown in Figure 3.

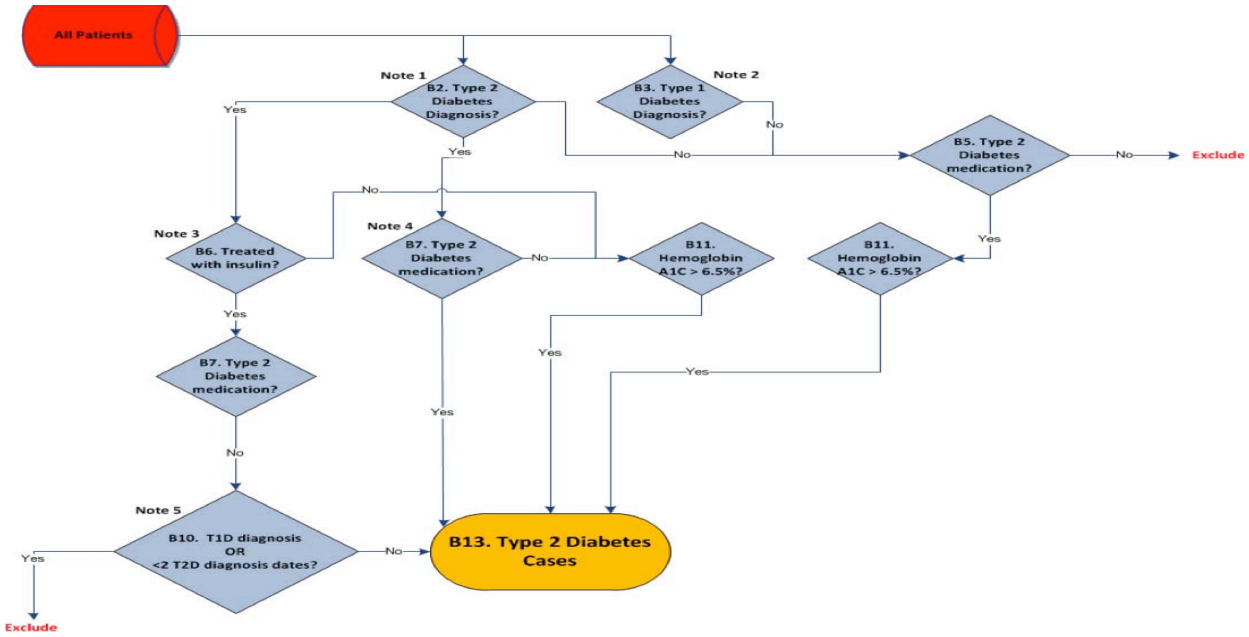


Figure 1. Type 2 Diabetes case algorithm from the eMERGE Network

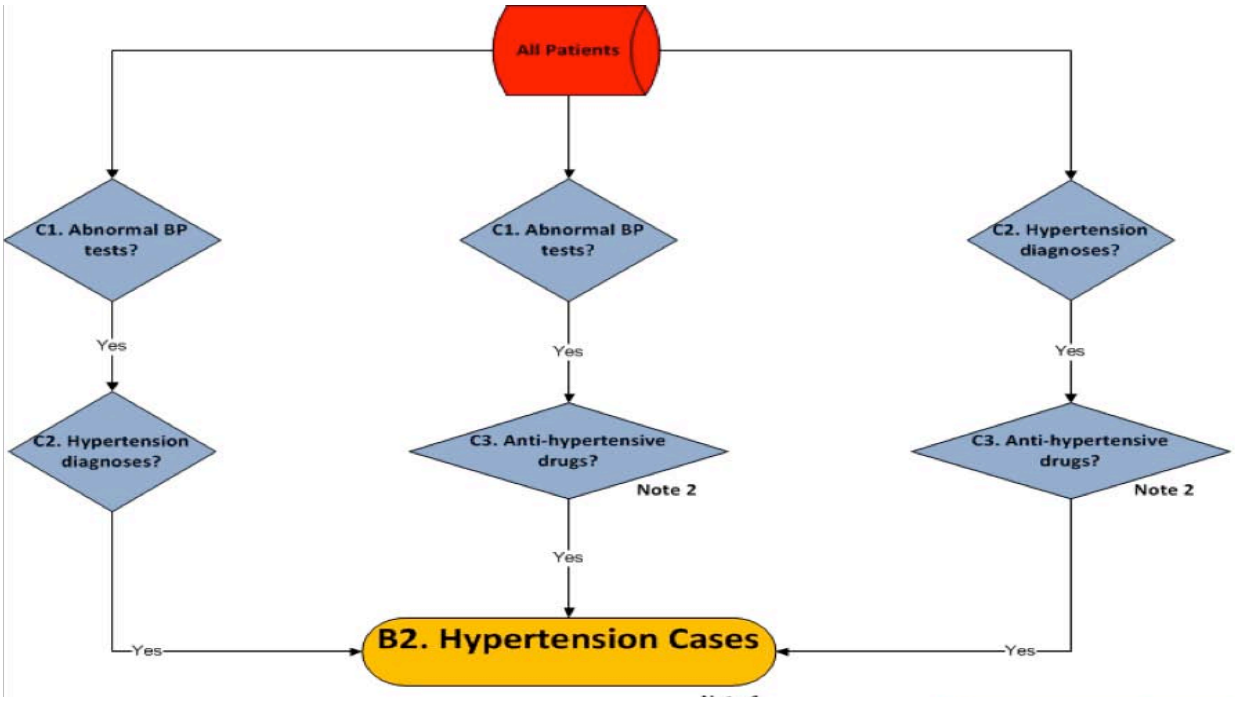


Figure 2. Hypertension case algorithm developed at Icahn School of Medicine at Mount Sinai

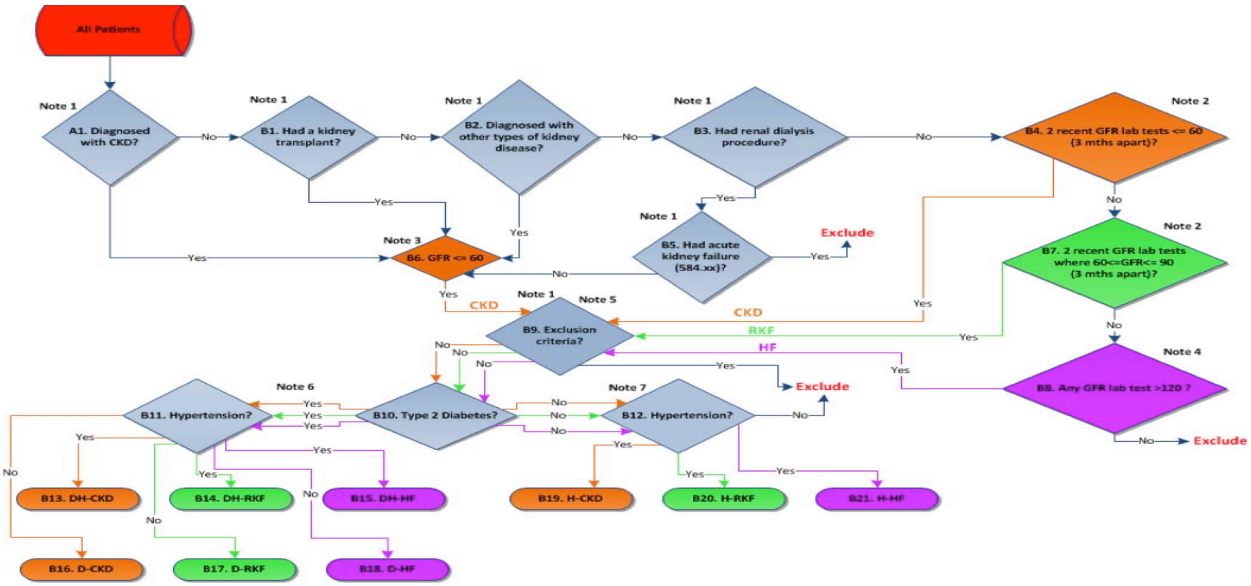


Figure 3. Phenotyping algorithm for chronic kidney disease cases

A similar approach was also adopted for the CKD controls algorithm with identical diagnostic codes, laboratory tests and text searches for exclusion criteria. Participants with 2 eGFR values from 60-89 ml/min/1.73m^2 (reduced kidney function) and at least one eGFR>120 ml/min/1.73m^2 (hyperfiltration) were also identified during this process. Since both reduced kidney function and hyperfiltration indicate an unrecognized early stage of kidney disease, including diabetic and hypertensive CKD they were excluded from all control sets. Once controls were identified, they were then filtered through eMERGE networks type 2 diabetes and hypertension algorithms and then classified as diabetic, hypertensive or diabetic/hypertensive controls. The algorithm for CKD controls is shown in Figure 4.

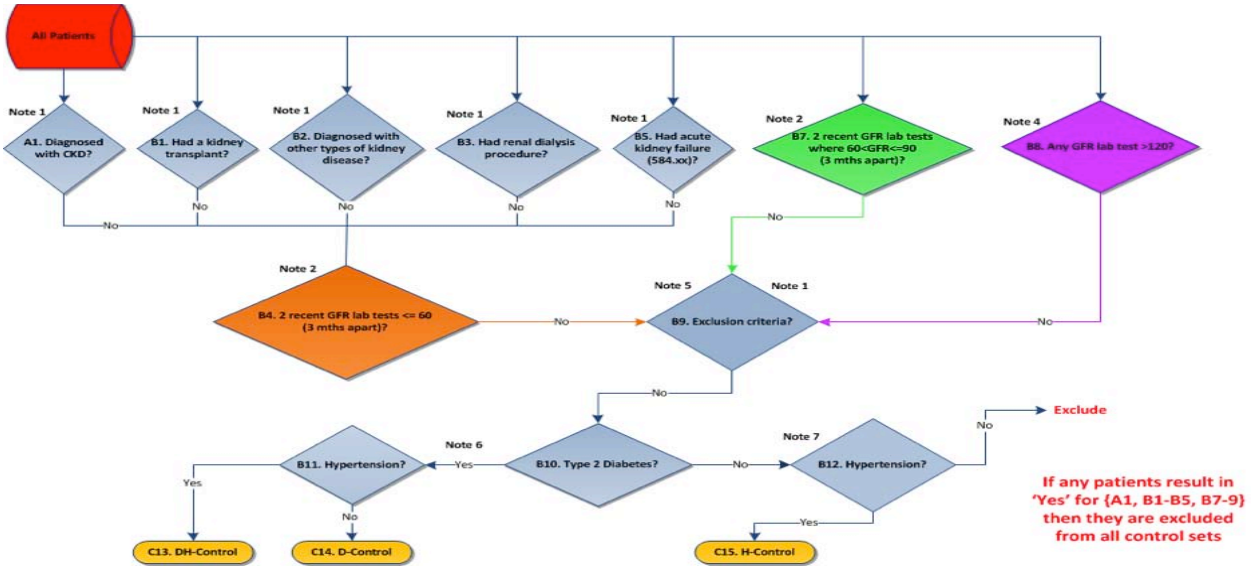


Figure 4. Phenotyping algorithm for chronic kidney disease controls

Implementation: Data elements needed for identification of cases or controls were extracted from the EMR and stored in a shadow relational database. Deidentificaton, pruning (elimination of redundant data points), transformations (for example, separating units from numeric values when they come from the EMR as a single string) and data cleansing were performed as needed. SQL queries were developed and unit-tested for each object in the cases and controls algorithms. The queries were then combined to implement the logic of the algorithms. In the Mount Sinai implementation, the subjects and observations qualifying at each step of the algorithm were saved,

facilitating subsequent validation reconciliation and algorithm refinements. A similar process of extracting EMR data to a shadow database and developing standalone queries was used at the validating sites. The decision logic for this algorithm and the use of specific terms extracted from textual clinical documentation (Table 2) exceed the clinical decision support capabilities embedded in commercial EMR's. However Mount Sinai has developed an external CDS engine that is capable of executing complex coded decision support logic and is integrated with the Epic EMR, thereby enabling the automated CKD algorithm to be integrated into clinical workflows to provide alerts to clinicians of patients at risk for CKD. (9)The validating sites have internally developed EMR's into which they can embed complex CDS code to provide alerts within existing clinical workflows. Sites using commercial EMR's would most likely not be able to implement the CKD algorithm with the available technology.

D**ata collection:** The algorithm was deployed to randomly select 600 each of CKD cases and controls from the EMR. Two independent physician reviewers (GNN and SF) at Mount Sinai Hospital manually reviewed each medical record. The gold standard for a case or a control was considered to be manual review by the physician reviewers. Any differences in agreement were arbitrated after discussion between the two reviewers. Identification of CKD in the EMR was ascertained using the CKD hierarchy of ICD9 codes (Table 3). A control was considered as correctly identified by ICD-9 codes if there was a diagnostic code identifying hypertension and/or diabetes as shown without an accompanying code for CKD (Table 3). While reviewing the charts, urine protein measurements as microalbumin/creatinine ratio and/or urine protein/creatinine ratios were also abstracted. When multiple measurements were available, the most recent measurement was recorded. In addition, documented referral to a nephrologist was abstracted. The algorithm was then deployed at Marshfield Clinic and Columbia University for secondary validation. Chart review by physician reviewers was used to validate 50 cases and 50 controls at each secondary site

Table 3. ICD-9 codes used for identifying cases and controls

Disease	ICD-9 code
End stage renal disease	585.1 to 585.9
Hypertensive chronic kidney disease, unspecified, with chronic kidney disease stage I through stage IV, or unspecified	403.90
Hypertensive nephropathy	403.10
Hypertensive renal disease	403
Hypertensive heart and renal disease	404
Diabetic nephropathy	583.81
Diabetic nephrosis	581.81
Diabetes with renal manifestations, type II or unspecified type, not stated as uncontrolled	250.40
Diabetes with renal manifestations, type II or unspecified type, uncontrolled	250.42
Diabetes with other specified manifestations, type II or unspecified type, not stated as uncontrolled	250.80
Diabetes with other specified manifestations, type II or unspecified type, uncontrolled	250.82
Intercapillary glomerulosclerosis	581.81
Kimmelstiel-Wilson syndrome	581.81
Hypertension	401-405 excluding 403 and 404
Diabetes Mellitus Type 2	250.00 to 250.93 excluding 250.40, 250.42,250.80 and 250.82)

Data analysis: After manually reviewing the cases, the inter-rater agreement/kappa statistic was calculated. Summary statistics (positive and negative predictive values) for identification of cases and controls with the algorithm and ICD-9 codes with manual as the gold standard were estimated. Medians and interquartile ranges and proportion of missing values for the urine protein measurements and the percentage of patients that had been referred to a nephrologist were also calculated for the primary site. Positive and negative predictive values were then calculated individually for both secondary sites. All analyses were performed using STATA SE version 12, College Station, TX.

Results: A total of 1200 medical records were reviewed. Out of these, 14(1.16%) were excluded due to confidential status or missing data, leaving 1186 patients included in the final analysis. The inter-rater agreement/kappa statistic between the two independent reviewers (SF and GNN) was 92%. After arbitration of disagreements there were a total of 609 cases (202 for Diabetic CKD [DCKD], 207 for Hypertensive CKD [HCKD] and 200 for Diabetic and Hypertensive [DHCKD]) and 577 controls (190 for DCKD, 190 for HCKD and 197 for DHCKD) by manual review. The summary of the comparison between the algorithm and ICD9 codes are presented in Table 4 for the primary site. The comprehensive algorithm correctly identified 569/609(93.43%) of cases and 553/577(95.84%) of controls. In contrast, conventional screening with ICD9 codes only identified 244/609(40.06%) of cases and 433/577(75.04%) of controls. The positive predictive value (PPV) for the algorithm was 95.95% and the negative predictive value (NPV) was 93.25% compared to a PPV of 62.89% and a NPV of 54.26% compared to identification using ICD9-CM diagnostic codes. The algorithm performed similarly at secondary sites with a PPV of 92% and a NPV of 100%[Table 5].

Table 4. Comparison of phenotyping algorithm and ICD-9 codes at primary site (Mount Sinai Hospital)

	Manual chart review			Manual chart review			
Phenotyping Algorithm	Case	Control	Total	ICD-9 Codes	Case	Control	Total
Case	569	24	593	Case	244	144	593
Control	40	553	593	Control	365	433	593
Total	609	577	1186	Total	609	577	1186
Positive Predictive Value (95% Confidence Interval)	95.95 (93.95 -97.33)			Positive Predictive Value (95% Confidence Interval)	62.89 (57.84-67.67)		
Negative Predictive Value (95% Confidence Interval)	93.25 (90.85-95.08)			Negative Predictive Value (95% Confidence Interval)	54.26 (50.73-57.75)		

Table 5. Performance of phenotyping algorithm at secondary sites (Marshfield Clinic and Columbia University Medical Center

	Marshfield Clinic			Columbia University Medical Center		
	Manual chart review			Manual chart review		
	Case	Control	Total	Case	Control	Total
Phenotyping Algorithm						
Case	46	4	50	46	4	50
Control	0	50	50	0	50	50
Total	46	54	100	46	54	100
Negative Predictive Value (95% Confidence Interval)	92 (79.89-97.41)			92 (79.89-97.41)		
Negative Predictive Value (95% Confidence Interval)	100 (91.11-100)			100 (91.11-100)		

As part of secondary analysis the urine protein/creatinine or the urine microalbumin/creatinine values for DCKD, HCKD and DHCKD cases (Table 6). For DCKD, the median microalbumin/creatinine ratio was 39 microgram/mg of creatinine. Similarly for HCKD the median microalbumin/creatinine ratio was 5.5 microgram/mg of creatinine and the protein/creatinine ratio and for DHCKD, the median microalbumin/creatinine and protein/creatinine ratios were 35 microgram/mg of creatinine and 15 mg/mg of creatinine, with 37% and 61% of patients respectively lacking measurements at any point of time. However, there was no record of a urine albumin or protein excretion for

a significant proportion of patients (ranging from 30-98%) depending on the subcategory of CKD. (Table 6) We also determined the proportion of participants that were referred to a nephrologist during any point of their clinical course in the EMR. Out of a total of 599 cases, only 112(18.7%) were referred to a nephrologist at any point during their course.

Table 6. Microalbuminuria and proteinuria measurements in CKD cases and controls

	Diabetic CKD	
	Median (IQR)	N (%) of missing values
Microalbumin/Creatinine	39(10-215)	59/200(30)
	Hypertensive CKD	
	Median (IQR)	N (%) of missing values
Urine microalbumin/creatinine	5.5(3-28)	196/200(98)
Urine protein/creatinine in mg/gm	30(30-300)	173/200(87)
	Diabetic and hypertensive CKD	
	Median (IQR)	N (%) of missing values
Urine microalbumin/creatinine	35(8-127)	74(37)

Discussion: In the near future, EMRs will become one of the most important sources of data for both clinical and genomic association studies. Since data are present longitudinally, it may facilitate studying natural history of a disease process as well as the response to treatment in a "real world" scenario. However, identification of particular phenotypes, especially chronic, complex diseases, is challenging because of the complexity of data itself and the way in which it is recorded in EMR. However, with government interest driving the widespread use and adoption of EMR's, this provides a vast and as-yet relatively untapped resource. (10) If robust phenotypes were constructed using meaningful information from various EMR sources, it would provide significant value for identifying patient cohorts that satisfy complex criteria. There has been significant debate about the optimal way to identify phenotypes in the EMR. (11) Automated approaches using electronic phenotyping and statistical analyses are popular as compared to simpler rule based systems. The utility of such phenotyping algorithms is manifold, including discovering novel genetic associations of complex diseases, tracking their natural history, isolating patients for clinical trials, and ensuring quality control in large institutions by ensuring that standard of care guidelines are met in these patients.

Kidney disease is a complex, common problem challenging modern healthcare. It is a major independent risk factor for all-cause mortality including cardiovascular mortality and adjusted rates of all-cause mortality are seven times greater for dialysis patients than for individuals in the general population. (3,12,13)As CKD is a significant health problem, accurate identification of diabetic and/or hypertensive CKD cases and controls for both research and clinical purposes is imperative. Accurate identification of individuals satisfying specific criteria from a large institutional population allows us to enroll for randomized trials, predict/track outcomes/progression, and perform retrospective cohort studies. (14,15) Studying the progression of complex diseases such as CKD is difficult as the recruitment of cohorts is a laborious process that creates a bottleneck in both clinical and translational research. In order to streamline this process, there has been an impetus to create EMR linked biobanks to enroll individuals in routine clinical care settings. The push from healthcare regulatory agencies for electronic medical records (EMRs) that provide a large amount of information available for research purposes has also been integral in improving the formation of research cohorts. (16) With appropriate patient consent and de-identifying data, the EMRs of patients are available and allow the studying of evolution and progression of disease. (17,18) In clinical care settings, a wealth of data is available through ICD-9 codes, discrete laboratory results, test reports, patient demographics, and notes written by the treating physicians. All of these data are available in a longitudinal form with multiple patient visits over several years. If matched to biobanks, the EMR can be used to identify traits/phenotypes in a large number of patients for biomarker/genomics research, thereby substantially reducing the effort and time needed to identify markers or variants that influence disease development, progression, or medication response. (18–20)

Although there are novel genetic associations including *UMOD*, *APOL1* and *SHROOM3*, there are other potential genetic associations that explain the differential rates of CKD in different ethnic populations. (21,22) Clinical decision making is challenging due to variability in the rates of progression and lack of widely-accepted guidelines to identify patients most at risk of progression to ESRD. (23,24) For studies to assess progression over the course of the patient's history in the EMR, accurate identification of large numbers of patients is needed. Currently the only way of identifying CKD cases/controls is by manually reviewing laboratory values, which is cumbersome, or through ICD9 codes. To accomplish these goals, researchers need robust phenotyping algorithms to effectively leverage disparate data sources in the EMR. To the best of our knowledge, this algorithm is one of the first automated phenotyping algorithms for diabetic/hypertensive CKD. It significantly outperformed conventional screening with ICD9 codes and could be deployed to different EMRs in various healthcare institutions. Thus, an integrated approach using diagnostic codes, medications, and laboratory tests yielded significant improvement over non-integrated approaches.
Although there are no recommended guidelines for nephrologist referral in CKD stage 3, there are studies suggesting that such referrals may improve prognosis. (25,26) However, we demonstrated that only 18.7% of confirmed CKD cases were referred to a nephrologist at any point during their EMR course. Thus, identifying appropriate patients for referral to a nephrologist is one of the many clinical applications of this algorithm.

Limitations: This algorithm does not include proteinuria or albuminuria measurements that are used to diagnose and stage CKD in addition to eGFR. This decision was based on the observation that in many instances urine albumin or protein excretion results are not recorded in the EMR. This approach was developed to classify co-existing prevalent CKD and T2D and/or HTN because in the vast majority of cases, T2D and/or HTN precede CKD, especially when other etiologies of kidney dysfunction are excluded (as they are in the algorithm). In a sub-analysis, we found that $\geq$ 95% of patients had a prior diagnosis of HTN and/or T2D before CKD, thus validating that the number of patients that might have CKD due to other etiologies are likely minimal. There was also a consideration of overfitting due to the high dimensionality of this problem under consideration. However, since the algorithm was replicated at multiple sites with near-identical results, this is likely to be negligible. Also, considering the low referral rate, it is possible that patients without EMR documented referral may have been referred to nephrologists outside the EMR.

Conclusions: In summary, we describe the development and validation of an automated algorithm for identifying diabetic/hypertensive CKD cases and controls and also demonstrate its superiority over traditional identification using ICD-9 diagnostic codes. We believe that this algorithm could be used to accurately and rapidly identify a specific target cohort within the EMR for both research and clinical purposes.

Acknowledgments/Funding: The eMERGE Network is funded by NHGRI, with additional funding from NIGMS through the following grants: U01HG004438 to Johns Hopkins University; U01HG004424 to The Broad Institute; U01HG004438 to CIDR; U01HG004610 and U01HG006375 to Group Health Cooperative; U01HG004608 to Marshfield Clinic; U01HG006389 to Essentia Institute of Rural Health; U01HG04599 and U01HG006379 to Mayo Clinic; U01HG004609 and U01HG006388 to Northwestern University; U01HG04603 and U01HG006378 to Vanderbilt University; U01HG006385 to the Coordinating Center; U01HG006382 to Geisinger Clinic; U01HG006380 to Icahn School of Medicine at Mount Sinai; U01HG006830 to The Children's Hospital of Philadelphia; and U01HG006828 to Cincinnati Children's Hospital and Boston Children's Hospital.

References:

1. Levey AS, Stevens LA, Coresh J. Conceptual model of CKD: applications and implications. Am J Kidney Dis Off J Natl Kidney Found. 2009 Mar;53(3 Suppl 3):S4–16.
2. Levey AS, Coresh J. Chronic kidney disease. Lancet. 2012 Jan 14;379(9811):165–80.
3. USRDS 2011 Annual Data Report. in Atlas of Chronic Kidney Disease and End-Stage Renal Disease in the United States (ed. National Institutes of Health) (National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD, 2011).
4. Appel LJ, Wright JT, Greene T, Agodoa LY, Astor BC, Bakris GL, et al. Intensive blood-pressure control in hypertensive chronic kidney disease. N Engl J Med. 2010 Sep 2;363(10):918–29.
5. Agrawal V, Jaar BG, Frisby XY, Chen S-C, Qiu Y, Li S, et al. Access to health care among adults evaluated for CKD: findings from the Kidney Early Evaluation Program (KEEP). Am J Kidney Dis Off J Natl Kidney Found. 2012 Mar;59(3 Suppl 2):S5–15.

6. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, et al. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet Med Off J Am Coll Med Genet. 2013 Oct;15(10):761–71.
7. Levey AS, Stevens LA, Schmid CH, Zhang YL, Castro AF, Feldman HI, et al. A new equation to estimate glomerular filtration rate. Ann Intern Med. 2009 May 5;150(9):604–12.
8. Kho AN, Hayes MG, Rasmussen-Torvik L, Pacheco JA, Thompson WK, Armstrong LL, et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. J Am Med Inform Assoc JAMIA. 2012 Apr;19(2):212–8.
9. Gottesman O, Scott SA, Ellis SB, Overby CL, Ludtke A, Hulot J-S, et al. The CLIPMERGE PGx Program: clinical implementation of personalized medicine through electronic health records and genomics-pharmacogenomics. Clin Pharmacol Ther. 2013 Aug;94(2):214–7.
10. Office of the National Coordinator for Health Information Technology. Electronic Health Records and Meaningful Use. 2011 [Internet]. Available from: http://healthit.hhs.gov/portal
11. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc JAMIA. 2014 Apr;21(2):221–30.
12. Chronic Kidney Disease Prognosis Consortium, Matsushita K, van der Velde M, Astor BC, Woodward M, Levey AS, et al. Association of estimated glomerular filtration rate and albuminuria with all-cause and cardiovascular mortality in general population cohorts: a collaborative meta-analysis. Lancet. 2010 Jun 12;375(9731):2073–81.
13. Gansevoort RT, Matsushita K, van der Velde M, Astor BC, Woodward M, Levey AS, et al. Lower estimated GFR and higher albuminuria are associated with adverse kidney outcomes. A collaborative meta-analysis of general and high-risk population cohorts. Kidney Int. 2011 Jul;80(1):93–104.
14. Mathias JS, Gossett D, Baker DW. Use of electronic health record data to evaluate overuse of cervical cancer screening. J Am Med Inform Assoc JAMIA. 2012 Jun;19(e1):e96–101.
15. Strom BL, Schinnar R, Jones J, Bilker WB, Weiner MG, Hennessy S, et al. Detecting pregnancy use of non-hormonal category X medications in electronic medical records. J Am Med Inform Assoc JAMIA. 2011 Dec;18 Suppl 1:i81–6.
16. Blumenthal D, Tavenner M. The “meaningful use” regulation for electronic health records. N Engl J Med. 2010 Aug 5;363(6):501–4.
17. McCarty CA, Nair A, Austin DM, Giampietro PF. Informed consent and subject motivation to participate in a large, population-based genomics study: the Marshfield Clinic Personalized Medicine Research Project. Community Genet. 2007;10(1):2–9.
18. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinforma Oxf Engl. 2010 May 1;26(9):1205–10.
19. Carroll RJ, Eyler AE, Denny JC. Naïve Electronic Health Record phenotype identification for Rheumatoid arthritis. AMIA Annu Symp Proc AMIA Symp AMIA Symp. 2011;2011:189–96.
20. Birman-Deych E, Waterman AD, Yan Y, Nilasena DS, Radford MJ, Gage BF. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. Med Care. 2005 May;43(5):480–5.
21. Köttgen A, Glazer NL, Dehghan A, Hwang S-J, Katz R, Li M, et al. Multiple loci associated with indices of renal function and chronic kidney disease. Nat Genet. 2009 Jun;41(6):712–7.
22. Parsa A, Kao WHL, Xie D, Astor BC, Li M, Hsu C, et al. APOL1 risk variants, race, and progression of chronic kidney disease. N Engl J Med. 2013 Dec 5;369(23):2183–96. Keane WF, Zhang Z, Lyle PA, Cooper ME, de Zeeuw D, Grunfeld J-P, et al. Risk scores for predicting outcomes in patients with type 2 diabetes and nephropathy: the RENAAL study. Clin J Am Soc Nephrol CJASN. 2006 Jul;1(4):761–7.
24. Keith DS, Nichols GA, Gullion CM, Brown JB, Smith DH. Longitudinal follow-up and outcomes among a population with chronic kidney disease in a large managed care organization. Arch Intern Med. 2004 Mar 22;164(6):659–63.
25. Orlando LA, Owen WF, Matchar DB. Relationship between nephrologist care and progression of chronic kidney disease. N C Med J. 2007 Feb;68(1):9–16.
26. Kim DH, Kim M, Kim H, Kim Y-L, Kang S-W, Yang CW, et al. Early referral to a nephrologist improved patient survival: prospective cohort study for end-stage renal disease in Korea. PloS One. 2013;8(1):e55323.

Using TURF to Understand the Functions of Interruptions

Vickie Nguyen, MA, MS[1,2], Nnaemeka Okafor, MD, MS[1,3], Jiajie Zhang, PhD[1,2], Amy Franklin, PhD[1,2]
[1]The National Center for Cognitive Informatics and Decision Making in Healthcare; [2]School of Biomedical Informatics; [3]Medical School, The University of Texas Health Science Center, Houston, TX

Abstract

Interruptions are an often lamented and frequently studied aspect of clinical practice. However, some interruptions, such as updates on patient care decisions and notifications of detrimental patient lab values, are in fact necessary to the work process. In this paper, we explore the interruptions as an emergent feature of communication in teams. Looking beyond the frequency of interruptions, we consider the source and intent of interruptions with the goal of discovering the functions served by such communications. Furthermore, in this study of an emergency department, we classify interruptions into those activities that support required work and those interruptions that create unnecessary breaks in workflow. The intent of our larger body of work is to develop health information technology systems that support team efforts including the functions currently served by interruptions.

Introduction

Interruptions

Interruptions, or a break in task to execute an unplanned task initiated by an internal or external source resulting in the pause or termination of the original task[1], are an often bemoaned and frequently studied aspect of clinical practice[2-7]. While unnecessary interruptions have the potential to lead to increased patient risk[6-9], some interruptions such as updates on patient care decisions by clinicians and notifications of detrimental patient lab values or irregular imaging results, are in fact necessary to the work process[1]. In this paper, we explore interruptions as an emergent feature of communication in teams. We focus on emergency medicine as this complex environment requires team based concurrent management of multiple patients coping with limited resources in a life-critical and interruption-laden environment[10].

U.S. emergency department physicians experience interruptions, on average, 10 times per hour during a shift[6, 8]. These routine breaks in task are more frequent in nursing, where emergency department nurses are interrupted every 4.5 minutes[6, 9]. Such interruptions not only impact the clinical care by potentiating medication errors or tasks on the wrong on the wrong patient[2, 4, 5, 11], and have also been correlated to lower patient satisfaction[12].

Interruptions are not always to the detriment of care. It has also been found that the act of interruptions can help increase shared knowledge amongst team members. With efficient communications between team members, new knowledge or work routines are exchanged and such conversations improve the team's performance on completing tasks[13]. Other positive impacts on task performance, the emotional state, and social attribution of people, depending on their current cognitive load, have also been found[14].

In addition to the above studies, much of the literature has focused on identifying the rate of interruptions, tasks that are interrupted, and the participants engaged in the interruption[1, 8, 9, 15, 16]. Understanding the reasoning for interruptions, or the function that interruptions serve, will allow us to create interventions, here in the form of a health information technology, that support required interruptions while mitigating unnecessary breaks in work. Inherent in our assumption is the need for this type of communication as part of ongoing clinical practice.

TURF

The TURF framework, a method of Work-Centered Design, provides a qualitative and quantitative means of developing useful and usable systems. TURF stands for Task, User, Representation, and Function (TURF) analyses[17, 18]. TURF is intended to aid in the creation of tools that provide a more useful and satisfying experience for system users while promoting patient safety and quality. This is accomplished by understanding the needs of users while supporting the functions to be accomplished as part of overarching goals and the tasks that are required in those efforts. Additionally, the users and their preferred visualizations make up other components of well thought out systems.

The objective of this study is to understand the function of interruptions by describing their place in the workflow. Namely, are the interruptions serving the functions of clinical work or do they reflect issues within the system (*e.g.*, the interface does not fulfill the needs of the clinician in providing patient care) or environment (*e.g.*, clinicians failing to understand roles and responsibilities required for completing overarching patient care goals). Our findings will help to inform our design of a team support tool to help clinical teams manage interruptions in the emergency department.

Domain versus Overhead Functions

The term overhead has traditionally been used in the business organizational context to refer to expenses required in operating a company such as paying rent for the facilities to manage production. Overhead is *not directly connected* to the items produced. In such examples, these extraneous outlays are part of "operating costs." Similarly, the term has been used in the engineering and computational domain to describe design and algorithm features that require additional features from the system in order to reach a goal such as expending additional energy to run a circuit or transmitting signals in addition to computing data. By contrast, in these contexts, domain functions are used to describe the elements *directly required in the work*. That is what is needed to create the product.

Zhang, Butler, and colleagues (2007 & 2011) expand on the use of domain and overhead terms to describe required and extraneous functions used to operate systems[17, 18, 19]. For example, a well designed interface would only display the domain, or required, operations users must click through in order to execute a command. A poorly designed system would have additional overhead, or unnecessary, operations interspersed throughout the system which may generate unnecessary effort. Couched more tangibly, when domain and overhead functions were applied to an electronic health record (EHR) allergy entry workflow, researchers discovered 99 overhead, and thus, extraneous functions. This is in stark comparison to the 28 domain functions[20]. This nearly 4:1 ratio of overhead to domain functions had the potential to dramatically impact the usability of this system.

For this paper, we further expand on the use of domain and overhead functions by using these terms to describe the interruptions within a clinical context. An overhead interruption refers to non-clinically dependent communications that disturb existing work processes. Some examples of overhead interruptions include: tracking down clinicians, repeating tasks already completed by other clinicians, completing tasks unrelated to the realms of your responsibilities, and performing workarounds of poorly implemented systems. Domain interruptions are therefore, interruptions whose communication fits within the needs of work. Examples of domain-type interruptions are providing updates on patient care plans, interpreting electrocardiography (EKG) or lab results, or verifying patient disposition decisions. Overhead interruptions can delay or impede the patient care process as they are not a necessary part of the required work while domain interruptions clearly support the patient care process as they are a part of the required work. Differentiating between these two types of interruptions may help us further understand how to better manage communications. We seek to support domain interruptions while mitigating the cognitive and workload burdens created by overhead-type interruptions.

Table 1 provides the definitions we used to classify interruptions by their function.

Table 1. Definitions used to classify interruptions by function.

Terms	Definitions
Interruption	"A break in the performance of a human activity initiated by a source internal or external to the recipient. This break results in the suspension of an initial task to perform an unplanned task which results in a break or termination of the primary task[1]."
Overhead	Interruptions that are not clinically dependent and create disturbed work processes (*e.g.*, Technology Workarounds). Often caused by the implementation of poorly designed systems or workflow methods[17, 18, 19].
Domain	Interruptions necessary for the completion of clinical tasks and create undisturbed work processes (*e.g.*, Interpreting EKGs)[17, 18, 19].

Methods

Data for this study was drawn from a larger project examining decision making and information foraging in the emergency department. The dataset includes ethnographic non-participant observations of emergency medicine work from multiple clinical perspectives at a Gulf Coast Trauma 1 Center Teaching Hospital. This hospital sees

about 6065 trauma cases per year and employs fifty-two attending physicians, fifty-three residents, and fifty nurses. The emergency department at this hospital uses an EHR system with clinical provider order entry functionalities, trackboard, and a traditional pager and phone communication procedure.

Here, we shadowed emergency department attending physicians, residents, and nurses. From this dataset, a convenience sample of six attendings, five residents, and five nurses were included in this secondary analysis. Each participant was observed on two different days for a time period of four hours each during each participant's start of shift. A total of 128 hours of observation are included in this study. Following our IRB approved procedure, participants gave written informed consent before observations began[a].

Two non-clinical graduate students shadowed the clinicians. Observations were collected using pen and paper, a digital wristwatch with the time in seconds, and an audio recorder. Information was collected using a field note taking procedure established in our previous ethnographic studies[21]. The audio recording device was placed within a pocket on the participant's person and its microphone clipped at the participant's white coat or scrub shirt collar[21].

Using a categorization framework created by Brixey et al. (2007) and Franklin et al. (2011), we identified the types of interruptions and task transition decision making (*i.e.*, interruptions, opportunistic decisions, or protocol based actions) that occurred among the emergency department team during[21, 22]. Interruptions were noted from all types of communications ranging from patient updates to general social behaviors during the observed work hours. Observation notes were supplemented with the audio recording for clarification. The observers had a reliability of κ=.80 indicating strong consistency in data collection and the reliability of initial categorization.

In this paper, we then further classified the interruptions from the above dataset. We identified whether the clinician observed was the recipient or initiator of the interruption[16], and whether or not the interruption was an overhead (*i.e.*, unnecessary communication for the completion of a task or maintaining team shared knowledge and awareness) or domain-type (*i.e.*, necessary for completing task or maintain team needs) interruption[17, 18, 19].

We were able to capture a snapshot of the emergency department team from the attending, resident, and nurse perspective on the directionality and interruption type by a member of the emergency department team.

Results

Attending Perspective

910 interruptions occurred during the 48 hours of observations of the six emergency department attending physicians (M = 18.96 interruptions per hour). First, considering the directionality of the interruptions, that is creator or receiver, we found that attending physicians are more often interrupted (91%) rather than the source of interruptions (9%). 52% of interruptions were initiated by residents with only 13% initiated by nurses.

When we consider the function, domain or overhead, served by these interruptions, we found 23% are overhead-type interruptions unnecessary for completing clinical tasks or maintaining teamwork needs. Common overhead-type interruptions that attending experienced were: being asked to help locate other clinicians, relaying information to other clinicians, and EHR problems (*i.e.*, technical issues). Again, residents were the primary source of these overhead interruptions (35%) with a small portion contributed by consultants (10%) and nurses (14%).

Resident Perspective

For the resident physicians, a total of 697 interruptions occurred during the 40 hours of observations (M = 17.43 interruptions per hour). 26% of interruptions received by the residents were initiated by attendings while nurses initiated 21% of these breaks in task. As initiators of interruptions, residents caused (52%) interruptions for attendings and (20%) interruptions for nurses.

For our functional analysis, we found 30% of the interruptions (n = 208) were classified as overhead-type interruptions. The overhead-type interruptions that residents experienced were: finding information for other clinicians, helping clinicians find patients, and helping other clinicians navigate the EHR. Overhead-type interruptions were initiated by attendings (24%), consultants (14%), other residents (15%), and nurses (12%).

[a] All clinical team members provided consent for each day of observation, including non-shadowed members of the group.

Nurse Perspective

Forty hours of nursing data was included in this secondary analysis. In that time frame, nurses experienced a total of 1053 interruptions (M = 26.33 interruptions per hour). Of the total interruptions, nurses initiated breaks in task on only 18% of the occasions observed. Residents contributed to 17% of the nurse-recipient interruptions.

Considering the function of the interruptions, 40% of the total interruptions observed were classified as overhead-type interruptions. These most often included inquiries by other nurses (45%).

Triangulation of Clinical Perspectives

After categorizing the interruptions by directionality and type of function from each clinical role's perspective, we performed a triangulation of the data comparing the occurrences of overhead and domain-type interruptions for a comprehensive analysis of the emergency department team's interactions (Figure 1). By comparing the difference perspectives of interruptions, we can validate each angle and further understand how each clinical role communicates and interacts with one another to determine how to better manage interruptions in the emergency department environment.

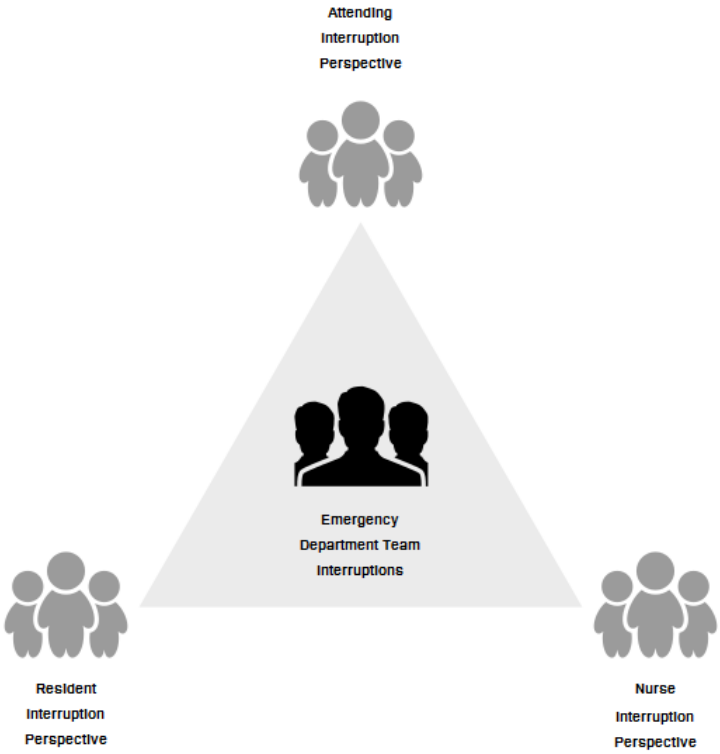


Figure 1. Triangulation of interruption perspectives by clinical role to validate the clinical team perspective.

Comparison of the different clinical role perspectives indicate an increasing number of domain interruptions from attending physician to nurse (Figure 2).

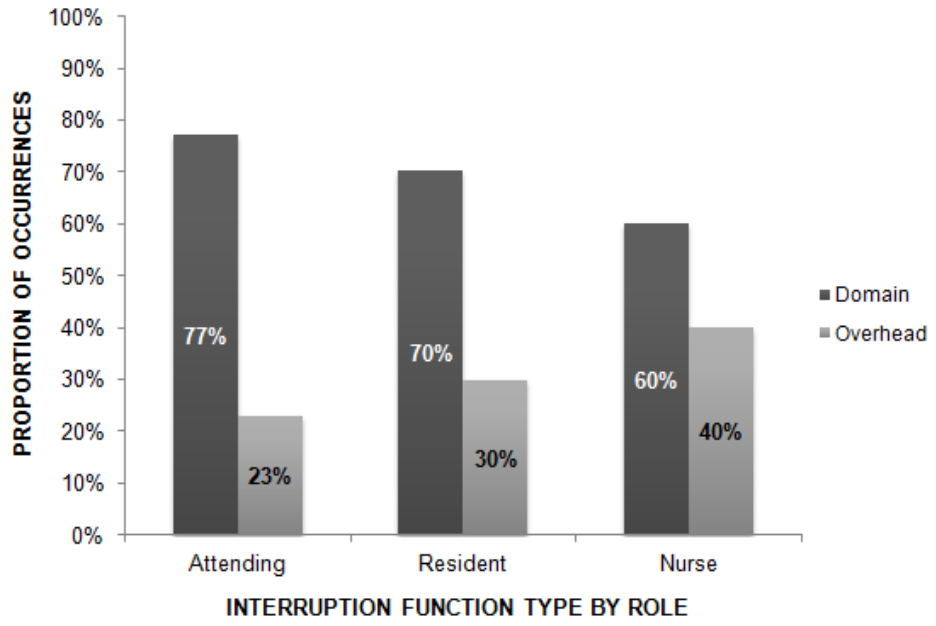


Figure 2. Percentage of domain and overhead types of interruptions experienced by role.

Breaking down the interruption functions by initiators and recipients of interruptions along with role demonstrates that although everyone received more interruptions than they initiated, there are differences in clinical role and interruption role (Figure 3). Nurses initiated more domain interruptions when compared to physicians, and though comparable, residents initiated more overhead breaks in task.

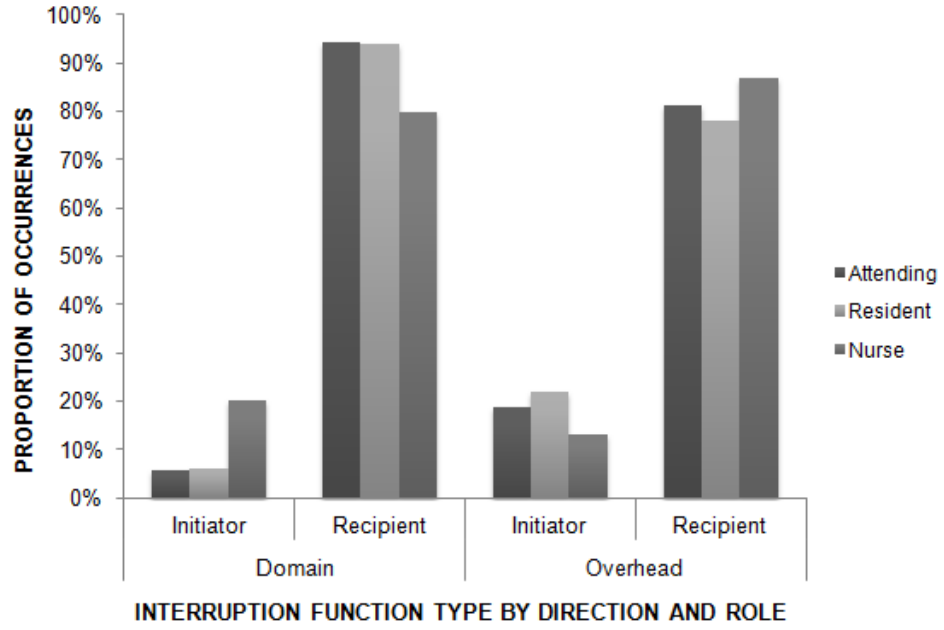


Figure 3. Classification of the identified interruptions by function, domain/overhead, direction of communication, and clinical role.

Discussion

In 128 hours of shadowing, 2660 interruptions were observed. Across clinical roles, this is the equivalent of 21 interruptions per hour. 61.1% of these breaks in task were classified here as received domain interruptions (n = 1626).

These findings demonstrate the ubiquitous nature of interruptions in emergency care and suggest a need for the management of interruptions, especially interruptions unnecessary to clinical tasks. Falling in line with expectation, residents and nurses received the majority of the domain-type interruptions as they are subordinates in the team and these types of interruptions may be natural to their workflow as task directives. For nurses to initiate a majority of the domain-type interruptions suggests that some communication gaps may be bridged by nurses between the attending, residents, and other nurses.

Residents were the main initiators of overhead-type interruptions. These types of interruptions may be used to manage responsibilities outside of clinical work, but related to duties of promoting standard social work behaviors and maintaining a team environment.

In addition to the overhead interruptions, the domain interruptions also require support. The use of interruptions as a communication strategy is an emergent feature of clinical teams. Currently, breaking the process of someone's effort may be necessary in some instances, but may be disruptive in other instances. We believe building from these initial classifications with additional effort, we will be able to design a team support tool to help better manage the necessary interruptions while mitigating the unnecessary interruptions for clinicians. These team support tools could be more informative dashboards or smart phone applications which could help clinicians obtain a better sense of when to push communications based on the overall workload of the emergency department.

Improving our understanding of team user's needs and interruption content and function may help in the interface design of systems in healthcare settings. Our application of the TURF framework in the design of a support tool managing the activity of interruptions can help provide further support in the importance of designing systems for clinicians. Patient safety and quality may improve through the enhancement of each clinician team's efficiency, effectiveness, coordination, communication, and, more broadly, future acceptance and use of possibly any implemented technology, such as health information technologies.

Limitations

This study used a limited convenience sample of participants within the trauma portion of a major emergency department. Consequently, these results may not be generalizable to other clinical settings. We were also unable to follow patients through their care outcomes and were therefore unable to verify if the impact of overhead interruptions greatly hampered the patient care process. Despite these limitations, we were able to outline the emergency department workflow, clinician communication and interaction patterns, and the function of interruptions.

Future Work

The scope of this work focuses on one emergency department. We are working to replicate the study in two other emergency departments using different health information technology systems, workflow, and patient sizes. Additionally, we are creating and testing more informative dashboards and communication tools (*e.g.*, smart phone applications) for clinicians to help better manage their workload and communications. We will compare the results of using these team support tools against the baseline findings of this study in determining how helpful the team support tools are to providing safer and faster patient care process outcomes.

Conclusion

In this paper, we explored interruptions between attendings, residents, and nurses in the emergency department environment. We looked beyond identifying the frequency of interruptions and considered the function of the interruptions. The interruptions we observed in the emergency department were classified into activities that supported required work and those that create unnecessary breaks in workflow. The results from this study will help inform our design of health information technology systems to support teams in better managing their communications.

Acknowledgements

We thank our participants and collaborators for facilitating and assisting us with this study. This research was supported by the Agency for Healthcare Research and Quality through a research grant (AHRQ R01HS021236-20) and a training grant (AHRQ T32HS017586).

References

1. Berg LM, Källberg AS, Göransson KE, Östergren J, Florin J, Ehrenberg A. Interruptions in emergency department work: An observational and interview study. BMJ Qual Saf. 2013;22:656-663.
2. Coiera E, Tombs V. Communication behaviours in a hospital setting: An observational study. BMJ. 1998;316:673-676.
3. France DJ, Levin S, Hemphill R, Chen K, Rickard D, Makowski R, Jones I, Aronsky D. Emergency physicians' behaviors and workload in the presence of an electronic whiteboard. Int J Med Inform. 2005;74:827-837.
4. Institute of Medicine. Health professions education: A bridge to quality. Washington (DC): National Academies Press; 2003.
5. Westbrook JI, Coiera E, Dunsmuir WTM, Brown BM, Kelk N, Paoloni R, Tran C. The impact of interruptions on clinical task completion. Qual Saf Health Care. 2010;19:284-289.
6. Grundgeiger T, Sanderson P. Interruptions in healthcare: Theoretical views. Int J Med Inform. 2009;78:293-307.
7. Rivera AJ, Karsh BT. Interruptions and distractions in healthcare: Review and reappraisal. Qual Saf Health Care. 2010;19(4):304-312.
8. Chisholm CD, Dornfeld AM, Nelson DR, Cordell WH. Work interrupted: A comparison of workplace interruptions in emergency departments and primary care office. Ann Emerg Med. 2011;38(2):146-151.
9. Li SYW, Magrabi F, Coiera E. A systematic review of the psychological literature on interruption and its patient safety implications. J Am Med Inform Assoc. 2012;19:6-12.
10. Crosskerry P, Cosby KS, Schenkel SM, Wears RL. Patient safety in emergency medicine. Philadelphia (PA): Lippincott Williams & Wilkins; 2009.
11. Kalisch BJ, Aebersold M. Interruptions and multitasking in nursing care. Jt Comm J Qual Patient Saf. 2010;36(3):126-132.
12. Jeanmonod R, Boyd M, Loewenthal M, Triner W. The nature of emergency department interruptions and their impact on patient satisfaction. Emerg Med J. 2010;27:376-379.
13. Zellmer-Bruhn ME. Interruptive events and team knowledge acquisition. Manag Science. 2003;49(4):514-528.
14. Adamczyk PD, Bailey BP. If not now, when?: The effects of interruption at different moments within task execution. CHI. 2004;6(1):271-278.
15. Allard J, Wyatt, Bleakley A, Graham B. "Do you really need to ask me that now?": A self-audit of interruptions to the 'shop floor' practice of a UK consultant emergency physician. Emerg Med J. 2012;29:872-876.
16. Brixey JJ, Robinson DJ, Turley JP, Zhang J. The roles of MDs and RNs as initiators and recipients of interruptions in workflow. Int J Med Inform. 2010;79:109-115.
17. Zhang J, Walji M. TURF: Toward a unified framework of EHR usability. J Biomed Inform. 2011;44(6):1056-1067.
18. Zhang J, Butler K. UFuRT: A work-centered framework and process for design and evaluation of information systems. HCI Int Proc. 2007;1-5.
19. Butler K, Zhang J, Esposito C, Bahrami A, Hebron R, Kieras D. Work-centered design: A case study of a mixed-initiative scheduler. ACM SIGCHI Int Proc. 2007;747-756.
20. Harrington C, Wood R, Breuer J, Pinzon O, Howell R, Pednekar M, Zhu M, Zhang J. Using a unified usability framework to dramatically improve the usability of an EMR module. Am Med Inform Assoc. 2011;549-558.
21. Franklin A, Liu Y, Li Z, Nguyen V, Johnson TR, Robinson D, Okafor N, King B, Patel VL, Zhang J. Opportunistic decision and complexity in emergency care. J Biomed Inform. 2011;44:469-476.
22. Brixey JJ, Robinson DJ, Johnson CW, Johnson TR, Turley JP, Patel VL, Zhang J. Towards a hybrid method to categorize interruptions and activities in healthcare. Int J Med Inform. 2007;76:812-820.

Pharmacovigilance on Twitter? Mining Tweets for Adverse Drug Reactions

Karen O'Connor[1], Pranoti Pimpalkhute[1], Azadeh Nikfarjam, MS[1], Rachel Ginn[1], Karen L Smith, PhD[2], Graciela Gonzalez, PhD[1]
[1]Arizona State University, Tempe, AZ; [2]Regis University, Denver, CO

Abstract

Recent research has shown that Twitter data analytics can have broad implications on public health research. However, its value for pharmacovigilance has been scantly studied – with health related forums and community support groups preferred for the task. We present a systematic study of tweets collected for 74 drugs to assess their value as sources of potential signals for adverse drug reactions (ADRs). We created an annotated corpus of 10,822 tweets. Each tweet was annotated for the presence or absence of ADR mentions, with the span and Unified Medical Language System (UMLS) concept ID noted for each ADR present. Using Cohen's kappa[1], we calculated the inter-annotator agreement (IAA) for the binary annotations to be 0.69. To demonstrate the utility of the corpus, we attempted a lexicon-based approach for concept extraction, with promising success (54.1% precision, 62.1% recall, and 57.8% F-measure). A subset of the corpus is freely available at: http://diego.asu.edu/downloads.

Introduction

Prescription drug usage continues to grow as part of health care strategies to increase health and improve quality of life. The National Center for Health Statistics (NCHS) reported that over a 10 year period, the percentage of those who were taking either 'one', 'two or more', or 'five or more' drugs increased 4%, 6% and 5% respectively.[2] While these drugs are prescribed for their therapeutic properties, their use may result in unintended or adverse effects.[3] An adverse drug reaction (ADR), as defined by the World Health Organization (WHO), is: "Any response to a drug which is noxious and unintended, and which occurs at doses normally used in man for the prophylaxis, diagnosis, or therapy of disease, or for the modifications of physiological function."[4] The morbidity and mortality rates associated with ADRs are considerable.

Pharmacovigilance is defined as "the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other possible drug-related problems".[5] ADRs are the primary focus of pharmacovigilance. ADRs may be detected in either the pre-marketing clinical trials or post-marketing surveillance of a drug. Post-marketing surveillance has traditionally been dependent on spontaneous reporting systems (SRS), which are maintained and monitored by various regulatory agencies, such as the FDA's Adverse Event Reporting System (FAERS). The FAERS provides consumers and providers with a system to report suspected ADRs. One of the main problems with SRS is that they are generally under-utilized due to the voluntary nature of reporting. A systematic review, conducted in 2006, estimated an under-reporting rate of 85-94%.[6] Recognizing the limitations of the SRS, the FDA launched an active surveillance initiative in 2008.

One of the FDA's Sentinel Initiative's aims is to use existing health data, such as electronic health records and claims data, to actively monitor for adverse event signals.[7] In addition to using existing health data to automatically detect ADRs, research has explored other data sources for this information. One such source is online social networks, in particular health related networks such as DailyStrength or PatientsLikeMe. Here, patients can freely discuss and share their experience with drug products. One of the first efforts to find ADRs in user posted comments was made by Leaman et al.[7] They used a modified lexicon based approach to extract drug and ADR relationships from user comments in the health networking and forum site DailyStrength.com. The authors concluded that user comments do contain extractable information regarding ADRs and these consumer reported effects do overlap with the known adverse effects. Generic social networking sites, such as Twitter, have also been found to contain valuable health related information. Recent research has shown that Twitter data analytics can have broad implications on public health research. Paul and Dredze demonstrated that public health topics, including syndromic information, geography based risk factors, and information about symptoms and medication, could be extracted from Twitter.[8] One limitation they note is the need for more data in order to produce better results.

Twitter is a valuable resource for researchers because the discussions are publicly available and can be accessed through the Twitter's streaming application programming interface (API). This unfettered access to volumes of up-to-date information is not without its challenges for natural language processing (NLP) researchers looking to

automatically extract relevant information. One difficulty is the paucity of relevant data, preliminary analysis demonstrated that less than 1% of tweets that included a drug name had an actual ADR mention. In contrast, our recent analysis of data from the health forum DailyStrength.com shows approximately 24% of the comments mention an ADR. Twitter would seem too sparse for exploration. However, with over 645 million users, 58 million tweets per day[9], and 80% of users posting tweets about themselves,[10] the value of Twitter as a source of similar data calls for careful reconsideration, even if one has to collect data over a longer period to get a large data volume.

Another challenge with Twitter is in the lack of structure in tweets. Tweets are restricted in character count, forcing users to condense their sentiments while still conveying the intended meaning. This leads to highly unstructured text that contains many abbreviations, some of which are unique to Twitter. They contain frequently misspelled words, colloquialisms, idioms, and metaphors that make automatic processing more difficult. For the task of mining for ADR mentions, these issues compound an already challenging problem.

Ritter et al. examined the performance of standard natural language processing tools (NLP) for named entity recognition (NER) in tweets.[11] They found that standard NLP tools performed poorly due to the compressed nature of the tweets, which eliminates the context needed to find an entity's type, and the relatively infrequent use of unique entity types, which makes it difficult to amass sufficient training data. They developed a set of NLP tools, specific to Twitter that outperformed the standard tools by a margin of 25%.

Research into the use of Twitter for detection of drug and adverse effect events has been sparse. Bian et al. utilized a data set of 2 billion tweets to explore the use of Twitter for real-time pharmacovigilance.[12] They chose five cancer drugs and mined the data set for mentions of those drugs. They then developed a SVM classifier to identify adverse effects caused by the drugs. Their results were qualified by the authors as "rather low", which they ascribe to: the nature of the tweets (unstructured, abbreviated words), the use of NLP tools that were created for more structured text and the use of nonmedical terms by the users that made matching difficult.

In this paper, we used a specifically selected corpus to evaluate the viability of Twitter as a source of ADR mentions and its potential value for pharmacovigilance. We anticipate that our findings could help augment other signals used to detect relationships between a drug and an adverse reaction.

Methods

To create a corpus of highly relevant data for our task, we collected tweets for 74 drugs (*target drugs*). A total of over 187,000 tweets that contained a mention of one of our target drugs were collected. After filtering out likely advertisements and balancing the data set, our final corpus consisted of 10,822 tweets. The corpus was annotated in two stages. The first was a binary annotation of each tweet indicating whether it contained an ADR mention or not. A combination of the tweets identified as containing an ADR and a randomly selected sample of 1,000 tweets that were marked as not containing a mention of an ADR were then annotated for specific concept spans and UMLS concept IDs. We annotated the spans and UMLS IDs for each adverse effect or indication mentioned, such that one tweet could have multiple mentions. We distinguished mentions about *adverse reactions* (an unexpected, negative effect resulting from the adequate use of the drug) from mentions about *indications* (the sign or symptom for which the drug was prescribed in the first place). This distinction is important, and it is a hard distinction to make even for annotators, let alone automated systems. A detailed description of these steps follows.

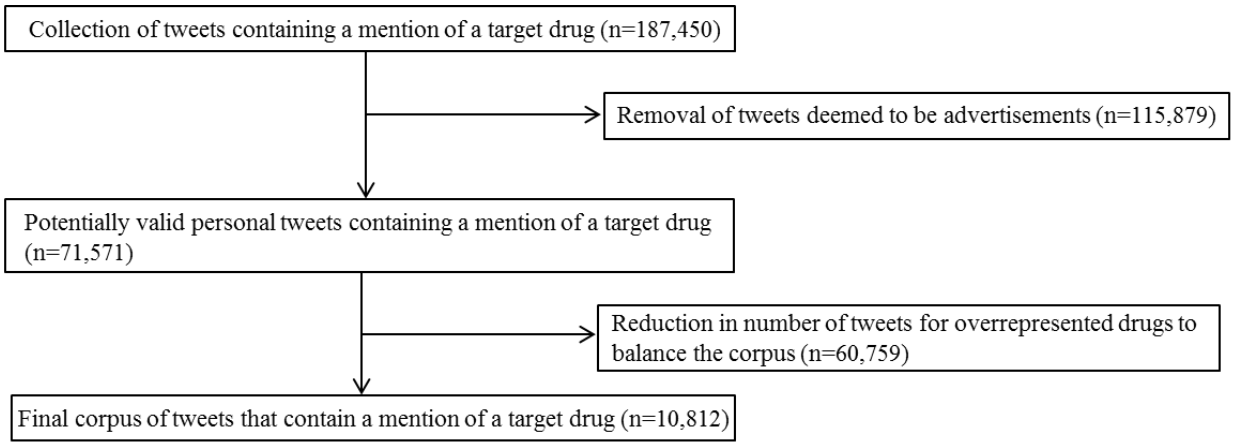
Data Acquisition: The 74 drugs included in this study are broadly used drugs whose adverse effects are well known (*truth set*), plus drugs released between 2007 and 2010, for which not all adverse effects are yet known and will only become visible as the drugs are more widely used (*test set*). Drugs in the truth set were selected on the basis of their widespread use, as demonstrated by their presence in the Top 200 products by volume in the US market published by IMS Health; and also based on whether they had a known adverse effect in at least one of the categories of interest. Many of the drugs for the truth set have come into widespread use recently, which allows for testing the capability of the NLP methods to confirm signals that are now known, but not so at the time of release. For the newer drugs, going back to 2007 allows for market growth leading to common prescribing and thus, likely presence in tweets. The list was narrowed based upon forecasts for widespread use, the prevalence of disease states and conditions, and on whether the drug was new in class. Major categories include drugs for central nervous system and mental health conditions such as Alzheimer's disease and schizophrenia. Treatments for age-related diseases (such as diabetes, cardiovascular diseases, urinary dysfunction, and musculoskeletal disorders) also met the criterion for potential widespread use, considering increased life expectancy. Additionally, biologics are an important class that was included because adverse events have traditionally been difficult to detect. Given the lag between FDA approval and recognition of

serious adverse events in this category, we selected four biologics to follow as part of the experimental set, and one is included in the truth set.

The drugs names, both the generic and brand, were used as keywords in the Twitter Search API. One of the first challenges we faced with acquiring the relevant tweets was accounting for a wide variety of drug name misspellings. To ensure that as much data as possible was obtained, we also used projected misspellings of drug names as keywords. The misspellings were created using a *phonetic spelling filter* that generates variants based on the phenomes of the correct spelling.[13] The filter generated many variants, and about 18% of them were added to the keyword list for reasonable coverage. Those were selected based on their Google search web prevalence.

The expanded drug list was used to search Twitter via the API. A total of 187,450 tweets were collected over a seven month period (August 2013 – February 2014). These required further filtering to remove advertisements and to balance the dataset among the drugs. This was accomplished by first removing texts containing URLs. Retweets were not removed. The number of tweets was thus reduced to 71,571. Next, to balance the number of tweets for each drug in the corpus, a maximum of 500 to 800 tweets for each drug name variant was randomly selected. This was necessary due to the vast imbalance of tweets containing mentions of some popular drug names as compared to others. Also, maintaining a soft limit, rather than a hard one, ensured that drugs that are highly commented upon also contain a higher number of samples in the corpus. This reduced the final corpus to 10,822 tweets. This process is summarized in Figure 1. A full description of the corpus appeared in Ginn et al.[14] A subset of the corpus is available online at http://diego.asu.edu/downloads.

Figure 1. Summary of tweets collected and subsequent reductions to obtain the final corpus.



Annotation. For annotation, we defined an adverse reaction as "an undesired effect of the drug experienced by the patient." This included mentions where the patient expressed the notion that the drug worsened their condition. An indication was defined as "the sign, symptom, syndrome, or disease that is the reason or the purpose for the patient taking the drug or is the desired primary effect of the drug. Additionally, the indication is what the patient, or prescriber believes is the main purpose of the drug." The annotated spans were mapped to UMLS concept IDs found in the lexicon.

Our lexicon was derived from our earlier work, presented in Leaman et al.,[7] and includes terms and concepts from four resources: the Coding Symbols for a Thesaurus of Adverse Reaction Terms (COSTART) vocabulary created by the U.S. Food and Drug Administration for post-market surveillance of adverse drug reactions (a subset of the UMLS Metathesaurus), which contains 3,787 concepts[15]; the side effect resource (SIDER), which contains 888 drugs linked with 1,450 adverse reaction terms extracted from pharmaceutical insert literature[16]; the Canada Drug Adverse Reaction Database, or MedEffect, which contains associations between 10,192 drugs and 3,279 adverse reactions.[17] We added terms from SIDER 2[16] and the Consumer Health Vocabulary (CHV)[18], which includes more colloquialisms.

The corpus was manually annotated in two stages by two expert annotators. The entire corpus was initially annotated for binary classification of the tweets as either '*hasADR*' for those with a mention of an adverse reaction, following the definition described above, or '*noADR*' for those without. The purpose of the binary classification was to validate our assumption that people would discuss their personal experiences with prescription drugs via tweets.

In the second stage of annotation, a combination of all tweets from the corpus classified as '*hasADR*' and a random selection of 1,000 tweets from the corpus that were classified as '*noADR*' were annotated to capture the spans in the

text that expressed either an adverse reaction or an indication. The spans of the mentions were annotated by the same annotators who performed the binary annotations following the rules set forth in the annotation guidelines. Some general guidelines include: *only capturing mentions if the concept was experienced by the patient*; *if the patient reported the concept as the reason for taking the drug than that mention would be labeled as an indication*; *all concepts were to be mapped to the UMLS id that most closely matched the meaning of the span*; and *when annotating a span, the smallest number of words needed to convey the meaning were chosen*. Table 1 illustrates a sample of tweets and their annotations.

Table 1. A sample of the tweets collected from Twitter and their annotations

Sample Comments	Classification	Annotations
20s 8th day with #Effexor still experiencing some side effects (drowsiness,sleepiness,GI effects). Moderate improvement in mood #depression	*hasADR*	"drowsiness" - *drowsiness: adverse effect*, "sleepiness" - *sleepiness: adverse effect*, "GI effect" – *gastro intestinal reaction: adverse effect*, "depression" – *depression: indication*
Over-eaten AGAIN just before bed. Stuffed. Good chance I will choke on my own vomit during sleep. I blame #Olanzapine #timetochange #bipolar	*hasADR*	"over-eaten" – *increased appetite: adverse effect*, "bipolar" – *bipolar disorder: indication*
@brokenmind_ Quetiapine was horrific for me in relation to wait gain. Such a horror story. But the weight will come off one day at a time.	*hasADR*	"wait gain" – *weight gain: adverse effect*
Tomorrow, my second infusion of Tysabri! Good luck for me! #Godblessme #MSLife	*noADR*	"MS" *multiple sclerosis:indication*
Do not take Cymbalta if you breathe – stolen from Tay	*noADR*	
Rules of Prozac: 1: You can never sleep, ever again. NEVER EVER 2: No you may NOT switch your brain off. Ever. 3: Exhaustion is your friend.	*hasADR*	"never sleep" – *insomnia: adverse effect*, "not switch your brain off" – *racing thoughts: adverse effect*, "exhaustion" – *exhaustion: adverse effect*
@FriarDanny I appreciate it. I gained over 30lbs with Paxil so I'm trying something different, tired of the appetite side effects.	*hasADR*	"gained over 30lbs" – *weight gain: adverse effect*, "appetite" – *increased appetite: adverse effect*
This cipro is totally "killing" my tummy .. hiks..	*hasADR*	""killing" my tummy" – *gastric pain: adverse effect*
Well played tysabri...kicking butt #nosleep.	*hasADR*	"nosleep" – *insomnia: adverse effect*
@Sectioned_ @bipolarlife7 @BBCWomansHour ah yes, I'm starting to think my paroxetine turns panic attacks into fat.	*hasADR*	"panic attacks" – *panic attack: indication*, "fat" – *weight gain: adverse effect*

The annotation of both adverse drug reactions and indications was necessary due to the fact that these two categories can have the same concepts associated with them. For example, a user may state "I take it for *insomnia*", while another may say, "Had to stop treatment, it was causing *insomnia*." The first mention is an indication and the second an adverse reaction but both would be mapped to the same UMLS concept id for *insomnia*. Thus, the type of the mention (adverse effect or indication) can vary, while the concept id can be the same.

Automatic Concept Extraction. We aimed at measuring the utility of lexicon-based techniques for the automatic concept extraction from Twitter data. Our proposed concept extraction method is based on an information retrieval technique that is flexible enough to deal with term variability in user posts. We used Apache Lucene[1] for both indexing and retrieval of the ADR lexicon concepts. A Lucene index was built from concepts and the associated UMLS IDs in the lexicon. Before indexing, we preprocessed the concepts which included removal of stop words and lemmatization. The lexicon entries were lemmatized to WordNet[2] roots using the Dragon toolkit[3].

To identify the indexed concepts present in a tweet, we generated a Lucene search from preprocessed tokens in the tweet sentences. We split tweets into sentences using Stanford Tokenizer[4]. The preprocessing of the sentences included spelling corrections, stop word removal, and lemmatization. For spelling corrections, we used Lucene SpellChecker that was customized to suggest a list of correct spellings for a given word based on ADR lexicon and a list of common English words from SCOWL[2] (Spell Checker Oriented Word Lists).

[1] http://lucene.apache.org
[2] http://wordnet.princeton.edu
[3] http://dragon.ischool.drexel.edu/features.as
[4] http://nlp.stanford.edu/software/tokenizer.shtml

The spans of the mentions associated with the retrieved lexicon concepts were then identified in the sentences using string comparison with regular expressions. The technique is flexible enough to identify both single and multi-word concepts, regardless of the order of the words in sentences or presence of other words in between.

Since the focus of this study is to extract adverse effects from the tweets, a manual filtering method was implemented to remove indications. This method filters the terms based on verbs that precede the term, such as "helps" or "works" for indications.

Results

Annotation statistics. After filtering out advertisements, the corpus contained tweets for 54 of the 74 target drugs. The number of tweets returned for four of these drugs was considerably higher than the rest and their numbers were reduced by making a random selection of tweets from each in order to balance out the dataset. From the resulting corpus of 10,822 tweets, 1,008 were determined during the binary phase of annotation to contain at least one mention of an adverse effect. These tweets were from 31 of the target drugs. The total number of adverse effects annotated was 1,285. The top ten drugs in order of the number of adverse effect mentions annotated are listed in Table 2.

Table 2. Top 10 drugs by sorted by the number of ADR mentions annotated. The generic name is indicated for those that were added to the original drug list for searching.

Drug Brand/Generic Name	ADR Mentions Annotated per Drug	Total Number of Tweets in Corpus
Seroquel/quetiapine	237	1,082
Effexor/venlafaxine	176	461
Vyvanse	122	800
Paxil/paroxetine	92	683
Prozac/fluoxetine	76	1,307
Lamictal/lamotrigine	73	395
Zyprexa/olanzapine	68	377
Humira	64	560
Cymbalta/duloxetine	63	832
Trazodone	58	530

Agreement between the annotators was measured by calculating inter-annotator agreement (IAA). We calculated IAA for span using a partial matching criterion. This means that the annotators were considered to be in agreement if there was some overlapping portion in the span selected by each annotator. IAA for concept ID and binary annotations was calculated using an exact matching criterion. The precision, recall and F-measures[19] associated with IAA of the annotations are shown in Table 3. For binary annotations, a kappa value of 0.69 was calculated.

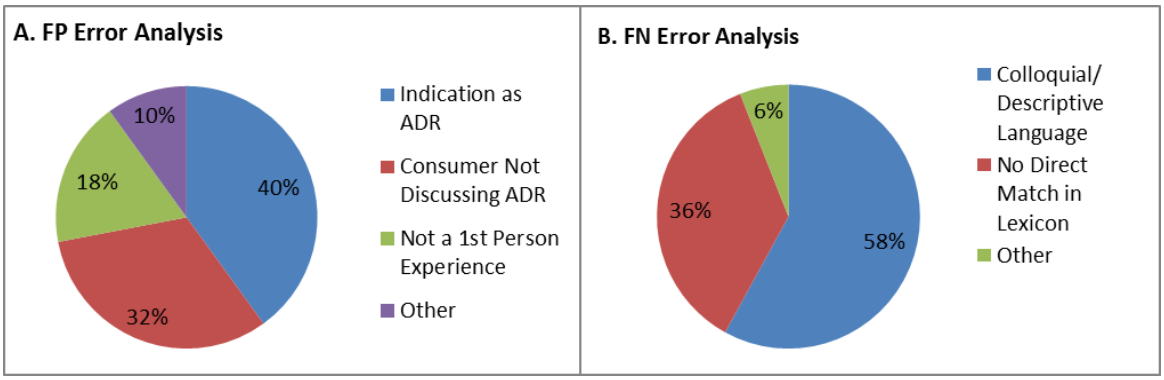
Table 3. IAA for span, concept ID and binary classification

IAA Type	Precision	Recall	F-measure
Span	0.8099	0.9802	0.8869
Concept ID	0.7278	0.9688	0.8311
Binary Classification	0.8155	0.8829	0.8448

Automatic concept extraction and identification. The system was evaluated on 1,873 annotated tweets, for adverse reactions only. We defined true positives as those extracted by the machine that match with the mentions labeled by annotators using partial matching. . Our system achieved 54.1% precision, 62.1% recall, and 57.8% F-measure for this experiment.

Error Analysis. For error analysis of the system, we chose a random sample of 50 false positive (FP) and 50 false negative (FN) results to review. We analyzed each group separately and categorized the errors by error type (Figure 2). For FPs, the errors fell into three main categories: the extraction of indications as ADRs, the extraction of terms from the tweets that were in the lexicon but not being used to discuss an ADR, and extraction of ADR mentions that were not experienced directly by the user. FN results were separated in to two main categories: the ADR was expressed using colloquial or descriptive terminology, or the ADR mention was expressed in a similar but not a direct match to the lexicon entry.

Figure 2. 2A shows the category and prevalence of the 50 FP errors analyzed. 2B shows the category and prevalence of the 50 FN errors analyzed.



Discussion

To assess the validity of finding ADRs in tweets, we compared the adverse effects that were found in the tweets to known and documented adverse effects. The results are reported in Table 4 for the top ten drugs as determined by the number of ADR mentions annotated. We found at least one corresponding adverse effects in all ten. In a few instances, the most frequent adverse event in the tweets aligned with the most frequently found adverse event in the clinical trials. We only assessed the relationship between the most frequent adverse reactions. Given the known adverse effects were from clinical trials, it may be interesting to examine less frequently occurring events in the trials as compared to the mentions. The relatively small sample size and short duration of the trial may not adequately represent such rare events, but these may get elucidated in the wide spread, long term Twitter evidence.

Compared to the 78.3% precision and 69.9% recall we reported in Leaman et al.[7], for the DailyStrength corpus for concept extraction, we see that the performance against the Twitter dataset is much lower. This could be due to the inherent characteristics of a tweet, with greater presence of "creative" expressions that do not match even the augmented lexicon. There is evidence of this when we look at the sources of FN errors. The largest source of these errors was the language used to describe the ADR.

Some of these issues stem from using inexact terminology to describe the effect, which has no match in the lexicon; examples include 'in a haze' or 'wired'. Not surprisingly, this issue also arises when the user uses an idiomatic expression or metaphors to describe the effect. For example, to express the adverse effect of dry mouth, a user tweeted "….*it feels like the Sahara desert in my mouth*". This was also the largest reason for errors reported by Leaman et al.[7], but with a slightly smaller percentage of their associated FNs (55%). In order to mitigate these errors, other approaches that are less reliant on lexicon matching are being explored. Nikfarjam and Gonzalez[21] have proposed using association rule mining to extract ADR mentions from colloquial text, extracting frequent patterns to find mentions instead of using a dictionary match. Pattern-based extraction showed improvements over the lexicon based approach, but the method requires a large amount of training data.

The FP errors were mainly due to misclassifying extracted terms as ADRs when they were not used as such. The biggest source of these errors was instances where an indication was misclassified as an ADR. There were also FP errors caused by the extraction of terms that were not used for discussing either an ADR or an indication. Some examples of this include a user name that contains a lexicon term ('*TScpCancer*') or the use of a term that is not necessarily related to an ailment ('*Sick* of meds'). Although the post-processing rules were effective in filtering some of the non ADR mentions, machine learning classification is needed in future research to examine the effectiveness of modeling the contextual and semantic features of the tweets in distinguishing different semantic types. In the future, we will explore the effectiveness of training a sequence labelling classifier such as Conditional Random Fields (CRFs) for this extraction task.

Another smaller, though notable, source of errors for FPs was the system's extraction of adverse effects that were not experienced by the user. Our guidelines for annotation stipulated that mentions of adverse effects were to be annotated only if they were experienced by the patient, in order to prevent annotations of song/movie quotes, social gossip, or potential advertisements. This presents a challenge when utilizing a social networking site that is not devoted solely to health topics. The unstructured, informal, and isolated nature of the tweets compounds this challenge by making some tweets unclear. It can be difficult for annotators to discern the relationship of the person writing the tweet with the drug they are mentioning. There are tweets that seem to be a commentary or report on the drug and its side effects ("Effexor XR side effects : - sucidal thoughts - insomnia - feeling high & unresponsible. -Dry mouth."), or of a song lyric ("I'm too numb to feel, blow out the candle, blindness #np #cymbalta") but without any context of why the person is tweeting about it, it is difficult to know. It could be a way to relay a personal experience or a repetition of something they heard. The determination of whether these are first person accounts requires the annotator to make a judgment call or the use of specialized knowledge about commonly used conventions in Twitter. In the song example above, using a common abbreviation to express a song you want to share on Twitter is #np ('*now playing*') and this is often followed by #name_of_artist. However, the user who used #Cymbalta possibly indicated they are ascribing the meaning of the lyric to the drug. Training a system on the nuances of Twitter conventions, or being able to discern 3rd person from 1st person accounts, is a very significant challenge — especially when tweets are preprocessed with techniques that affect symbol and punctuation placement.

From Table 2, we can see that the number of tweets collected for a drug is not necessarily indicative of the number of ADRs that we can expect to find. This may be attributed to the fact that Twitter is a general social networking site and will contain a lot of noise when attempting to utilize its data for a specific research purpose. The prevalence of tweets may be due to the drug being the news, or the drug may have a catchy slogan used in its advertisement that gets tweeted frequently (such as Cymbalta's "Depression hurts, Cymbalta can help"). To try to eliminate some of the noise, the binary data in the annotated corpus will be used in future work to refine our methods for searching and filtering for tweets that may contain ADRs.

In this study we aimed at exploring Twitter as a resource for the automatic extraction of ADRs, and the effectiveness of the commonly used lexicon-based techniques for ADR extraction from tweets. In the future, with more annotated data available, we will examine the utility of state-of-the-art machine learning classifiers such as CRF for concept extraction. Our current study has a number of limitations, which are as follows:

- Modified distribution of comments: Although we attempted to enforce a soft limit on the number of comments for each drug in the annotation set, their distributions in the annotation set are not representative of their actual distribution in real life data. A thorough analysis is required to identify if this will affect the performance of our system when applied to real life data.
- For some drugs, we were only able to collect a small number of tweets. More tweets associated with those drugs will be required to make reliable predictions in future tasks.
- We have not performed detailed experimentation to assess how the data imbalance problem posed by Twitter data affects performance of ADR detection systems. We will address this aspect of research in the future.

Conclusion

We have shown that people do tweet about their adverse effects experiences with their prescription drug use. In these tweets, they mention the drug name, along with the adverse effect(s), making it possible to automatically extract the drug and adverse effect relationship. We were, however, only able to achieve moderate success with the lexicon based system used to automatically extract the adverse effect mentions. A large part of this difficulty is due to the problem of using a formal lexicon to match to colloquial text. Another significant problem is the large data imbalance associated with data collected from Twitter. In the future, we will also explore machine learning algorithms that take into account the data imbalance issue with Twitter data. Our continuing efforts at annotation of data makes the application of supervised learning approaches a lucrative future possibility.

Table 4. List of drugs included in preliminary analysis**,** with their most common adverse reactions, frequency of incidence in adults, as listed in the FDA's online drug library.[22] The last column shows the most frequent adverse effects extracted from the Twitter data using our automated system. Effects found in both are highlighted in bold.

Drug Brand/Generic Name	**Primary Indications**	**Documented Adverse Effects (no order)**	**Adverse Effects Found in Tweets (Frequency)**
Seroquel/ Quetiapine	Schizophrenia, Bipolar I Disorder: manic episodes, Bipolar Disorder	**somnolence**, dry mouth, headache, **dizziness**, asthenia, constipation, fatigue	**somnolence (22.2%)**, abnormal dreams (9.6%), feel like a zombie (8.1%), weight gain (6.6%), restless leg syndrome (6.6%), increased appetite (5.9%), sleep paralysis (2.9%), **dizziness (2.2%)**, psychosis (2.2%), tremors (2.2%)
Effexor/ venlafaxine	Major Depressive Disorder (MDD)	**nausea**, **headache**, somnolence, dry mouth, dizziness	withdrawal syndrome (21.3%), insomnia (11.1%), **headache (4.3%)**, malaise (4.3%), abnormal dreams (4.3%), **nausea (3.4%),** shaking (3.4%), fatigue (3.4%)
Vyvanse	ADHD	decreased appetite, **insomnia**, dry mouth, diarrhea, nausea	**insomnia (38.2%)**, OCD (9.3%), anger (5.6%), heart racing (5.6%), depression (3.6%), psychosis (3.6%), headache (3.6%), feel weird (3.6%)
Paxil/ Paroxetine	MDD, Obsessive Compulsive Disorder (OCD), Panic Disorder, Social Anxiety Disorder, Generalized Anxiety Disorder (GAD), PTSD	nausea, **somnolence**, abnormal ejaculation, asthenia, tremor, insomnia, sweating	withdrawal syndrome (27.7%), weight gain (12.8%), depression (8.5%), headache (6.4%), **somnolence (6.4%)**, allergic (6.4%), feel sick (6.4 %), emotional (6.4%)
Prozac/ Fluoxetine	MDD, OCD, Bulimia Nervosa Panic Disorder	nausea, headache, insomnia, nervousness, anxiety**, somnolence**	**somnolence (22.2%),** withdrawal syndrome (8.9%), feeling ill (8.9%), abnormal dreams (6.7%), suicidal thoughts (6.7%), tremors (6.7%), allergic reaction (4.4%)
Lamictal lamotrigine	Epilepsy, Bipolar Disorder	vomiting, coordination abnormality, dizziness, rhinitis, dyspepsia, nausea, headache, diplopia, ataxia, **insomnia**, fatigue, back pain	**insomnia (17.9%)**, rash (12.8%), lethargy (7.7%), joint pain (5.1%), feel like a zombie (5.1%), feel sick (5.1%)
Zyprexa/ olanzapine	Schizophrenia, Bipolar I Disorder	dizziness, constipation, personality disorder, **weight gain**, akathisia, **somnolence**, dry mouth, asthenia, dyspepsia	**weight gain (40.0%)**, **somnolence (11.4%)**, increased appetite (8.6%), dependence (5.7%)
Humira	Rheumatoid Arthritis, Juvenile Idiopathic Arthitis, Psoriatic Arthitis, Crohn's Disease, Ulcerative Colitus, Plaque Psoriasis	**upper respiratory infection**, **rash**, **headache**, sinusitis, accidental injury	somnolence (24%), feel sick (8%), palpitations (8%), ache/pains (8%), joint pain (4%), **headache** (4%), **rash** (4%), **respiratory disorder (4%)**
Cymbalta/ Duloxetine	MDD, GAD, Diabetic Peripheral Neuropathy, Fibromyalgia, Chronic Muscoskeletal Pain	nausea, headache, **dry mouth**, **fatigue**, **somnolence**	withdrawal syndrome (16.3%), **fatigue (14.0%)**, **somnolence (7.0%)**, dizziness (7.0%), **dry mouth (4.7%)**, depression (4.7%), rash (4.7%), migraine (4.7%)
Trazodone	MDD	**somnolence**, **headache**, dry mouth, dizziness, nausea	**somnolence (24.3%)**, abnormal dreams (16.2%), hangover effect (8.1%), **headache (5.4%)**, insomnia (5.4%), withdrawal syndrome (5.4%)

.

References

1. Carletta J. Assessing agreement on classification tasks: the kappa statistic. *Comput Linguist*. 1996;22(2):249–254.

2. Gu Q, Dillon CF, Burt V. *Prescription Drug Use Continues to Increase: U.S. Prescription Drug Data for 2007-2008*. Available at: http://www.cdc.gov/nchs/data/databriefs/db42.pdf. Accessed March 8, 2014.

3. Edwards IR, Aronson JK. Adverse drug reactions: definitions, diagnosis, and management. *Lancet*. 2000;356(9237):1255–9. doi:10.1016/S0140-6736(00)02799-9.

4. Safety of Medicines - A Guide to Detecting and Reporting Adverse Drug Reactions - Why Health Professionals Need to Take Action: References. *World Heal Organ*. 2002. Available at: http://apps.who.int/medicinedocs/en/d/Jh2992e/12.html. Accessed March 9, 2014.

5. Lindquist M. The Need for Definitions in Pharmacovigilance. . *Drug Safety 2007*. 30(10).

6. Hazell L, Shakir SA. Under-Reporting of Adverse Drug Reactions: A Systematic Review. *Drug Safety 2006*. 2006;29(5):385. 12p. 4 Charts.

7. Platt R, Wilson M, Chan KA, Benner JS, Marchibroda J, McClellan M. The New Sentinel Network — Improving the Evidence of Medical-Product Safety. *N Engl J Med*. 2009;361(7):645–647. doi:10.1056/NEJMp0905338.

8. Paul M, Dredze M. You are what you tweet: analyzing Twitter for public health. In: *Proceedings of the Fifth International AAAi Conference on Weblogs and Social Media*.; 2011:265–72. Available at: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2880/3264. Accessed March 9, 2014.

9. Twitter Statistics | Statistic Brain. Available at: http://www.statisticbrain.com/twitter-statistics/. Accessed March 9, 2014.

10. Naaman M, Boase J, Lai C-H. Is it really about me? In: *Proceedings of the 2010 ACM conference on Computer supported cooperative work - CSCW '10*. New York, New York, USA: ACM Press; 2010:189. doi:10.1145/1718918.1718953.

11. Ritter A, Clark S, Mausam, Etzioni O. Named entity recognition in tweets: an experimental study. 2011:1524–1534. Available at: http://dl.acm.org/citation.cfm?id=2145432.2145595. Accessed March 13, 2014.

12. Bian J, Topaloglu U, Yu F. Towards large-scale twitter mining for drug-related adverse events. In: *Proceedings of the 2012 international workshop on Smart health and wellbeing - SHB '12*. New York, New York, USA: ACM Press; 2012:25. doi:10.1145/2389707.2389713.

13. Pimalkhute P, Patki A, Nikfarjam A, Gonzalez GH. Phoenitic Spelling Filter for Keyword Selection in Drug Mention Mining from Social Media. In: *AMIA Summit*.; 2014.

14. Ginn R, Pimalkhute P, Nikfarjam A, et al. Mining Twitter for adverse drug reaction mentions: A corpus and classification benchmark. In: *BioTexM*.; 2014.

15. Coding Symbols for Thesaurus of Adverse Reaction Terms (COSTART) Source Information. Available at: http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/CST/. Accessed April 9, 2014.

16. Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P. A side effect resource to capture phenotypic effects of drugs. *Mol Syst Biol*. 2010;6:343. doi:10.1038/msb.2009.98.

17. MedEffect Canada - Health Canada. (n.d.). Available at: http://hc-sc.gc.ca/dhp-mps/medeff/index-eng.php. Accessed April 11, 2014.

18. Zeng-Treitler Q, Goryachev S, Tse T, Keselman A, Boxwala A. Estimating consumer familiarity with health terminology: a context-based approach. *J Am Med Inform Assoc*. 2008;15(3):349–56. doi:10.1197/jamia.M2592.

19. Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *J Am Med Inform Assoc*. 2013;20(5):806–13. doi:10.1136/amiajnl-2013-001628.

20. Leaman R, Wojtulewicz L, Sullivan R, Skariah A, Yang J, Gonzalez G. Towards Internet-Age Pharmacovigilance : Extracting Adverse Drug Reactions from User Posts to Health-Related Social Networks. *Proc 2010 Work Biomed Nat Lang Process*. 2010;(July):117–125.

21. Nikfarjam A, Gonzalez GH. Pattern mining for extraction of mentions of Adverse Drug Reactions from user comments. *AMIA Annu Symp Proc*. 2011;2011:1019–26. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3243273&tool=pmcentrez&rendertype=abstract. Accessed March 5, 2014.

22. FDA. Drugs@FDA. Available at: http://www.accessdata.fda.gov/scripts/cder/drugsatfda/. Accessed March 13, 2014.

Computerization of Mental Health Integration Complexity Scores at Intermountain Healthcare

Thomas A. Oniki, PhD[1], Drayton Rodrigues, MBBS, MS[1], Noman Rahman[1], Saritha Patur[1], Pascal Briot, MBA[1], David P. Taylor, PhD[1], Adam B. Wilcox, PhD[1], Brenda Reiss-Brennan, PhD[1], Wayne H. Cannon, MD[1]
[1]Intermountain Healthcare, Salt Lake City, UT

Abstract

Intermountain Healthcare's Mental Health Integration (MHI) Care Process Model (CPM) contains formal scoring criteria for assessing a patient's mental health complexity as "mild," "medium," or "high" based on patient data. The complexity score attempts to assist Primary Care Physicians in assessing the mental health needs of their patients and what resources will need to be brought to bear. We describe an effort to computerize the scoring. Informatics and MHI personnel collaboratively and iteratively refined the criteria to make them adequately explicit and reflective of MHI objectives. When tested on retrospective data of 540 patients, the clinician agreed with the computer's conclusion in 52.8% of the cases (285/540). We considered the analysis sufficiently successful to begin piloting the computerized score in prospective clinical care. So far in the pilot, clinicians have agreed with the computer in 70.6% of the cases (24/34).

Introduction

Frequently in today's primary care setting, providers are confronted with diagnosing and treating mental health concerns. Gatchel and Oordt estimated that 70% of primary care visits stem from psychosocial issues.[1] And Kessler et al. found that almost 50% of mental health disorders were treated in primary care.[2] Unfortunately, as Collins et al. lament, "most primary care doctors are ill-equipped or lack the time to fully address the wide range of psychosocial issues that are presented by the patients."[3] Indeed, Vermani et al. cite evidence of low detection rates and poor quality of care for a variety of mental health disorders in the primary care setting.[4]

In this context, researchers at Intermountain Healthcare have sought to assist the Primary Care Physician (PCP) with the task of managing mental health issues in the primary care setting. Intermountain has organized itself around key clinical processes and developed "Clinical Programs"[5] in nine clinical areas to address those processes. In multi-disciplinary fashion, the clinical programs identify best practices and standardize the care delivered within the organization. The Primary Care Clinical Program is one such program. Within the Primary Care Clinical Program, Intermountain researchers and clinicians have developed a Mental Health Integration (MHI) program[6,7] that espouses a patient-centered, team-based approach to diagnosing and treating mental illness in the primary care setting. The central concept of the program is to eliminate "the chasm between physical health care and mental health care—rolling both into a comprehensive whole that is addressed by a high-functioning team."[6] The program standardizes processes and supports the PCP with expanded team support (including mental health professionals, community advocates, and care management) to address the mental health needs of their patients and families. Each member of the team is trained in specific responsibilities (communication, shared decision-making, etc.) that contribute to a collective whole health care plan. A standardized process is outlined via a Care Process Model (CPM).

A primary objective of the CPM is to determine which of three levels of care is appropriate for the patient -- routine (PCP-based) care, collaborative care (which adds care manager and mental health specialist participation), or enhanced care (which further increases care manager and mental health specialist participation). The care determination is made by assessing "Mental Health Complexity." A complexity of "mild" suggests routine care is appropriate, a complexity of "moderate" suggests collaborative care, and a complexity of "high" suggest enhanced care.

To assess complexity, the PCP asks the patient to fill out a paper-based "MHI packet" as a screening mechanism when a patient presents to the PCP with possible mental health concerns. The packet solicits 47 pieces of data from the patient on 21 facets related to mental health (some facets involve more than one piece of data) such as:

- Previous mental health treatment
- Level of chronic pain

- Family Rating Scale (assessing level of support of family relationships)
- Validated tools for diagnosing and treating patients with mental health conditions, such as the Patient Health Questionnaire (PHQ-9)

Our researchers have developed classification criteria for each facet that classify a patient as "mild," "moderate," or "high" complexity for the facet. (These will be referred to as "sub-scores" for the rest of the paper.) For illustration purposes, Table 1 shows the criteria for several sub-scores. Some sub-score classification criteria, such as those for the Patient Health Questionnaire, are drawn from published, validated instruments. Other sub-score criteria, such as those for chronic pain severity, are based on the experience of Intermountain clinicians and researchers. Over the past decade, Intermountain MHI researchers have refined and formalized the sub-score complexity classification criteria and have documented and published them internally.

Currently, a portion of the patient responses (corresponding to 14 of the 21 facets) is transcribed by clinical staff into Intermountain's HELP2[8] clinical information system. This portion represents those items most frequently filled in by patients and most influential in determining complexity based on our experience. The PCP is then asked to apply the sub-score criteria to the information to obtain sub-scores. Finally, the PCP aggregates the sub-scores into an overall "high"/"moderate"/"mild" complexity score that will dictate the initial level of care for the patient. The guidelines for aggregation have heretofore been incompletely specified; examples of sub-scores and corresponding overall complexity had been published internally, but logically-complete criteria had not been defined. The PCP reviews the sub-scores and the examples and makes his or her best judgment of an overall complexity.

The MHI program has shown positive clinical and financial outcomes.[6,7] However, improving the consistency and appropriateness of complexity scores is critical to the long-term success of the program. Consequently, the MHI researchers sought the assistance of the Intermountain Medical Informatics group to automate the scoring process. Together, we reasoned that computerization would accrue three benefits:

1. More scores will be generated. Clinician scoring has not been mandatory. Thus, in the busy clinic environment, some patients do not receive a score, even though the patient has filled out a Mental Health packet. We theorized that more scores would result in more appropriately directed care.
2. Scores will be more consistent. Undoubtedly, there has been variability in clinicians' scoring – especially in the aggregation of sub-scores into an overall complexity. Computerization would process inputs objectively and consistently, promoting standardization of care and enabling more reliable stratification of patients for research purposes.
3. Scores will be more appropriate. There is no "correct" or "gold standard" overall complexity; our clinician researchers have conceived of it in experience-based fashion to assist the PCPs in managing issues outside their expertise. We anticipate, though, that we can continually improve the complexity criteria as our researchers and the PCPs learn from each other through an iterative process of applying the criteria to patients, evaluating the care levels dictated, and refining again.

Table 1. An extract of sub-score criteria for several sub-scores.

Mild	Moderate	High
Previous mental health treatment: 1 episode	**Previous mental health treatment:** ≥ 2 episodes	**Previous mental health treatment:** multiple treatment failures
Chronic Pain Severity: 0-3 on a scale of 10	**Chronic Pain Severity:** 4-6 on a scale of 10	**Chronic Pain Severity:** 7-10 on a scale of 10
Family Rating Scale: Style III (balanced/secure)	**Family Rating Scale:** Style II (confused/chaotic)	**Family Rating Scale:** Style I (disconnected/avoidant)
Patient Health Questionnaire: Symptom score: <5 (out of a list of 9) Severity score: 10-14 (maximum possible=27)	**Patient Health Questionnaire**: Symptom score: >=5 (out of a list of 9) Severity score: 15-19 (maximum possible=27)	**Patient Health Questionnaire**: Symptom score: >=5 (out of a list of 9) Severity score: >=20 (maximum possible=27)

Our plan was as follows:

- Implement the classification criteria in Intermountain's decision support system
- Execute the classification criteria on retrospective data
- Analyze results and refine the classification criteria
- Pilot the classification criteria with several primary care providers and solicit their feedback
- Make necessary refinements

Once these steps are completed, the intent is to implement the classification in Intermountain's production system. When the patient's packet information is entered, the computer will deduce a complexity and present it to the patient's PCP, thus suggesting a level of care to begin with. The PCP will be able to either accept that level of care or set another level based on his or her clinical judgment.

We report here on our efforts through the piloting step.

Materials and Methods

Availability of data in the system

The first step in implementation of the algorithm was to verify the availability of the input data. We verified that the 24 pieces of data needed to conclude the 14 sub-scores were available in HELP2. Table 2 shows the 14 sub-scores included in the computerized algorithm, with the number of data items involved in concluding each sub-score. These

Table 2. The 14 sub-scores addressed by the computerized algorithm

Sub-score	Category	Number of Data Items (Data Items)
Number of Somatic Complaints	S1	1 (Number of Somatic Complaints, 0-9)
Chronic Pain Severity	S1	1 (Chronic Pain Severity, 0-10)
Sleep Problem Severity	S1	1 (Sleep Problem Severity, 0-10)
Substance Use	S1	1 (Current Substance Use, Y/N)
Overall Impairment	S2	1 (Impairment Rating Score, 1-7)
Overall Health	S2	1 (Overall Health, 1-10)
Family Relational Style	O1	3 (Family Style I Score, Family Style II Score, Family Style III Score)
Family Pattern Profile (Support Person)	O1	1 (Most Common Support)
Patient Health Questionnaire (PHQ-9)	O1	2 (PHQ-9 Symptom Count, PHQ-9 Severity Score)
Suicide Assessment	O2	3 (PHQ-9 Q9 response, Suicide State, Suicide Risk)
Anxiety/Stress Disorders (GAD-7)	O3	3 (GAD-7 Q1 Score, GAD-7 Q2-5 Score, GAD-7 Q6-7 Score)
Mood Disorder Questionnaire (MDQ)	O3	3 (MDQ Q1 Score, MDQ Q2 Response, MDQ Q3 Response
Mood Regulation	O3	2 (Symptom Score, Impairment Score)
Attention Deficit Hyperactivity Disorder (ADHD)	O3	1 (ASRS-v1.1 Part A Score)
Total number of sub-scores: 14		**Total number of data items: 24**

PHQ-9: Patient Health Questionnaire-9[9]
GAD-7: Generalized Anxiety Disorder-7[10]
MDQ: Mood Disorder Questionnaire[11]
ASRS-v1.1: Adult ADHD Self-Report Scale (ASRS) Version 1.1[12]

24 data items became the basis of the computerization.

Adequately expressive logic

Morris[13] asserts, "An adequately explicit clinical method is one that contains adequate detail to generate specific instructions (patient-specific orders) without requiring judgments by the clinician." Adequate explicitness is essential in the computerization of clinical decision support criteria.[13-16] We held several discussions in which we strengthened the explicitness of the sub-score algorithms. For example, the last row of Table 1 shows the original logic for deducing one particular sub-score – the sub-score for the PHQ-9.

Examination found that cases such as the following would have indeterminate results:

- A symptom count < 5 with a severity score < 10
- A symptom count of 5 with a severity score of 10-14
- A symptom count of 5 with a severity score < 9

As we completed the logical "decision table," we recognized that the "symptom" component was actually superfluous to the logic. Refining the criteria resulted in the simplification shown in Table 3.

Discussions regarding boundary conditions, logical operations (clarification of ORs vs. ANDs), and logical completeness resulted in fully computerizable sub-score classification criteria.

The next step was to fully specify the criteria for aggregating sub-scores into an overall complexity. In the clinical judgment of our MHI researchers, subjective data items contributed differently to complexity than did objective items. They consequently categorized each sub-score as either a subjective (S) or an objective (O) measure, depending on whether the data contributing to the sub-score were subjective or objective in nature. They further divided the subjective scores into two subcategories (S1 and S2), the S1 scores being more critical (based on their clinical experience) to the overall complexity. The resulting categories are shown in Table 2. This allowed the researchers to develop the empiric rules shown in Figure 1.

Note the rules were arranged in waterfall fashion (IF-THEN-ELSEIF), the last condition being

ELSE overall complexity = "mild"

In essence, the "default" complexity in the absence of data was "mild," i.e., the patient should receive routine care.

Table 3. A simplification of the sub-score criteria for the Patient Health Questionnaire**.**

Mild	Moderate	High
Patient Health Questionnaire: Severity score: <15 (maximum possible=27)	**Patient Health Questionnaire:** Severity score: 15-19 (maximum possible=27)	**Patient Health Questionnaire:** Severity score: >=20 (maximum possible=27)

IF O2 = HIGH THEN OVERALL = HIGH
ELSE IF O2 = MODERATE AND 1 or more O1s = HIGH THEN OVERALL = HIGH
ELSE IF O2 = MILD AND 1 or more O1s = HIGH AND 2 or more O3s = HIGH THEN
OVERALL = HIGH
ELSE IF 2 or more S1s = HIGH AND 1 or more S2s = HIGH AND 1 or more Os = HIGH THEN
OVERALL = HIGH
ELSE IF 1 or more SUB-SCORES = HIGH THEN OVERALL = MODERATE
ELSE IF 2 or more S1s = MODERATE AND 1 or more S2s = MODERATE AND 1 or more Os
= MODERATE THEN OVERALL = MODERATE
ELSE OVERALL = MILD

Figure 1. Logic for aggregating sub-scores into overall complexity.

Computerization and execution

Two medical knowledge engineers converted the classification criteria into Java code within Foresight, Intermountain's existing Computerized Decision Support (CDSS) system.[17] Foresight is a J2EE-compliant, Intermountain-developed decision logic execution engine coupled with a sophisticated clinical data monitor. Foresight rules are written in Java and deployed in an Oracle Weblogic server.

We executed the classification criteria on 500 patients for whom providers had entered a complexity score in HELP2 between August and November of 2012. The results were exported into Oracle tables and made available to the clinical experts for review and analysis.

Preliminary Analysis

We narrowed our preliminary analysis to 273 patients who had suicide assessment (the most critical facet to complexity) or greater than 80% of the 24 items. In this population, the clinician agreed with the computer in 53.1% of cases. An MHI researcher (BR) reviewed the results of the classification criteria for twenty randomly sampled cases. Upon reviewing the patient data, she judged that the classification was not weighing two factors heavily enough: an MHI packet question regarding the patient's level of thoughts of death and suicide, and a question regarding the patient's access to a support person. We recognized this would result in a general increase in complexity – possibly higher than the clinicians might tend to score – but we reasoned that at this learning and improvement stage it was better to "overstate" complexity (and hence "over-involve" the mental health specialists than "under-involve" them). If, as a result, the specialists told the PCPs their involvement was unnecessary, learning would occur and the criteria could be adjusted. We adjusted the criteria to be more sensitive to these factors.

We re-ran the classification on the 2012 patients and on 500 patients whose complexity was entered between June and October of 2013. Again, we focused analysis on patients who either had suicide data entered or who had greater than 80% of the 24 data items entered. These criteria left 540 patients to analyze.

Results

Retrospective Data Analysis Results

Clinician agreement frequency on the set of 540 patients is shown in Table 4. The clinician agreed with the computer in 285 of the 540 cases (52.8%). Of the 255 cases in which the provider and the computer disagreed, the algorithm was "higher" 179 times (70.2%). A graphical view of the frequencies, comparing the frequency of cases in which the clinician agreed with the computer, cases in which the clinician was "higher," and cases in which the computer was "higher" is shown in Figure 2.

We (BR and PB) performed a case-by-case examination of the forty cases in which the clinician gave a complexity score of "high" while the computer gave a lower score, to gain greater insight into such cases and ensure that the algorithm was not "missing" complexity it should have been detecting. We reviewed charts to determine what data might have influenced the providers to score the patients as "high." In most of the cases, we found evidence of comorbidities (diagnoses or multiple medication orders). Comorbidities are considered in the paper version of the algorithm, but were not included in this first version of the computerized algorithm, i.e., they were not included in the "top 14" sub-scores that were included in HELP2 and computerized.

In some cases, chart review did not reveal the reasons for the provider's score of "high." We reasoned that

Table 4. Clinician and computer agreement in the retrospective data.

		Computer			
		Mild	Moderate	High	TOTAL
Clinician	Mild	47	89	14	150
	Moderate	36	178	76	290
	High	4	36	60	100
	TOTAL	87	303	150	540

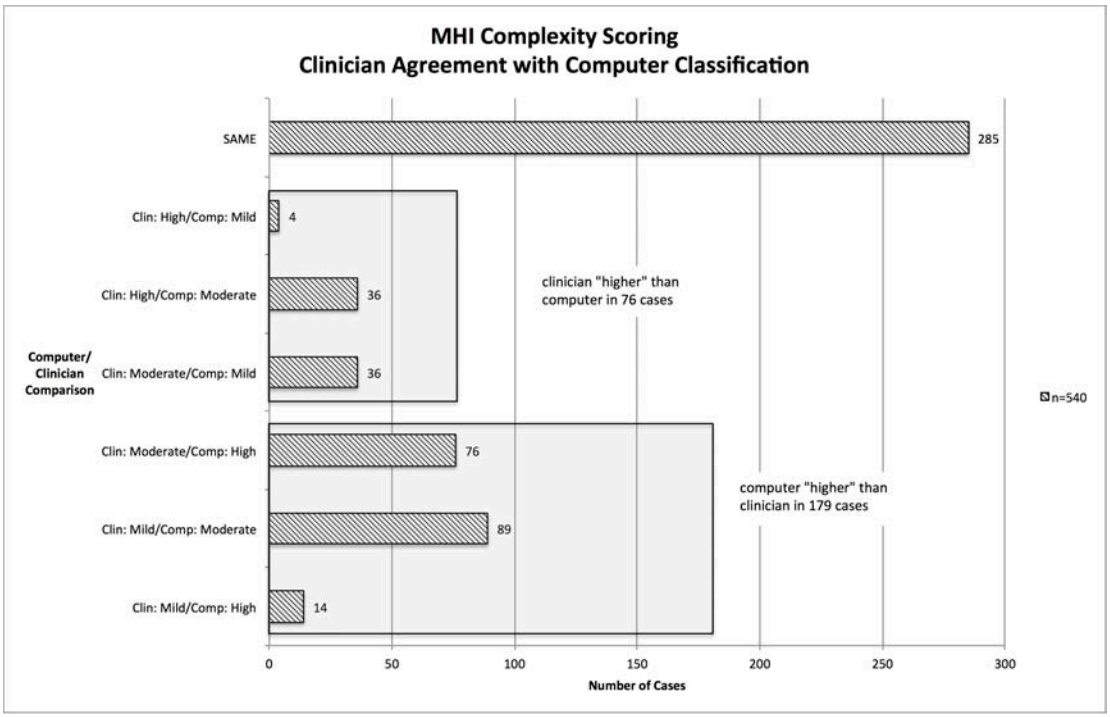


Figure 2. Clinician agreement with computer classification.

occasionally the computer algorithm might not be able to generate a "high enough" score because it did not have access to the proper data elements, and in such situations, the provider seemed to be properly disagreeing.

In general, we hypothesized that disagreements could have stemmed from several sources:

- The clinician and the computer may have been using different sets of inputs. There may be factors that are salient to complexity that the clinician was privy to but that were not available to the classification criteria.
- The clinicians may have disagreed with the MHI researchers' judgment that was reflected in the criteria logic.
- The clinicians may have agreed with the criteria logic, but may have committed errors, e.g., overlooked data or misapplied the criteria.
- There may be certain combinations of inputs which were not considered in the development of our classification criteria.

The rate of agreement was not a major concern at this point. There have been extensive examples of computerizing guidelines and protocols at Intermountain in which compliance is initially low. But as clinicians become more comfortable with the logic of the computer and the logic is adjusted according to clinician feedback, compliance increases. We view this effort as akin to those experiences.

Pilot

We moved forward with a pilot to learn why the clinicians were disagreeing and to see which of the hypothesized reasons for disagreement were in operation. To date, two licensed providers – a psychiatric nurse practitioner and a PhD psychologist – have participated. When they sign a Mental Health Integration note in HELP2, an alert is sent to their HELP2 message queue, indicating the complexity score the algorithm has generated. The participants acknowledge the alerts, either accept or reject the scores, and provide feedback (especially what a rejected score should have been in their judgment). This functionality is only operational for the participating clinicians.

To date, the two participants have responded to 34 scores. They agreed with the computer scores in 24 of the 34 cases (70.6%). Table 5 presents the agreement results. Table 6 presents data availability data for the ten cases in which the clinician disagreed with the computer. For each case, the table shows the computer's assessment vs. the clinician's and the number of data items (out of the maximum of 24 possible) that the computer actually found entered in HELP2 for that patient.

Discussion

Summary and Findings

In the retrospective data, the clinicians would have agreed with the classification only a little more than half the time. When they disagreed, the classification tended toward greater complexity. We purposely tended toward overstating complexity because at this stage of the algorithm's development we thought we could learn more by identifying all potential contributors to complexity, even at the cost of some "false positives." In the pilot, the clinician has agreed with the classification even more (70.6%).

Our process of defining and refining the classification criteria naturally aligned with the hypothetico-deductive process[18,19] common in medical problem solving. From experience and data, the researchers formulated a hypothesis (the original criteria), exhibiting data-driven inductive reasoning characteristic of experts.[19] By identifying different classes of data that would be helpful in formulating the criteria (S1, S2, O), our researchers were using "schemes," a characteristic of expert problem solving.[20] We then tested the hypothesis against data. Computerization assisted us here, allowing us to more easily validate all the criteria against real patient data. We identified test cases where the resulting complexity score did not match the researchers' expert judgment (i.e., they conflicted with the hypothesis)

Table 5. Clinician and computer agreement in the pilot.

		Computer			
		Mild	Moderate	High	TOTAL
Clinician	Mild	13	0	0	13
	Moderate	6	9	1	16
	High	0	3	2	5
	TOTAL	19	12	3	34

Table 6. Data availability for disagreement cases in the pilot.

Case	Computer	Clinician	Number of data items available (out of 24)
1	Moderate	High	24
2	Mild	Moderate	3
3	Mild	Moderate	3
4	Mild	Moderate	6
5	Mild	Moderate	6
6	High	Moderate	6
7	Mild	Moderate	3
8	Moderate	High	24
9	Mild	Moderate	3
10	Moderate	High	22

and adjusted the criteria to produce better results for those cases without impairing its performance with regard to the rest of the cases. We will continue to repeat the cycle.

The effect of "missing data" on criteria execution warrants further examination. Others have noted the impacts of incomplete data on inference and suggested strategies, both in the general case[21-24] and in the case of electronic health records and decision support.[25-27] In the pilot so far, as shown in Table 4, in six of the ten cases in which the clinician disagreed with the computer, the computer concluded "mild" while the clinician viewed the patient as "moderate" complexity. As shown in Table 5, in two of the six cases, clinicians had only entered six out of 24 possible elements into HELP2 while in the other four cases, they had only entered three of the 24 elements. As has been explained, the criteria conclude "mild" if not presented with any data that drive another conclusion. Consequently, as might have been predicted, the pilot showed that, in the face of sparse data, the criteria are unable to conclude anything but "mild." The clinician, on the other hand, with more data available to him or her, is able to make a more educated assessment of complexity. We need to investigate:

- Why did they enter so few elements? Is it because the ones they entered are most important to complexity? Or were they the most conveniently collected from the patient?
- What are the most salient elements that the clinician is using that the computer does not have access to, either because 1) the MHI packet is not addressing them, 2) the MHI packet is addressing them but HELP2 is not storing them, or 3) the MHI packet is addressing them, HELP2 is storing them, but the clinicians are not entering them?

We recognize that interpreting missing data as "mild" complexity may appear inconsistent with our earlier stated preference to err on the side of "over-involving" the mental health specialist. However, in our tendency to over-state complexity, we were satisfied based on our data analysis that the criteria would result in a slight over-statement, but not so much that would cause skepticism toward the result (i.e., "alert fatigue"). In contrast, we felt that making sparse data case generate "moderate" or "high" would indeed cross that threshold. A better solution may have been to not generate a result at all if some pre-determined number of items were not present, with an "insufficient data" notification. A variant of this idea would be to always conclude a complexity, but also generate an accompanying confidence measure based on the amount of data available. Still another possibility is to improve MHI data entry via reminders.

Computerization of scoring algorithms has been undertaken in Intensive Care Unit (ICU) acuity, readmission risk, community acquired pneumonia risk, and other areas.[28-32] Little if any computerization of algorithms has been performed in the mental health complexity scoring. One reason is likely the subjective, self-reported, and often sensitive nature of the data. But we venture to capture and utilize this data in our automated scoring because patient engagement is a critical component of Intermountain's MHI efforts. They in fact reflect the enterprise's Shared Accountability strategy, which aims to engage all stakeholders – the healthcare organization, employers, caregivers, the community, and patients – in providing better care and better health at sustainable cost levels. This underscores a cultural shift at Intermountain – and in the nation as a whole – in which patients and their families are being invited to become more engaged and invested in their care. In this climate, researchers in informatics and the social sciences will need to develop innovative, engaging, and nonintrusive technical and cultural solutions for interacting with patients and collecting reliable data from them.

Limitations and Future Work

A thorough, formal examination of the criteria before computerization may have isolated computerization issues from criteria issues. But our motivation for omitting this step may be found in our view of the criteria and in the environment at Intermountain. It would be preferable to validate a diagnostic algorithm (i.e., ensure the algorithm generates a sensitive and specific outcome by comparing its result to the "gold standard") before computerizing. In contrast, the classification criteria have been developed based on MHI researchers' experience as an aid to the PCP. No "gold standard" exists. Our computerization is actually a quality improvement exercise aimed at refining the criteria and, when clinicians disagree, learning more about why they do. Further, continuous quality improvement and computerized data capture and decision support are so firmly engrained in our culture at Intermountain that it was natural and convenient to expedite the implementation of the criteria in HELP2, allowing us to easily collect feedback and continuously improve clinician understanding on the one hand and the criteria on the other.

After the pilot, we intend to do more thorough analysis on both the agreement cases and disagreement cases. We will explore whether there is correlation between individual clinicians or clinician roles and agreement/disagreement and whether certain sub-scores are responsible for an inordinate amount of disagreement.

The work described is analysis of retrospective data and pilot results to date. Prospective evaluation of refined criteria on larger numbers of patients is anticipated after incorporation of pilot findings. The criteria specifically address a narrow subject – MHI complexity. The classification is not intended to be applicable to any other clinical domain. However, the Foresight architecture has been used in a variety of domains.[33-35] And we can conceive of a Foresight-based framework to more generically address scoring/classification problems like the MHI complexity problem.

We performed the retrospective analysis only on patients for whom most of the 24 data elements had been entered. This did not give us insight into those cases in which data were sparse. The pilot is our passage into exploration of how to appropriately interpret and address missing data.

We focused our criteria on the data elements that experience, best practices, and external sources suggested. Starting with the pilot, we can endeavor to identify which of the elements truly contribute most to MHI complexity and which elements are not contributing.

The capstone of the effort will be incorporating the criteria into Intermountain's production system and the clinician workflow. During the pilot, the generation of a complexity score results in an alert message to the clinician. The clinician must navigate to his or her message queue to view the generated score. The message displays the complexity score and prompts for feedback, but does not self-explain. In contrast, when implemented in production, the system will ideally generate and display the score at the point the clinician enters or reviews the data, display the sub-scoring that constituted the complexity, track clinician responses, and assist the clinician in launching workflow (for example, orders for consults) if the clinician desires. Also envisioned is direct patient-entry of the MHI data at our patient portal instead of the current transcription process.

Conclusion

We have been able to effectively computerize classification criteria to assess a primary care patient's mental health complexity. Over time, as we analyze our data, further our understanding of the nature of MHI complexity, refine the criteria to incorporate the most important data elements, and educate clinicians on our findings, we hope the scoring will yield a true representation of the patient's MHI complexity and assist PCPs in an area outside their expertise. The computerization of the MHI complexity scoring will provide more scores, more appropriate scores, and more consistent scoring, which we anticipate will further the MHI objective of integrating more appropriate and directed mental health care in the primary care environment and facilitate future research.

References

1. Gatchel RJ, Oordt MS. Clinical health psychology and primary care: Practical advice and clinical guidance for successful collaboration. Washington, DC: American Psychological Association; 2003.
2. Kessler RC, Demler O, Frank RG, et al. Prevalence and treatment of mental disorders, 1990 to 2003. N Engl J Med. 2005;352(24):2515-2523.
3. Collins C, Hewson D, Munger R, Wade T. Evolving models of behavioral health integration in primary care. New York: Millbank Memorial Fund; 2010. PDF available at: http://www.milbank.org/reports/10430EvolvingCare/EvolvingCare.pdf
4. Vermani M, Marcus M, Katzman M. Rates of detection of mood and anxiety disorders in primary care: a descriptive, cross-sectional study. Prim Care Companion CNS Disord. 2011;13(2): PCC.10m01013. doi: 10.4088/PCC.10m01013
5. James BC, Lazar JS. Health system use of clinical programs to build quality infrastructure. In: Nelson EC, Batalden PB, Lazar JS, editors. Practice-based learning and improvement: a clinical improvement action guide. Oakbrook Terrace (IL): Joint Commission Resources, 2007; p. 95–108.
6. Reiss-Brennan B, Mental health integration: normalizing team care. J Prim Care Community Health. 2014 Jan 1;5(1):55-60.
7. Reiss-Brennan B, Briot PC, Savitz LA, Cannon W, Staheli R. Cost and quality impact of Intermountain's mental health integration program. J Healthc Manag. 2010 Mar-Apr;55(2):97-113; discussion 113-4.
8. Clayton PD, Narus SP, Huff SM, et al. Building a comprehensive clinical information system from components. The approach at Intermountain Health Care. Methods Inf Med. 2003;42(1):1–7.
9. Kroenke K, Spitzer RL. The PHQ-9: A new depression and diagnostic severity measure. Psychiat Ann, 2002;32:509-521.

10. Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med. 2006 May;166 (10):1092–7. doi:10.1001/archinte.166.10.1092. PMID 16717171.
11. Hirschfeld RM, Williams JB, Spitzer RL, et al. Development and validation of a screening instrument for bipolar spectrum disorder: the Mood Disorder Questionnaire. Am J Psychiatry. 2000 Nov;157(11):1873-5.
12. Kessler RC, Adler L, Ames M, et al. The World Health Organization Adult ADHD Self-Report Scale (ASRS). Psychol Med. 2005 Nov; 35(2):245-25.
13. Morris AH. Treatment algorithms and protocolized care. Curr Opin Crit Care. 2003 Jun;9(3):236-40.
14. Morris AH. Developing and implementing computerized protocols for standardization of clinical decisions. Ann Intern Med. 2000 Mar 7;132(5):373-83.
15. McDonald CJ, Overhage JM. Guidelines you can follow and can trust. An ideal and an example. JAMA. 1994 Mar 16;271(11):872–873.
16. Tierney W M , Overhage JM, Takesue BY, et al. Computerizing guidelines to improve care and patient outcomes: the example of heart failure. J Am Med Inform Assoc. 1995;2:316-22.
17. Rocha RA, Bradshaw RL, Hulse NC, Rocha BH. The clinical knowledge management infrastructure of Intermountain Health Care. In: Greenes RA, editor. Clinical Decision Support - The Road Ahead. Boston: Academic Press. 2006; 469-502.
18. Hardin LE. Research in medical problem solving: a review. J Vet Med Educ. 2003 Fall;30(3):230-5.
19. Patel VL, Yoskowitz NA, Arocha JF, Shortliffe EH. Cognitive and learning sciences in biomedical and health instructional design: A review with lessons for biomedical informatics education. J Biomed Inform. 2009;42(1):176-97.
20. Mandin H, Jones A, Woloschuk W, Harasym P. Helping students learn to think like experts when solving clinical problems. Acad Med. 1997 Mar;72(3):173-9.
21. Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods. 2002;7:147-177.
22. Rubin DB. Inference and missing data. Biometrika. 1976;63:581-592.
23. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychol Methods. 2001;6:330-351.
24. Royston P. Multiple imputation of missing values. Stata Journal. 2004;4:227-241.
25. Lin JH, Haug PJ. Exploiting missing clinical data in Bayesian network modeling for predicting medical problems. J Biomed Inform. 2008;41:1-14.
26. Wells BJ, Nowacki AS, Chagin K, Kattan MW. Strategies for handling missing data in electronic health record derived data. eGEMs (Generating Evidence & Methods to improve patient outcomes). 2013;1(3):Article 7.
27. Berner ES, Kasiraman RK, Yu F, Ray MN, Houston TK. Data quality in the outpatient setting: impact on clinical decision support systems. AMIA Annu Symp Proc. 2005:41-5.
28. Baillie CA, VanZandbergen C, Tait G, Hanish A, Leas B, French B, Hanson CW, Behta M, Umscheid CA. The readmission risk flag: using the electronic health record to automatically identify patients at risk for 30-day readmission. J Hosp Med. 2013 Dec;8(12):689-95.
29. Deleger L, Brodzinski H, Zhai H, Li Q, Lingren T, Kirkendall ES, Alessandrini E, Solti I. Developing and evaluating an automated appendicitis risk stratification algorithm for pediatric patients in the emergency department. J Am Med Inform Assoc. 2013 Dec;20(e2):e212-20.
30. Chandra S, Agarwal D, Hanson A, Farmer JC, Pickering BW, Gajic O, Herasevich V. The use of an electronic medical record based automatic calculation tool to quantify risk of unplanned readmission to the intensive care unit: a validation study. J Crit Care. 2011 Dec;26(6):634.
31. Harrison AM, Yadav H, Pickering BW, Cartin-Ceba R, Herasevich V. Validation of computerized automatic calculation of the Sequential Organ Failure Assessment Score. Crit Care Res Pract. 2013;2013:975672.
32. Aronsky D, Haug PJ. Assessing the quality of clinical data in a computer-based record for calculating the Pneumonia Severity Index. J Am Med Inform Assoc 2000;7(1):55-65.
33. Morris AH, Orme J, Rocha BH, et al. An electronic protocol for translation of research results to clinical practice: a preliminary report. J Diabetes Sci Technol. 2008 Sep;2(5):802-8.
34. Rocha BH, Langford LH, Towner S. Computerized management of chronic anticoagulation: three years of experience. Stud Health Technol Inform. 2007;129(Pt 2):1032-6.
35. Staes CJ1, Evans RS, Rocha BH, Sorensen JB, Huff SM, Arata J, Narus SP. Computerized alerts improve outpatient laboratory monitoring of transplant patients. J Am Med Inform Assoc. 2008 May-Jun;15(3):324-32.

A Template for Authoring and Adapting Genomic Medicine Content in the eMERGE Infobutton Project

Casey L. Overby, PhD[1,2]; Luke V. Rasmussen[3]; Andrea Hartzler, PhD[4]; John J. Connolly, PhD[5]; Josh F. Peterson, MD MPH[6,7]; RoseMary E. Hedberg, MA[3]; Robert R. Freimuth, PhD[8]; Brian H. Shirts, MD PhD[9]; Joshua C. Denny, MD MS[6,7]; Eric B. Larson, MD MPH[10]; Christopher G. Chute, MD DrPH[8]; Gail P. Jarvik, MD PhD[11]; James D. Ralston, MD MPH[10]; Alan R. Shuldiner, MD[1]; Justin Starren, MD PhD[3,12]; Iftikhar J. Kullo, MD[13], Peter Tarczy-Hornoch, MD[14], Marc S. Williams, MD[15]

[1]Program for Personalized and Genomic Medicine and Department of Medicine, [2]Center for Health-related Informatics and Bioimaging, University of Maryland, Baltimore, MD; [3]Department of Preventive Medicine, [12]Medical Social Sciences, Northwestern University, Chicago, IL; [4]The Information School, [9]Laboratory Medicine, [14]Biomedical Informatics and Medical Education, [11]Department of Medical Genetics, University of Washington, Seattle, WA; [5]Center for Applied Genomics, The Children's Hospital of Philadelphia, Philadelphia, PA; [6]Department of Biomedical Informatics, [7]Department of Medicine, Vanderbilt University, Nashville, TN; [10]Group Health Research Institute, Seattle, WA; [8]Department of Health Sciences Research, [13]Division of Cardiovascular Diseases, Mayo Clinic, Rochester MN; [15]Genomic Medicine Institute, Geisinger Health System, Danville, PA

Abstract

The Electronic Medical Records and Genomics (eMERGE) Network is a national consortium that is developing methods and best practices for using the electronic health record (EHR) for genomic medicine and research. We conducted a multi-site survey of information resources to support integration of pharmacogenomics into clinical care. This work aimed to: (a) characterize the diversity of information resource implementation strategies among eMERGE institutions; (b) develop a master template containing content topics of important for genomic medicine (as identified by the DISCERN-Genetics tool); and (c) assess the coverage of content topics among information resources developed by eMERGE institutions. Given that a standard implementation does not exist and sites relied on a diversity of information resources, we identified a need for a national effort to efficiently produce sharable genomic medicine resources capable of being accessed from the EHR. We discuss future areas of work to prepare institutions to use infobuttons for distributing standardized genomic content.

Introduction

The Electronic Medical records and Genomics (eMERGE) Network[1] is a national consortium funded by the National Human Genome Research Institute (NHGRI) to develop methods and best practices for using the electronic health record (EHR) for genomic medicine and research. Methods and best practices span from using EHR data to discover genotype-phenotype associations, to incorporating genotype data into the EHR for clinical use. There are however significant challenges to using genotype data to make informed healthcare decisions. These include lack of evidence-based clinical guidelines for using genotype data and barriers to implementation (i.e., practice change).

In collaboration with the Pharmacogenomics Research Network[2], the eMERGE network is using clinical guidelines prepared by the Clinical Pharmacogenetics Implementation Consortium (CPIC)[3, 4] in patients who are preemptively genotyped and who have an increased probability of receiving target drugs including clopidogrel, warfarin and simvastatin in the next 3 years. In order to achieve impact, however, changes in healthcare practices are critical. Institutions participating in this eMERGE Pharmacogenomics (PGx) project are therefore establishing various implementation strategies that encourage healthcare providers to individualize drug therapy using CPIC guidelines.

While implementation strategies must be localized[5-7], developing sharable tools and resources increases the likelihood of a successful local implementation. To this end, the eMERGE Network is exploring use of infobuttons[8-10] as a decision support tool to provide context-specific links within the EHR to relevant genomic

medicine content about drug-gene associations. A previous study showed providing infobuttons that link to specific content topics was more effective than links that point to general overview content[11]. We therefore aimed to design an infobutton-compatible information resource[12] that is likely to impact decision-making. Our goal was to: (a) characterize the diversity of information resource implementation strategies among eMERGE-PGx project participants; (b) develop a master template containing content topics important for genomic medicine practices; and (c) assess the coverage of content topics among information resources developed by eMERGE sites.

Methods

Characterize the diversity of eMERGE-PGx information resource implementation strategies

Through early research team discussions, we found that many eMERGE-PGx project participants had already established local approaches to synthesize evidence and manage content for healthcare providers and/or patients. We therefore characterized implementation strategies among eMERGE-PGx project participants by assessing the target audience, purpose and format for their information resources. We also assessed authoring, editing and reviewing, and updating processes. In addition to efforts related to eMERGE-PGx, for comparison purposes, we evaluated resources from three national efforts: two to synthesize genomic medicine evidence (GeneReviews®[13-15] and PLoS Currents Evidence on Genomic Tests[16, 17]) and one to publish clinical guidelines (CPIC guidelines). In order to design an information resource that accounts for diverse implementation strategies, we developed a master template for authoring and adapting genomic medicine content from existing resources.

Develop a master template for eMERGE-PGx content using DISCERN-genetics criteria

The research team developed an initial list of content topics based on the quality assessment tool, DISCERN-genetics.[18, 19] The DISCERN criteria was developed for assessing information on treatment[20, 21], and is widely used for both appraisal[21-25] and to guide the production of health information on treatment for the public[26, 27]. The DISCERN methodology was more recently used to develop the DISCERN-genetics criteria to assess the quality of information on genetic screening and testing for a lay audience.[18] We chose to use this tool given the criteria were developed to facilitate applying information covering a broad range of genetic screening and testing situations to a variety of settings and conditions. The DISCERN-genetics criteria have also undergone an extensive evaluation involving testing with health information consumers, producers and providers. Moreover, the focus of DISCERN-genetics is on information content, rather than the way information is packaged, presented and made accessible. These characteristics are appropriate given the diversity of implementation strategies among eMERGE institutions.

One author (CO) developed a master template from the initial list of content topics. Initial revisions were then made to the template based upon discussions among three authors on this manuscript with expertise in clinical and biomedical informatics (CO, LR, AH). Following revisions, we piloted the template by adapting content from one information resource (MyResults.org[28]) developed by the Children's Hospital of Philadelphia. Specifically, three authors (CO, LR, JC) each adapted content for one of three drug-gene pairs (clopidogrel/*CYP2C19*, warfarin/*CYP2C9* & *VKORC1* and simvastatin/*SLC01B1*) to the master template. Pilot efforts informed our approach to provide training to eMERGE network participants interested in using the master template.

Assess the coverage of template content topics among eMERGE institutions

One author (CO) led a 20 minute training session that provided eMERGE participants with an overview of the master template, the process for adapting content, and study questions. At the conclusion of this session we identified individuals from eMERGE institutions interested in contributing to this study and each identified one or two PGx scenarios for which they would either author or adapt content from an existing information resource to the master template.

We used the Qualtrics online survey platform (http://qualtrics.com/) to collect descriptive data (i.e., institution, name of existing information resource, topic/scenario), content that was adapted to the master template and that was uploaded as an attachment, data on the adaption process, and data that was useful for configuring information resources for infobuttons (e.g.,where in the EHR should the resource be accessed). Quantitative data were downloaded and saved as an Excel file, and then read into STATA[29] for summary statistics. We report summary statistics for descriptive data and data on the adaption process. Two questions about the adaption process were asked for each content topic (i.e., "Please indicate how you were able to adapt text from the original resource?" and "Was there an equivalent section in the original resource?"). Data collected for questions relevant for configuring information resources were not within the scope of this manuscript and therefore were not analyzed.

We also included an open-ended comment field for each content topic. Qualitative narrative responses were used to provide a more thorough understanding of challenges faced when using the master template.

Results

Diversity of eMERGE PGx resource implementation strategies

We characterized seven information resources, including four developed by eMERGE institutions and three from national efforts to synthesize genomic medicine evidence. **Table 1** summarizes the target audience, purpose and format of the information resource. With the exception of MyResults.org that targets a patient audience, all of the resources we reviewed target healthcare professionals. Other target audiences among the resources we evaluated include researchers, public health professionals and decision-makers (providers, patients, managers and policy makers). The purpose of two information resources (MyDrugGenome.org and MyResults.org) was to provide education. The purpose of five information resources (AskMayoExpert, CPIC Guidelines, GeneReviews, MyDrugGenome.org, and Northwestern Clopidogrel Fact Sheet) was to allow healthcare providers to manage and treat patients that have specific genetic conditions or genetic test results. One information resource (PLoS Currents Evidence on Genomic Tests) aimed to provide a forum for sharing genomic medicine content with a wide audience.

Table 1. Scope of existing genomic medicine resources.

Information Resource (Institution)	Target audience	Purpose	Format (how information is made accessible)
AskMayoExpert* (Mayo Clinic)	Mayo Clinic healthcare providers	To provide access to expert clinical information on a wide variety of topics related to patient care including prevention, diagnosis and treatment.	Custom web-based content system; in the context of pharmacogenomics, the content is also used to develop clinical decision support logic.
CPIC Guidelines (PGRN and PharmGKB)	Clinical professionals (MD, PharmD, NP), other healthcare professionals, researchers and genetics professionals.	To provide guidance to healthcare providers about how genetic tests should be interpreted to improve drug therapy.	Published in the journal *Clinical Pharmacology and Therapeutic*s and made available at PharmGKB[30, 31]
GeneReviews® (University of Washington)	Genetics professionals (MD, PhD, MS), and other healthcare professionals.	To allow healthcare providers who are not experts in a given disorder to manage the first encounter with a patient with a given diagnosis.	Online web-based resource publically accessible on the National Center for Biotechnology Information (NCBI) Bookshelf website[32]
MyDrugGenome.org* (Vanderbilt University)	Clinical professionals (MD, NP), other healthcare professionals, Researchers and genetics professionals	To educate healthcare providers on pharmacogenetics and to allow healthcare providers to interpret the impact of specific pharmacogenetic diplotypes on target medications.	Online web resource linked to implemented clinical decision support available in order entry and e-prescribing applications
MyResults.org* (Children's Hospital of Philadelphia)	Patients in the eMERGE network, resources for professionals planned for future implementation.	To educate patients on the purpose of the eMERGE network and eMERGE projects relevant to their healthcare.	Online web-based resource
Northwestern Clopidogrel Fact Sheet* (Northwestern University)	Northwestern University clinical professional (MD, NP), other healthcare professionals, researchers and genetics professionals.	To allow healthcare providers to interpret new genomic test results.	Custom web-based content system accessible through infobuttons in the EHR
PLoS Currents Evidence on Genomic Tests	Public health and healthcare practitioners and decision-makers.	To bring together data from multiple sources (e.g., basic research, clinical trials, epidemiological and clinical studies), and provide a forum for sharing data, with the goal of making it actionable.	Published using the PLOS Currents Annotum publishing platform[33] and archived at PubMed Central[34]

*Resources developed by eMERGE institutions

Table 2 (on the next page) summarizes content management processes for these resources including authoring, editing and reviewing, and updating processes. Four different approaches to authoring materials are described among the seven resources we reviewed. For four resources (CPIC Guidelines, GeneReviews, MyDrugGenome.org[35], and MyResults.org) content is team-authored, for one resource (Northwestern Clopidogrel Fact Sheet) a single individual authors materials, and for two resources (AskMayoExpert and PLoS Currents Evidence on Genomic Tests) one or more individuals author materials. In terms of editing and reviewing, three resources are externally peer-reviewed (CPIC Guidelines, GeneReviews and PLoS Currents Evidence on Genomic Tests). In all three instances reviewers are provided detailed guidance on how to review the materials. Conversely, a team of individuals reviews content for the four information resources that are developed by eMERGE institutions (AskMayoExpert, MyDrugGenome.org, MyResults.org and Northwestern Clopidogrel Fact Sheet). Prior to

publishing, some institutions require a formal approval (e.g., as with AskMayoExpert) and some require team consensus (e.g., as with MyDrugGenome.org). Three resources we evaluated, CPIC Guidelines, GeneReviews and MyResults.org, have established a formal process to regularly update their content.

Table 2. Content management processes for genomic medicine resources

Information Resource (Institution)	Authoring	Editing and reviewing	Updating
AskMayoExpert* (Mayo Clinic)	Internally authored by one primary author. Additional authors/contributors to information are optional.	The Mayo Clinic Pharmacogenomics Task force, comprised of a multidisciplinary team of subject experts reviews and approves pharmacogenomics modules. (Internal review)	A formal review is conducted 1 year from publication, unless stated otherwise by the authors.
CPIC Guidelines (PGRN and PharmGKB)	Distributed (international); a multidisciplinary writing committee composed of scientists, pharmacologists, and healthcare providers that have expertise in the topic area; a PharmGKB scientific curator; and a biomedical informatics professional.	The writing committee and members of CPIC review the guideline manuscript internally. The manuscript then undergoes external scientific peer-review by the journal *Clinical Pharmacology and Therapeutics* prior to publication. (Internal and external review)	A formal review is conducted every two years or as needed based on important new information that would modify prescribing recommendations. All versions are archived and separately citable.
GeneReviews® (University of Washington)	Distributed (international); experts include at least one healthcare provider (target audience member).	Editors with expertise in clinical genetics, laboratory genetics, and genetic counseling review disease descriptions for accuracy. Descriptions are then peer-reviewed by internationally acknowledged subject experts. (Internal and external review)	A formal review is conducted every two or three years, depending on the topic. The editorial staff or authors initiate time-critical revisions when a clinically significant development needs to be published.
MyDrugGenome.org* (Vanderbilt University)	Internally authored by members of the Vanderbilt pharmacogenomics implementation team.	The Vanderbilt pharmacogenomics team, made up of a multidisciplinary group of subject experts, review clinical and pharmacogenomics content related to each drug-gene interaction. (Internal review)	Updates are concurrent with release of new drug-gene interactions within the PREDICT program.[36]
MyResults.org* (Children's Hospital of Philagelphia)	Led the research/outreach team at Children's Hospital of Philadelphia and supported by other eMERGE members.	A genetic counselor from Northwestern University reviews the content. (External review)	Several components are still under development. The website is updated every three weeks. There is a formal review by two individuals that occurs annually.
Northwestern Clopidogrel Fact Sheet* (Northwestern University)	Primarily developed by one genetic counselor as a synthesis of existing resources.	A physician reviews the content to ensure wording is succinct, informative and actionable. Genetic counselors also review content for accuracy and appropriate wording. (Internal review)	No formal process is established at this time.
PLoS Currents Evidence on Genomic Tests	Manuscript submission. Most authors are researchers in genetics, knowledge synthesis, or related fields.	A board or reviewers determine whether contribution is intelligible, relevant, ethical and scientifically credible. The reviewers focus on 5 criteria: Methodology, Results and interpretation, Quality of the written English, Data availability, and Ethical standards. (External review)	Revisions to contributions are possible with approval by the Reviewer Board. All versions are archived and separately citable.

*Resources developed by eMERGE institutions

Master template developed for eMERGE-PGx content

Following discussions among three research team members, 19 DISCERN-genetics question themes[19], were translated into 13 content topics that comprise the eMERGE master template (see **Box 1**). All content topics correspond to DISCERN-genetics question themes, with the following exceptions: (1) five themes about testing ("purpose of the test," "testing procedure," "test accuracy," "after the test," and "access to test results") were translated into three content topics (see **Box 1**, #5-7); (2) three themes ("discrimination," "psychosocial consequences," and "consequences for others") were collapsed into one content topic (see **Box 1**, #9); (3) two themes ("aims are clear" and "sources of information used") were replaced (see **Box 1**, #1 and #12, respectively); and (4) two themes ("aims achieved," and "balance

Box 1. Content topics captured by eMERGE template

1. Clinical scenario/Overview
2. Background and effects of the condition
3. Treatment and management choices for the condition
4. Risk of developing, carrying or passing on the condition
5. Types of tests available or being offered
6. Testing procedure
7. Test accuracy and reliability
8. Shared decision making
9. Potential risks (psychosocial consequences, implications of discrimination, potential consequences for others)
10. Local information
11. Additional sources of support and information
12. Content contributors
13. Date of the information

and bias") were dropped all together. We also used hints provided for questions in the DISCERN-genetics questionnaire in the master template to help individuals identify relevant content to adapt from another information resource (e.g., Question #1: Are the aims clear?; Hints: What is it about? What is it meant to cover [and what topics are excluded]? Who might find it useful?).

Piloting the master template with the MyResults.org resource informed the design of reference material used in our training session. This material was shared by email to all individuals who volunteered to adapt content to the master template. In addition, our pilot effort informed changes to MyResults.org format and content. The MyResults team continues to revise their template to align with the 13 content topics listed in **Box 1**, in an effort to achieve consistency across eMERGE informational projects.

Individuals from six eMERGE institutions contributed pharmacogenomics information resources. Resources and drug-gene pairs we covered are summarized in **Table 3**. All together we assessed eleven information resource/drug-gene pair combinations (e.g. AskMayoExpert/Clopidogrel[*CYP2C19*] is one information resource/drug-gene pair combination). One resource provided clinical guidelines for using pharmacogenomics data (CPIC guidelines). The other six resources summarized evidence and recommendations from systematic reviews and evidence-based clinical guidelines. Five were existing information resources prepared by eMERGE institutions (AskMayoExpert, MyDrugGenome.org, MyResults.org, Northwestern Patient Handout, Northwestern Clopidogrel Fact Sheet). One resource contained content that was authored using the master template (Geisinger-authored).

Table 3. Information resources and sample of pharmacogenomic drug-gene pairs adapted to the eMERGE template

Information Resource	**Pharmacogenomics Drug-Gene Pairs Adapted to the Master Template**				
	Carbamazepine (HLA-B)	***Clopidogrel (CYP2C19)***	***Simvastatin (SLC01B1)***	***Thiopurine (TPMT)***	***Warfarin (CYP2C9 & VKORC1)***
AskMayoExpert		X	X		
CPIC Guidelines on PharmGKB	X				X
Geisinger-authored			X		
MyDrugGenome.org		X		X	
MyResults.org			X	X	
Northwestern Patient Handout		X			
Northwestern Clopidogrel Fact Sheet		X			

NOTE: empty cells denote that they were not included in this study (not that they do not exist for a given information resource).

Coverage of template content topics among eMERGE institutions

We assessed the coverage of content topics in our template by authoring or adapting content for seven resources (eleven information resource/drug-gene pair combinations). **Table 4** summarizes for each content topic whether relevant text existed and whether a specific sub-section existed in the resource for the drug-gene pair.

All information resources contained text relevant for four content-topics ("Clinical scenario/Overview", "Background and effects of the condition", "Treatment and management choices for the condition", and "Additional sources of support and information"). All but one information resource contained text relevant for the "Risk of developing, carrying or passing on the condition" content topic. Text on "Types of tests available or being offered," "Testing procedure," "Test accuracy and reliability" existed for seven (64%), eight (73%), and eight (73%) of the information resource/drug-gene pair combinations, respectively. Text on "Content contributors" existed for eight (73%) of information resource/drug-gene pair combinations.

Less than half of the information resource/drug-gene pair combinations covered each of the remaining content topics: 45% covered "Shared decision making," 45% covered "Potential risks (psychosocial consequences, implications of discrimination, potential consequences for others)," 45% covered "Local information" (e.g., information about re-imbursement), and 45% covered "Date of the information."

Two resources had only a few sub-sections (Northwestern Patient Handout and Northwestern Clopidogrel Fact Sheet). The remaining five resources (AskMayoExpert, CPIC Guidelines on PharmGKB, Geisinger-authored, MyDrugGenome.org, MyResults.org) included a minimum of eight sub-sections that corresponded to template content topics. For MyDrugGenome.org we also assessed other local resources for content (e.g., the PREDICT project Smarter Prescriptions webpage[37], See [a] in **Table 4**).

Table 4. Resource text and sub-sections corresponding with eMERGE template content topics.

Resource (Drug-gene pair)	Content topic number (See Box 1 for a brief description)												
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
AskMayoExpert (Clopidogrel/CYP2C19)	T S	T S	T S	T S	T S	T S				T S	T S	T S	S
AskMayoExpert (Simvastatin/SLC01B1)	T S	T S	T S	T S	T S	T S				T S	T S	T S	S
CPIC Guidelines on PharmGKB (Warfarin/CYP2C9 & VKORC1)	T S	T S	T S	T	T S		T S	T	T S		T S	T S	T S
CPIC Guidelines (Carbamazepine/HLA-B)	T S	T S	T S	T	T S		T S		T S		T S	T S	T S
Geisinger-authored[b] (Simvastatin/SLC01B1)	T S	T S	T S	T S	T S	T S	T S	T S	T S	S	T S	T S	T S
MyDrugGenome.org (Clopidogrel/CYP2C19)	T S	T S	T S	T S	T S[a]	T S[a]	T S			T	T S	T S[a]	
MyDrugGenome.org (Thiopurine/TPMT)	T S	T S	T S	T S	T S[a]	T S[a]	T S				T S	T S[a]	
MyResults.org – (Simvastatin/SLC01B1)	T S	T S	T S	T S		T S	T S	T S	T	T	T S		
MyResults.org (Thiopurine/TPMT)[c]	T S	T S	T S	T S		T S	S	S			T S		
Northwestern Patient Handout (Clopidogrel/CYP2C19)	T	T	T				T S	T			T S		T
Northwestern Clopidogrel Fact Sheet (Clopidogrel/CYP2C19)	T	T	T	T S	T	T	T S	T	T	T	T S	T	T

T=text relevant for the content topic exists; S: sub-section relevant for the content topic exists; [a]content exists within another related webpage; [b]authored materials using the master template; [c]content had not yet been published to MyResults.org at the time of this analysis.

Discussion

This work characterized the diversity of information resource implementation strategies among eMERGE-PGx project sites; developed a master template containing content topics important for genomic medicine practices (as identified by the DISCERN-Genetics tool); and assessed the coverage of content topics by authoring and adapting content to our master template at six eMERGE institutions.

Implementation strategies were diverse, despite our common start and end points. While the target audience and purpose for information resources we reviewed were comparable across sites, we found that authoring, reviewing and editing processes were different. In addition, with the exception of CPIC Guidelines, GeneReviews and MyResults.org, updating processes were largely informal and in one case had not yet been established. This finding is consistent with the finding that despite adopting a common sequencing platform among Clinical Sequencing Exploratory Research (CSER) institutions[38], there were a range of approaches to annotate and prioritize genomic variants and generating whole-genome clinical reports for integration into the EHR.[6]

The master template contains 13 content topics that are important for genomic medicine practices. We successfully demonstrated our ability to adapt content to this template for a range of pharmacogenomics scenarios and information resources in this study. This task, however, did not come without challenges. Given we kept the language general for a broad range of genomic medicine scenarios, we found that the hints and content sections did not always appear directly relevant for pharmacogenomics scenarios (e.g., "genetic condition" implies a "*disease-causing* genetic condition"). This finding is not surprising given the DISCERN-genetics questionnaire was evaluated previously with a sample of information on cystic fibrosis, Down's syndrome, familial breast cancer, familial colon cancer, hemochromatosis, Huntington's disease, sickle cell disease and thalassemia.[18] To these authors knowledge, however, the DISCERN-genetics criteria is the only criteria that has been assessed with such a broad range of genetic conditions that represent different populations, disease pathways and treatment decisions, and that has been evaluated with a diverse audience including health information consumers, producers and providers.

When assessing the coverage of content topics among information resources developed by eMERGE institutions we found that many existing resources have sections or sub-sections that map directly to the content topics we defined in our template. These findings reinforce the appropriateness of our decision to use the DISCERN-genetics criteria

as the basis for our master template, given the criteria focus on information content, rather than the way information is packaged, presented and made accessible. For example, MyResults.org and AskMayoExpert both have structured templates for content relevant for drug-gene pairs of interest to eMERGE. The MyResults.org template is organized into tabs that allow patients to explore answers to common questions and resources (e.g., the video titled "Does Everyone Respond to Statins in the Same Way?"[39] under the "Media and Recommended" tab). Alternatively, the AskMayoExpert template provides content on a single document organized into major sections including "FAQs", "Key Facts", "Guidelines", "Publications and Resources", and "Patient Education". Even though the templates are quite different from each other, they both include sections or sub-sections that align with the content topics listed in **Box 1** (see **Table 4** e.g., "Pub/Resources" and "Patient Education" sections of AskMayoExpert and the "Media and Recommended" tab of MyResults.org map to the "Additional sources of support and information" eMERGE template content topic).

We also identified important aspects of genetic testing that are not addressed well by existing information resources in their current form. For example, "Shared decision making," and "Potential risks (psychosocial consequences, implications of discrimination, potential consequences for others)," were covered by fewer than half of the information resource/drug-gene pair combinations we reviewed. In addition, we found that few resources captured content on "Date of the information," which directly impacts our ability to define formal processes for updating materials. We also identified aspects of genetic testing that are covered very well by the resources we analyzed. For example, content relevant for "Clinical scenario/Overview", "Background and effects of the condition", "Treatment and management choices for the condition", and "Additional sources of support and information" template content topics were available in all of the resources we evaluated. These findings help guide areas of focus for developers of genomic medicine content. Toward improving the coverage of important content areas and achieving consistency across eMERGE informational projects, the MyResults team has already made revisions to their template to align with the 13 content topics included in the master template. At the time of our analysis, MyResults targeted a patient audience, and is now being considered for expansion to a healthcare provider audience.

Conclusions and Future Directions

Overall, we gathered genomic medicine information resources in preparation for a network-wide infobutton implementation across eMERGE institutions. Given that a standard implementation does not exist and sites relied on a diversity of information resources, there is a need for methods and approaches to help institutions efficiently produce sharable genomic medicine resources. Our findings highlight two potential areas for developing tools as part of the eMERGE infobutton project: (a) collaborative authoring and (b) adapting existing information resources. Exploring both areas, in turn, there are several considerations and opportunities to collaborate with existing projects.

First, our finding that teams of individuals often author and review information resource content motivates future work to design a collaborative writing application. We also suspect that the primary literature, clinical guidelines, and systematic reviews that authors draw from to prepare materials are similar among eMERGE institutions. It may therefore be useful to provide access to these resources in a common authoring environment. We hope to draw from the experiences of others developing software (e.g., the Librarian Infobutton Tailoring Environment, LITE[9]) to provide content for the Infobutton Manager knowledge base that creates a set of context-specific links to information resources. As an important note for designing collaborative writing applications, we will also need to address open questions about the safety, reliability, divergence from traditional authoring approaches and legal implications for decision-making.[40]

Second, our finding that many resources already have sub-sections that map to our master template motivates the need for a tool to assist with adapting existing resources. Given the challenges we encountered in interpreting content-topics for pharmacogenomics scenarios, we will also need to incorporate flexibility to tailor language to the genomic medicine scenario of focus (e.g., the option to substitute "genetic condition" with "genomic predisposition" for pharmacogenomics scenarios). We expect to encounter organizational and technical challenges in our pursuits. From an organization point of view there may be many barriers that exist such as a lack of rewards to motivate sharing, little communication about the benefits and values of sharing knowledge, and a shortage of appropriate infrastructure to support sharing practices. Collaborative projects such as eMERGE provide opportunities to identify and develop models and approaches to overcome such challenges to sharing genomic medicine knowledge. From a technical standpoint, free-text is not ideal for implementing clinical decision support. We are therefore collaborating with the OpenInfobutton development team that has released the Infobutton Responder[41] to enable healthcare organizations to index their local clinical content so that it is compliant with the HL7 infobutton standard[12, 42]. Such tools could help institutions to efficiently produce sharable genomic medicine resources in a form that can be

leveraged by infobuttons in the EHR that link to specific content topics. Exploration of open source content management systems that might be customized for these purposes is underway within the eMERGE Network.

Acknowledgments

We are grateful for valuable input from Dr. Guilherme Del Fiol (University of Utah), Dr. James Cimino (NIH Clinical Center), Dr. Roberta A. Pagon (University of Washington), and Dr. Mary Relling (St. Jude Children's Research Research Hospital). The authors would also like to thank Ms. Sarah Stallings (Vanderbilt University), Vivian Pan (Northwestern University) and the PREDICT Internal Development Team (Vanderbilt University) for their contributions.

This research was supported in part by the Columbia Training in Biomedical Informatics (NIH/NLM #T15 LM007079), the University of Maryland Program for Personalized and Genomic Medicine, NIH/NIGMS (U19GM61388; the Pharmacogenomics Research Network), NIH/NHLBI PAPI (U01HL105198), UW NEXT grant (U01HG006507), UW CSER (U01HG006507), UW CTSA (UL1TR000423), and the eMERGE network [U01HG006828 (Cincinnati Children's Hospital Medical Center/Harvard); U01HG006830 (Children's Hospital of Philadelphia); U01HG006389 (Essentia Institute of Rural Health); U01HG006382 (Geisinger Clinic and University of Maryland, Baltimore); U01HG006375 (Group Health Cooperative and the University of Washington); U01HG006379 (Mayo Clinic).); U01HG006380 (Mount Sinai School of Medicine); U01HG006388 (Northwestern University); U01HG006378 (Vanderbilt University); and U01HG006385 (Vanderbilt University serving as the Coordinating Center]. Information resources we cover are supported in part by Mayo Clinic (AskMayoExpert), Vanderbilt University Medical Center (MyDrugGenome.org), Children's Hospital of Philadelphia (MyResults.org), and Northwestern University (Northwestern Patient Handout and Northwestern Clopidogrel Fact Sheet).

References

1. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, Sanderson SC, Kannry J, Zinberg R, Basford MA *et al*: **The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future**. *Genet Med* 2013, **15**(10):761-771.
2. Shuldiner AR, Relling MV, Peterson JF, Hicks JK, Freimuth RR, Sadee W, Pereira NL, Roden DM, Johnson JA, Klein TE *et al*: **The Pharmacogenomics Research Network Translational Pharmacogenetics Program: overcoming challenges of real-world implementation**. *Clin Pharmacol Ther* 2013, **94**(2):207-210.
3. Relling MV, Klein TE: **CPIC: Clinical Pharmacogenetics Implementation Consortium of the Pharmacogenomics Research Network**. *Clin Pharmacol Ther* 2011, **89**(3):464-467.
4. Caudle KE, Klein TE, Hoffman JM, Muller DJ, Whirl-Carrillo M, Gong L, McDonagh EM, Sangkuhl K, Thorn CF, Agundez JA *et al*: **Incorporation of Pharmacogenomics into Routine Clinical Practice: the Clinical Pharmacogenetics Implementation Consortium (CPIC) Guideline Development Process**. *Curr Drug Metab* 2014.
5. Overby CL, Kohane I, Kannry JL, Williams MS, Starren J, Bottinger E, Gottesman O, Denny JC, Weng C, Tarczy-Hornoch P *et al*: **Opportunities for genomic clinical decision support interventions**. *Genet Med* 2013, **15**(10):817-823.
6. Tarczy-Hornoch P, Amendola L, Aronson SJ, Garraway L, Gray S, Grundmeier RW, Hindorff LA, Jarvik G, Karavite D, Lebo M *et al*: **A survey of informatics approaches to whole-exome and whole-genome clinical reporting in the electronic health record**. *Genet Med* 2013, **15**(10):824-832.
7. Peterson JF, Bowton E, Field JR, Beller M, Mitchell J, Schildcrout J, Gregg W, Johnson K, Jirjis JN, Roden DM *et al*: **Electronic health record design and implementation for pharmacogenomics: a local perspective**. *Genet Med* 2013, **15**(10):833-841.
8. Cimino JJ, Elhanan G, Zeng Q: **Supporting infobuttons with terminological knowledge**. *Proc AMIA Annu Fall Symp* 1997:528-532.
9. Cimino JJ, Jing X, Del Fiol G: **Meeting the electronic health record "meaningful use" criterion for the HL7 infobutton standard using OpenInfobutton and the Librarian Infobutton Tailoring Environment (LITE)**. *AMIA Annu Symp Proc* 2012, **2012**:112-120.
10. Del Fiol G, Curtis C, Cimino JJ, Iskander A, Kalluri AS, Jing X, Hulse NC, Long J, Overby CL, Schardt C *et al*: **Disseminating context-specific access to online knowledge resources within electronic health record systems**. *Stud Health Technol Inform* 2013, **192**:672-676.

11. Del Fiol G, Haug PJ, Cimino JJ, Narus SP, Norlin C, Mitchell JA: **Effectiveness of topic-specific infobuttons: a randomized controlled trial**. *J Am Med Inform Assoc* 2008, **15**(6):752-759.
12. **Context-aware knowledge retrieval (infobutton) product brief. HL7 International Wiki Site.** [http://wiki.hl7.org/index.php?title=Product_Infobutton]
13. **GeneReviews™** [http://www.ncbi.nlm.nih.gov/books/NBK1116/]
14. Pagon RA: **GeneTests: an online genetic information resource for health care providers**. *J Med Libr Assoc* 2006, **94**(3):343-348.
15. Pagon RA, Tarczy-Hornoch P, Baskin PK, Edwards JE, Covington ML, Espeseth M, Beahler C, Bird TD, Popovich B, Nesbitt C *et al*: **GeneTests-GeneClinics: genetic testing information for a growing audience**. *Hum Mutat* 2002, **19**(5):501-509.
16. Gwinn M, Dotson WD, Khoury MJ: **PLoS currents: evidence on genomic tests - At the crossroads of translation**. *PLoS Curr* 2010, **2**.
17. **PLOS Currents Evidence on Genomic Tests** [http://currents.plos.org/genomictests]
18. Shepperd S, Farndon P, Grainge V, Oliver S, Parker M, Perera R, Bedford H, Elliman D, Kent A, Rose P: **DISCERN-Genetics: quality criteria for information on genetic testing**. *Eur J Hum Genet* 2006, **14**(11):1179-1188.
19. **DISCERN - The DISCERN Genetics Instrument** [http://www.discern-genetics.org/quality_criteria_and_references.php]
20. Charnock D, Shepperd S, Needham G, Gann R: **DISCERN: an instrument for judging the quality of written consumer health information on treatment choices**. *J Epidemiol Community Health* 1999, **53**(2):105-111.
21. Griffiths KM, Christensen H: **The quality and accessibility of Australian depression sites on the World Wide Web**. *Med J Aust* 2002, **176 Suppl**:S97-S104.
22. Godolphin W, Towle A, McKendry R: **Evaluation of the quality of patient information to support informed shared decision-making**. *Health Expect* 2001, **4**(4):235-242.
23. Gimenez-Perez G, Caixas A, Gimenez-Palop O, Gonzalez-Clemente JM, Mauricio D: **Dissemination of 'patient-oriented evidence that matters' on the Internet: the case of Type 2 diabetes treatment**. *Diabet Med* 2005, **22**(6):688-692.
24. Molassiotis A, Xu M: **Quality and safety issues of web-based information about herbal medicines in the treatment of cancer**. *Complement Ther Med* 2004, **12**(4):217-227.
25. Jefford M, Tattersall MH: **Informing and involving cancer patients in their own care**. *Lancet Oncol* 2002, **3**(10):629-637.
26. Jefford M, Gibbs A, Reading D: **Development and evaluation of an information booklet/decision-making guide for patients with colorectal cancer considering therapy in addition to surgery**. *Eur J Cancer Care (Engl)* 2005, **14**(1):16-27.
27. Sanger S, Nickel J, Huth A, Ollenschlager G: **[Well-informed on health matters--how well? The German 'Clearinghouse for Patient Information'--objective, background and methods]**. *Gesundheitswesen* 2002, **64**(7):391-397.
28. **MyResults.Org - EMERGE** [myresults.org]
29. StataCorp: **Stata Statistical Software: Release 13**. In. College Station, TX: StataCorp LP; 2013.
30. Thorn CF, Klein TE, Altman RB: **Pharmacogenomics and bioinformatics: PharmGKB**. *Pharmacogenomics* 2010, **11**(4):501-505.
31. Thorn CF, Klein TE, Altman RB: **PharmGKB: the Pharmacogenomics Knowledge Base**. *Methods Mol Biol* 2013, **1015**:311-320.
32. Hoeppner MA: **NCBI Bookshelf: books and documents in life sciences and health care**. *Nucleic Acids Res* 2013, **41**(Database issue):D1251-1260.
33. **Annotum hompage** [annotum.org]
34. Roberts RJ: **PubMed Central: The GenBank of the published literature**. *Proc Natl Acad Sci U S A* 2001, **98**(2):381-382.
35. **MyDrugGenome - PREDICT DGI** [mydruggenome.org/dgi.php]
36. Pulley JM, Denny JC, Peterson JF, Bernard GR, Vnencak-Jones CL, Ramirez AH, Delaney JT, Bowton E, Brothers K, Johnson K *et al*: **Operational implementation of prospective genotyping for personalized medicine: the design of the Vanderbilt PREDICT project**. *Clin Pharmacol Ther* 2012, **92**(1):87-95.
37. **Smarter Prescriptions** [http://www.vanderbilthealth.com/main/41143 - how]

38. **Clinical Sequencing Exploratory Research (U01)** [http://grants.nih.gov/grants/guide/rfa-files/RFA-HG-10-017.html]
39. **Dr. Iftikhar Kullo (Mayo Clinic): Different Responses to Statins** [http://www.youtube.com/watch?v=NM9zS3bgVJE&feature=player_embedded]
40. Archambault PM, van de Belt TH, Grajales FJ, 3rd, Faber MJ, Kuziemsky CE, Gagnon S, Bilodeau A, Rioux S, Nelen WL, Gagnon MP *et al*: **Wikis and collaborative writing applications in health care: a scoping review**. *J Med Internet Res* 2013, **15**(10):e210.
41. **Phase II - Infobutton Responder - OpenInfobutton** [http://www.openinfobutton.org/project-overview---phase-ii-infobutton-responder]
42. Del Fiol G, Huser V, Strasberg HR, Maviglia SM, Curtis C, Cimino JJ: **Implementations of the HL7 Context-Aware Knowledge Retrieval ("Infobutton") Standard: challenges, strengths, limitations, and uptake**. *J Biomed Inform* 2012, **45**(4):726-735.

Examining the Multi-level Fit between Work and Technology in a Secure Messaging Implementation

Mustafa Ozkaynak, PhD[1], Sharon Johnson, PhD[2], Stephanie Shimada, PhD[3], Beth Ann Petrakis, MPA[3], Bengisu Tulu, PhD[2], Cliona Archambeault, MS[4], Gemmae Fix, PhD[3], Erin Schwartz, PhD[5], Susan Woods, MD[5]
[1]University of Colorado, Aurora, CO; [2]Worcester Polytechnic Institute, Worcester, MA; [3]Center for Healthcare Organization and Implementation Research (CHOIR), Bedford, MA; [4]New England Veterans Engineering Resource Center, Boston, MA; [5]Portland VA Medical Center, Portland, OR

Abstract

Secure messaging (SM) allows patients to communicate with their providers for non-urgent health issues. Like other health information technologies, the design and implementation of SM should account for workflow to avoid suboptimal outcomes. SM may present unique workflow challenges because patients add a layer of complexity, as they are also direct users of the system. This study explores SM implementation at two Veterans Health Administration facilities. We interviewed twenty-nine members of eight primary care teams using semi-structured interviews. Questions addressed staff opinions about the integration of SM with daily practice, and team members' attitudes and experiences with SM. We describe the clinical workflow for SM, examining complexity and variability. We identified eight workflow issues directly related to efficiency and patient satisfaction, based on an exploration of the technology fit with multilevel factors. These findings inform organizational interventions that will accommodate SM implementation and lead to more patient-centered care.

Introduction

Secure messaging (SM) is increasingly available as a component of patient portals and allows patients to communicate online with their providers about non-urgent issues through a secure website. SM allows healthcare organizations to provide personal service in an effective and efficient manner, and provides patients with an additional communication channel that is complementary to existing channels (i.e. phone calls and walk-in clinics). The use of SM is positively associated with health outcomes[1–3], patient satisfaction[2,4], perceived improved patient knowledge and self-care[5], adherence[6], efficiency[7] and cost of care[1,8]. However, there are also various challenges regarding SM. For example, SM adoption rates among patients are still low[9].

Integrating SM into existing practice is of widespread interest, given its inclusion in meaningful use criteria (http://www.healthit.gov/providers-professionals/achieve-meaningful-use/core-measures-2/use-secure-electronic-messaging, Last accessed on 08/03/2014), yet it poses significant challenges for clinicians[10,11]. Heyworth et al[10] reported that providers' inability to access the SM system easily was an important challenge possibly leading to less enthusiastic system adoption. Wakefield et al.[11] developed a seven-step SM implementation framework. A critical step of the framework was integration with workflow. Organizations that implement SM need to determine who will be involved in SM, who will be responsible for responding to patient messages, and what policies should be in place regarding patient access rights and staff responsibilities. These decisions, specific to organizational context, can lead to either better or worse integration of SM in workflow[12,13]. These decisions can be carefully planned, but they can also result in unintended consequences if there is a lack of fit between the technology and the work being done[11].

The importance of the fit between information technologies and workflow has been discussed in the context of a variety of health information technologies such as electronic health records (EHR)[14], barcoding[15] and computerized provider order entry (CPOE)[16]. The boundaries of these technologies are limited to clinical settings and the active users are clinicians and staff. However, workflow considerations for SM are more complex because the system boundaries extend into daily living settings, with patients being active users as well. The design of SM systems should consider these broader boundaries and wider range of users. Clinical workflow for SM (i.e., the component of overall workflow occurring in the clinic) should be sensitive to the workflow in the daily living environment.

Conceptual models that address the design and integration of technology for a work system stress the need to assess the fit between the new technology and workflow at multiple levels. The Technology Acceptance Model (TAM) and its variations mainly focus on individual factors such as perceived usefulness and ease of use[17]. Organizational

theories such as sociotechnical systems theory highlight the importance of the joint optimization of social (e.g. policies) and technical systems for effective functioning of organizations[18]. The broader cultural, political, and economic environment in which organizations operate also affect technology implementation[19]. SM implementation may also be affected by factors at each of these levels. However, successful implementation is challenged by requiring attention to all factors simultaneously[20].

The purpose of this study is to describe clinical workflows associated with the use of SM, and to identify issues (at multiple levels such as individual, task, organizational and environment) that affect clinical efficiency and effectiveness in primary care settings. Although several studies have demonstrated the positive contribution of SM, workflow challenges may limit the full benefit from SM from being achieved, particularly if adoption rates and use increase. This study examined the research question "What challenges arise when integrating SM use with workflows at different organizational levels?". Based on our data, many challenges were related to technology fit, and we then explored these challenges using a multilevel framework[21] to inform system design and organizational interventions that can lead to more effective SM use and more patient centered care.

Methods

The Veterans Health Administration (VHA) began broad enterprise implementation of SM in 2011, after a pilot implementation at several sites. Clinical staff members used an integrated EHR called the Computerized Patient Record System (CPRS). The SM application is tethered to, but not fully integrated into CPRS. Patients who were in-person authenticated for the VHA's patient portal and personal health record (My Health***e***Vet) were able to opt-in to use SM. In the VHA, patients can use SM to contact their primary care providers (and more recently specialists) about non-urgent matters by logging into their My Health***e***Vet accounts. They are told that a response from the clinical team will occur within three business days. Teams at two study locations, four teams each in New England and the Pacific Northwest, initiated SM use in 2008 as part of a pilot implementation conducted prior to national rollout. While not all providers at these locations participated in SM until the full rollout in 2011, because of their participation as pilot sites, the early adopters had more experience with SM than those at other facilities. This experience allowed us to evaluate how they had incorporated SM into their workflow, defined in this study from an operations perspective, focused on processes and allocated resources. Specifically we defined workflow as a sequence of activities by multiple individuals related to SM utilization[22]. We conducted twenty-nine individual interviews with the members of eight primary care teams. The eight teams were selected based on variation in the volume of secure messages received during the six months prior to selection for the study and the proportion of messages completed by the provider (physician or nurse practitioner) versus other team members (Table 1).

Data was collected by MO and BAP from the New England teams in March-June 2013 and by MO, SS, BAP, ES and SW from the Northwest teams in April 2013. Interviews were guided by a semi-structured interview protocol. Interview questions addressed general information about each interviewee's role on the healthcare team, how SM was used, the integration of SM with daily practice, and team members' attitudes towards and experiences with SM. Interviews took place in private rooms at each interviewee's clinic. Each interview lasted approximately 30-45 minutes. All interviews were audiotaped, except for one, and transcribed verbatim. All data collection methods were approved by the VHA Central Institutional Review Board.

Data were qualitatively analyzed in two steps. In the first step, all interviews for a team were read by two of the authors, and each author separately created a summary describing the team's use of SM. The summary was completed using a semi-structured template, focused around the elements of workflow. The sections of the template included who was interviewed, their roles and responsibilities with respect to secure messages, their impressions of the types of messages received, their description of major process steps in handling secure messages, perceptions of the value of SM, comments about organizational decisions and factors that affected SM workflow, and any technology-related issues that arose. A third author synthesized the other two researchers' notes to create an overall site summary. In the second step, the authors worked from the site summaries to develop a general workflow diagram. The researchers who analyzed data have backgrounds in medical informatics, industrial engineering, management information systems, health administration, qualitative research and primary care medicine.

Table 1. Description of the study sample

Team No	Location	Incoming Message Volume in 6 Months	Provider Completion Rate*	Roles of the Interviewees
1	New England	303	0%	Physician, licensed practical nurse (LPN), registered nurse (RN), Pharmacist**
2	New England	197	0%	Physician, LPN, Pharmacist**
3	New England	283	20%	Nurse Practitioner, LPN, RN
4	New England	329	58%	Physician, LPN, RN, Pharmacist
5	Northwest	385	0%	Physician, LPN, RN, Pharmacist
6	Northwest	491	0%	Physician, RN, Medical assistant
7	Northwest	503	73%	Physician, LPN, RN, Social worker
8	Northwest	539	32%	Physician, LPN, RN, Medical Assistant

* Provider completion rate is the percentage of the secure messages completed by provider (physician or nurse practitioner) by clicking the "***Complete***" button in the SM system.

** The same pharmacist serves both Teams 1 and 2.

Results

Secure Messaging Workflow

The clinical workflow for SM as described by interviewees and synthesized across the eight teams is illustrated in Figure 1. Figure 1 shows the main activities, the order of these activities and information flow at a high level once a team receives a message from a patient. Messages are usually focused on one request and can be addressed by one staff member. However, some messages included multiple requests and were handled by multiple people.

Once a message is opened, the team member reviews the content of the message and identifies the staff member who can best respond to the message based on their scope of practice. The team member who first opened the message often collects additional information (e.g. reviewing patient data in the CPRS system or contacting the patient for further details) and takes necessary action (e.g., request an appointment or refill a prescription). Alternatively, the message might be triaged or assigned through the SM system to another team member. In some cases, the issue was handled outside of SM by notifying another staff in person or via an electronic alert in CPRS.

If a patient requested more than one item for action (e.g. scheduling a new appointment and also providing information about physical therapy resources), then the message was first be handled by one of the team members appropriate to the request (e.g. scheduling clerk) and then reassigned to another team member (e.g. physician).

Once necessary actions were taken, a response was sent to the patient, typically either via SM or by phone. Based on the team member's judgment, the secure message was copied into CPRS as a clinical note and signed by the team member. All secure messages were required to be completed. This was accomplished by clicking the "***Complete***" button in the SM system. Replying to a message also automatically generated the option to simultaneously complete the message. Completing a message was designed to signal that the issue raised by the patient was resolved, and the message no longer appeared in the team member's Inbox. VHA policy requires completing messages within three business days; otherwise, the message is flagged as "***Escalated***".

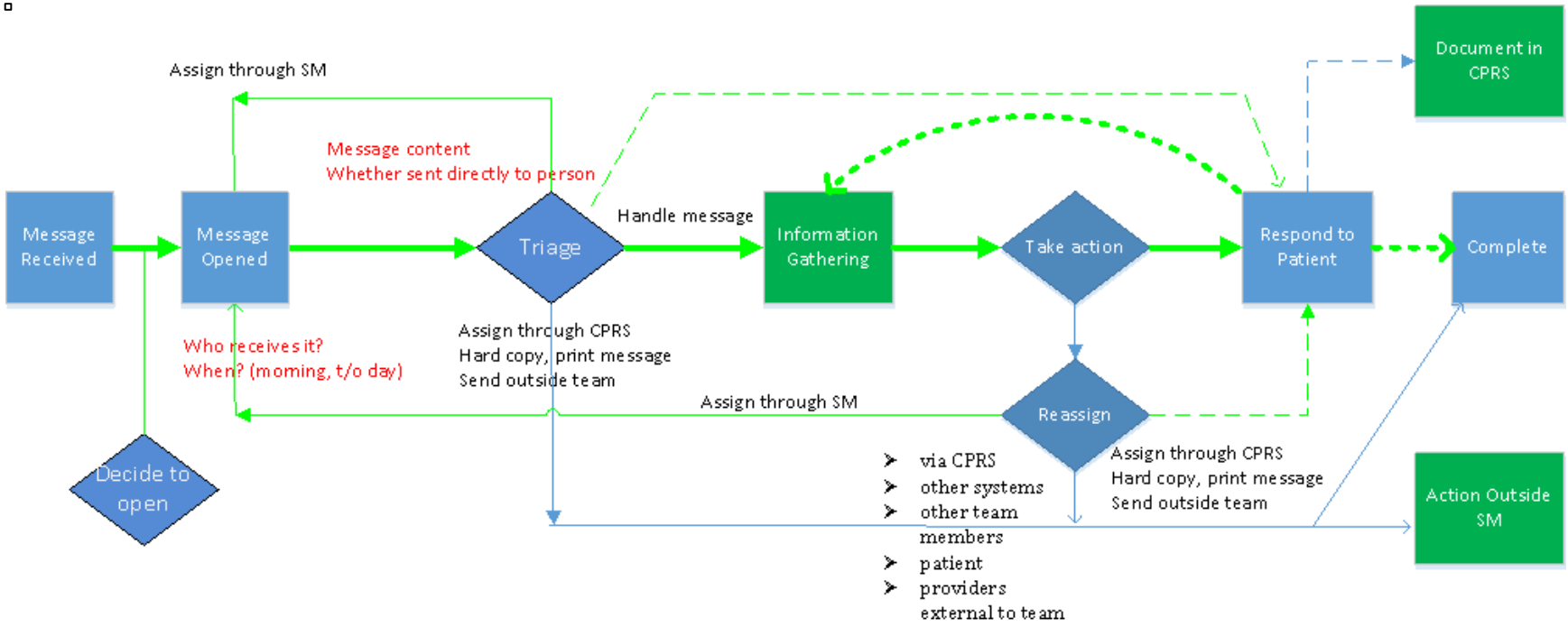


Figure 1. Overview of SM workflow across teams

The clinical workflow for SM can be complex because it includes various decision points, requires the cooperation of clinical team members, and exists in tandem with other workflows focused on responding to patients (e.g., telephone calls, walk-ins). On occasion, it was necessary to reach the patient for further information and clarification in order to proceed. If the patient's question was multifaceted, interviewees noted that it was often more expeditious to follow up with patients directly by phone, although patients were not always available to ensure timely contact. While the SM platform was the same across sites, clinics varied in terms of size, locale, patient demographics, physical layout, and organizational structure. In addition, each clinic and the teams within these clinics made local organizational decisions about how to embed SM into their existing workflows. These differences and decisions led to variability in the clinical workflow for SM across the care delivery teams. Our analysis revealed two major workflow variations across teams. Variations occurred in how team members used the SM system, and how teams communicated with other staff members and patients outside of the SM system.

Differences were found both in how messages were received and how team members responded to messages. Depending on the team and situation, there was significant variation in which team member was responsible for opening messages. For example, on one team an LPN was the first to open all incoming messages. On another, registered nurses and LPNs shared the responsibility of opening messages. In yet other settings, physicians opened messages first. In four of the teams, the providers did not personally use the SM system, relying on team members to relay the information. (Note that these team members were intentionally selected to be interviewed based on their low provider completion rate.) In other teams, physicians acted on and directly responded to patient messages using the SM system

In part because of differences in how roles and responsibilities were defined with respect to SM, message resolution was handled using a variety of means of communication. For example, providers who did not use the SM application accessed information through in-person consultations with staff, via printed copies of messages, or through CPRS notifications. Other team members called patients to respond to the issues raised in a secure message, rather than replying within the application. Although the use of multiple channels provides convenience and facilitates communication, it can also affect the team's ability to track message handling via the SM application. For example, if the staff member who opened the message forwards the message verbally and then closes the message, reviewing a secure message thread within the system may not reveal when or whether the issue was actually resolved.

Workflow Issues

Most interviewees believed the SM system was of value to patients, providing an alternate access channel and a potentially richer way to communicate. However, they were more mixed in their assessment of how well the system worked. We examined several workflow issues raised by interviewees and conceptualized the source of these issues as a lack of fit between the technology and multiple levels of work. Karsh et al.[21] discussed technology-work misfit at four interrelated levels: user, task, organization and environmental levels. In Table 2, we describe technology-work fit at different levels from Karsh et al.'s model, and identify specific workflow issues from our findings.

User-technology fit: At the user level, we observed two workflow issues related to technology fit. First, some nurses and providers preferred not to use the SM system while others were enthusiastic users. Some team members were not comfortable with computers, while others found the system burdensome (i.e. using SM required them to log in to another system and to keep several applications open simultaneously). Enthusiastic users often took on a greater role within the SM workflow (e.g. opening all messages), while others used alternative communication channels, such as telephones to respond to patients. Second, several types of messages described by interviewees suggested patients may use the system inappropriately, thereby impacting workflow. We identified two broad areas of inappropriate use by patients: 1. sending a message for an urgent issue such as chest pain, and 2. sending long messages that also included personal information not directly relevant to their health condition and request.

Task-technology fit: At the VHA, SM is not integrated with the primary existing clinical information system, CPRS. Use of the SM system therefore requires opening an additional window in the computer screen and an additional login. Moreover, because SM is not integrated into CPRS, clinicians juggle between windows to locate pertinent information such as test results to copy and paste to the SM screen. This not only is additional work, but also presents potential safety and privacy concerns[23,24]. Several interviewees also raised the issue of the inconvenience of not being able to send information (e.g. lab results) from CPRS. Other aspects of the technology also interfere with workflow tasks. For example, for privacy and security reasons, the SM software automatically times out when not in use.

Table 2. Levels of technology fit, description of levels and findings related to workflow issues at different work levels

Level	Description of the Level from Karsh et al.	Workflow Issues Identified
User-technology fit	"Fit between technology and user characteristics (e.g., values, attitudes, abilities)"	1. Use among team members varies due to their abilities, attitudes and values 2. Inappropriate use of messaging by patients
Task-technology fit	"Fit between technology and health care task characteristics (e.g., complexity, time constraints) "	3. SM was tethered, but not integrated, into the electronic record 4. Technology-related issues (e.g., frequent log offs, time required to log onto second system)
Organization-technology fit	"Fit between technology and organizational characteristics (e.g., policies, practices, social climate, resources)"	5. Need for additional policies (e.g. access by family members, identification of surrogates) 6. Additional workload 7. Despite the significant impact on workload, there was no workload credit for SM.
Environment-technology fit	"Fit between technology and the external (e.g. politics, culture) or internal (e.g., lighting, layout, noise) environment"	8. Patient expectations of early response

Organization-technology fit: Three workflow issues were identified. First, because SM is a new communication channel between patients and providers, new situations arise that require additional policies. For example, SM accounts might be used by family members (i.e., informal caregivers) on behalf of patients. This situation has not been officially addressed by organizational policies, and staff members used their judgment in how to handle confidentiality when they communicated about with family members about patients. Another example was ensuring timely response to SM when a provider was off duty. Although CPRS enables identification of surrogates for CPRS alerts, a duplicate, separate arrangement must be made in the SM system. Second, some organizational practices create additional workload. Some sites had decided that every secure message should be copied by the team user into CPRS – a task requiring manual intervention. Finally, the use of SM was not formally counted as a part of workload credit. Several interviewees saw this as a significant barrier to use, particularly as the volume of messages has grown.

Environment-technology fit: At the level of the environment, we identified one workflow issue. VHA expects clinical teams to respond and complete patient messages within three business days. On the other hand, patients may expect a same-day or faster response. The expectations of patients for quick responses has also been documented in previous studies[25]. Patients who do not get a response on the same day may be upset, and they may send additional messages or follow up by telephone to repeat their requests and reflect their frustration.

Discussion

We studied the implementation of SM in eight primary care delivery teams within the VHA. The complexity of the clinical workflow for SM, shown in Figure 1, makes it challenging to develop a SM system that fits with existing work at all levels, across many sites. Across sites, variation occurred at the user, task, organization, and environmental levels, both before and in response to implementation. Based on our data, we examined the issue of fit between the technology and four levels of work, and categorized workflow issues described by interviewees in terms of these levels. A limitation of our work is that it is based on a single organization, and a study sample of eight teams, which may limit the range of workflow challenges.

The value of using a multi-level framework to examine technology fit is that it underscores the need for strategies to mitigate issues in both the technical and organizational spheres, and at different organizational levels. Design and implementation strategies for health information technologies should recognize these levels, and the relationships between them.

Implementation strategies should seek to determine where lack of fit between technology and workflow is occurring, then develop responses to improve the fit. Responses can include both changes in technology design as well as organizational practices. In terms of technology design, in the context of this study, several refinements to the SM system would improve task-technology fit, eliminating what interviewees' perceived as difficulties or time-wasting workarounds. For example, a key improvement from the interviewees' perspective would be integrating the SM system with existing information systems. Such a change might increase the engagement of providers with the system, reduce total workload in the clinic and potentially reduce variability across sites. In terms of user-technology fit focused on patients, systems might be developed that flag key words in messages – such as chest pain – that direct the patient to call rather than to send a secure message.

To increase fit, implementation can also address organizational practices at multiple levels. For example, in our context, user-technology fit might be improved by more effective training of clinicians and patients. Effective training may increase adoption among staff members and decrease inappropriate use of the SM by patients. Organization-technology fit can be improved by developing policies or guidance to address areas of uncertainty, such as when a secure message should be copied to the EHR or who will open messages first. As the number of patients using SM and message volume increase, a policy to address workload is particularly important. Environment-technology fit can be improved by providing guidance to patients about the best use of the system, with personalized training to patients when needed.

Conclusion

The clinical workflow for SM is complex and includes variations as an SM system is implemented in different clinic settings. The direct involvement of patients as users also contributes to its complexity. Based on interview data, we identified eight workflow issues that emerged in an SM implementation, and used a multi-level technology-fit framework to examine them. To improve fit, design and implementation strategies should encompass both technical improvements and organizational practices, and be sufficiently comprehensive to address workflow issues at different levels. While this study was based on data from a single organization and SM system, which influenced to some extent the issues identified, the multi-level framework and analysis of fit are applicable for any system and organization. With the emphasis on SM in meaningful use criteria, which is likely to increase the number of patients using SM and the overall volume of messages, organizations need insights about how to implement SM systems to ensure that SM increases access to providers, improves communication and eventually leads to more effective, timely, patient centered and safe care.

Workflow issues have been explored in other technology studies (e.g. EHR, CPOE), but SM workflow differs because patients are also active users. We identified two workflow issues related to patient use, including inappropriate use of messaging by patients and patients' expectation of quick (often 24 hour or sooner) response. Both issues relate directly to patient satisfaction and influenced staff workflow. Because the study design focused on staff interviews, our results emphasize the importance of effective communication between patients and staff, to provide relevant information without requiring additional rounds of messages or phone calls. Our future research includes an analysis of messages between the patients and clinics to better understand communication patterns. From a system design perspective, templates that can structure patient interactions have been shown to be effective[26]. Examining SM use from the patients' perspective, in daily living environments, is also likely to be a fruitful area for additional research.

Acknowledgement

This study was funded by the Veterans Health Administration (QUERI RRP 11-409, PI: Woods), with additional support from the New England Veterans Engineering Resource Center. We would like to thank the members of the eight PACT teams who agreed to be interviewed.

References

1. Zhou YY, Kanter MH, Wang JJ, Garrido T. Improved quality at Kaiser Permanente through e-mail between physicians and patients. Health Aff. 2010 Jul;29(7):1370–5.

2. Wade-Vuturo AE, Mayberry LS, Osborn CY. Secure messaging and diabetes management: experiences and perspectives of patient portal users. J Am Med Inform Assoc. 2013 May 1;20(3):519–25.

3. Harris LT, Koepsell TD, Haneuse SJ, Martin DP, Ralston JD. Glycemic control associated with secure patient-provider messaging within a shared electronic medical record: a longitudinal analysis. Diabetes Care. 2013 Sep;36(9):2726–33.

4. Lin C-T, Wittevrongel L, Moore L, Beaty BL, Ross SE. An Internet-based patient-provider communication system: randomized controlled trial. J Med Internet Res. 2005;7:e47.

5. Woods SS, Schwartz E, Tuepker A, Press NA, Nazi KM, Turvey CL, et al. Patient experiences with full electronic access to health records and clinical notes through the My HealtheVet Personal Health Record Pilot: qualitative study. J Med Internet Res. 2013;15:e65.

6. Muller D, Logan J, Dorr D, Mosen D. The effectiveness of a secure email reminder system for colorectal cancer screening. AMIA Annu Symp Proc. 2009;2009:457–61.

7. Liederman EM, Morefield CS. Web Messaging: A New Tool for Patient-Physician Communication. J Am Med Informatics Assoc. 2003;10:260–70.

8. Reid RJ, Fishman PA, Yu O, Ross TR, Tufano JT, Soman MP, et al. Patient-centered medical home demonstration: a prospective, quasi-experimental, before and after evaluation. Am J Manag Care. 2009;15:e71–e87.

9. Shimada SL, Hogan TP, Rao SR, Allison JJ, Quill AL, Feng H, et al. Patient-provider secure messaging in VA: variations in adoption and association with urgent care utilization. Med Care. 2013 Mar;51(3 Suppl 1):S21–8.

10. Heyworth L, Clark J, Marcello TB, Paquin AM, Stewart M, Archambeault C, et al. Aligning medication reconciliation and secure messaging: qualitative study of primary care providers' perspectives. J Med Internet Res. 2013 Jan;15(12):e264.

11. Wakefield DS, Mehr D, Keplinger L, Canfield S, Gopidi R, Wakefield BJ, et al. Issues and questions to consider in implementing secure electronic patient-provider web portal communications systems. Int J Med Inform. 2010 Jul;79(7):469–77.

12. Ancker JS, Kern LM, Abramson E, Kaushal R. The Triangle Model for evaluating the effect of health information technology on healthcare quality and safety. J Am Med Inform Assoc. 19(1):61–5.

13. Bohmer RMJ. The Four Habits of High-Value Health Care Organizations. N Engl J Med. 2011;365:2045–7.

14. Ash JS, Bates DW. Factors and forces affecting EHR system adoption: report of a 2004 ACMI discussion. J Am Med Inform Assoc. 2005;12(1):8–12.

15. Patterson ES, Cook RI, Render ML. Improving patient safety by identifying side effects from introducing bar coding in medication administration. J Am Med Inform Assoc. 2002;9(5):540–53.

16. Ash JS, Sittig DF, Poon EG, Guappone K, Campbell E, Dykstra RH. The extent and importance of unintended consequences related to computerized provider order entry. J Am Med Inform Assoc. 2007;14(4):415–23.

17. Venkatesh V, Morris MG, Davis GB, Davis FD. User Acceptance of Information Technology: Toward a Unified View. MIS Q. 2003;27(3):425–78.

18. Pasmore WA. Designing Effective Organizations: The Sociotechnical Systems Perspective. New York: John Wiley and Sons; 1988.

19. Blumenthal D. Stimulating the adoption of health information technology. N Engl J Med. 2009;360:1477–9.

20. Holden RJ, Karsh B-T. A Review of Medical Error Reporting System Design Considerations and a Proposed Cross-Level Systems Research Framework. Hum Factors J Hum Factors Ergon Soc. 2007 Apr 1;49(2):257–76.

21. Karsh B-T, Escoto KH, Beasley JW, Holden RJ. Toward a theoretical approach to medical error reporting system research and design. Appl Ergon. 2006 May;37(3):283–95.

22. Ozkaynak M, Brennan PF, Hanauer D a, Johnson S, Aarts J, Zheng K, et al. Patient-centered care requires a patient-oriented workflow model. J Am Med Inform Assoc. 2013;20(e1):e14–6.

23. Thielke S, Hammond K, Helbig S. Copying and pasting of examinations within the electronic medical record. Int J Med Inform. 2007 Jun;76 Suppl 1:S122–8.

24. Siegler EL, Adelman R. Copy and paste: a remediable hazard of electronic health records. Am J Med. 2009 Jun;122(6):495–6.

25. Couchman GR, Forjuoh SN, Rascoe TG, Reis MD, Koehler B, Walsum KL Van. E-mail communications in primary care: what are patients' expectations for specific test results? Int J Med Inform. 2005 Jan;74(1):21–30.

26. Leventhal T, Taliaferro JP, Wong K, Hughes C, Mun S. The patient-centered medical home and health information technology. Telemed J E Health. 2012 Mar;18(2):145–9.

Visualization of Patient Prescription History Data in Emergency Care

Selcuk Ozturk, MS, Mehmet Kayaalp, MD, PhD, Clement J McDonald, MD
Lister Hill National Center for Biomedical Communications,
U.S. National Library of Medicine, National Institutes of Health, Bethesda, MD

Abstract

Interpreting patient's medication history from long textual data can be unwieldy especially in emergency care. We developed a real-time software application that converts one-year-long patient prescription history data into a visually appealing and information-rich timeline chart. The chart can be digested by healthcare providers quickly; hence, it could be an invaluable clinical tool when the rapid response time is crucial as in stroke or severe trauma cases. Furthermore, the visual clarity of the displayed information may help providers minimize medication errors. The tool has been deployed at the emergency department of a trauma center. Due to its popularity, we developed another version of this tool. It provides more granular drug dispensation information, which clinical pharmacists find very useful in their routine medication-reconciliation efforts.

Introduction

Preventing Medication Errors, a report published by Institute of Medicine, drew the attention of the healthcare community, policy makers, and the general public to the gravity of serious clinical mistakes due to preventable medication errors.[1] As recognized by the Joint Commission, taking an accurate and current medication history can be difficult to accomplish, especially when patients are severely ill, injured, disabled, extremely young, or unable to articulate themselves in English.[2] These conditions occur more frequently in emergency rooms. In the case of a disaster, even a well-prepared emergency department can become overwhelmed.

We developed a software system that aims at helping clinicians obtain necessary patient medication history quickly and accurately both in disaster and in routine emergency care. Our effort is part of a joint enterprise called Bethesda Hospitals' Emergency Preparedness Partnership (BHEPP).[3] Other members of BHEPP include Clinical Center at National Institutes of Health (NIH), Suburban Hospital Johns Hopkins Medicine and Walter Reed National Military Medical Center. We developed and deployed our system at the Emergency Department (ED) of Suburban Hospital.

The system first receives patient registration information from the hospital's Health Level 7 (HL7) messaging system and then queries the prescription history of the registered patient at Surescripts. Surescripts, in turn, searches that patient's prescription history in the databases of its network of pharmacy benefit managers (PBMs).[4] If the patient is found, we receive the full-year of prescription history of the patient. Our software system parses the list of prescription fulfillment activity, converts it into a visual timeline chart, and sends it to a printer in the ED. The entire process takes approximately 4 seconds.

Surescripts usually provides a short summary of the report, which lists each drug's fulfillment count and the first and last fulfillment dates within the last twelve month period. However, some pharmacy benefit managers do not deliver NDC (National Drug Code) numbers with each drug name. In those cases, the summary reports become long lists of individual drug dispensations. In a high polypharmacy case of an elderly patient, drug history data can easily be 15- to 20-pages long. Such a long list of medication dispensation data is difficult to read and interpret. When quick emergency care response is required, decoding such a document can become a burden. Our system ameliorates this situation by grouping the drugs with the same name and dose together, making it an easily interpretable summary as it is done by Surescripts when drug NDC numbers are available.

Emergency care providers usually do not have time to read long drug lists containing various doses, dates, dispensation amounts and so on. Our system creates a timeline chart (see Figure 1). On the right column, drugs grouped by name and dose are ordered from the most to the least recent. The drugs that are currently in use are printed in boldface. The time flows from left to right as indicated with the dates on the horizontal axis at the top of the chart. With this chart, care providers can glean a year-long prescription history of the patient very quickly The example presented in Figure 1 was produced from a de-identified two-page, more than 100-line-long summary of drugs without NDC numbers.

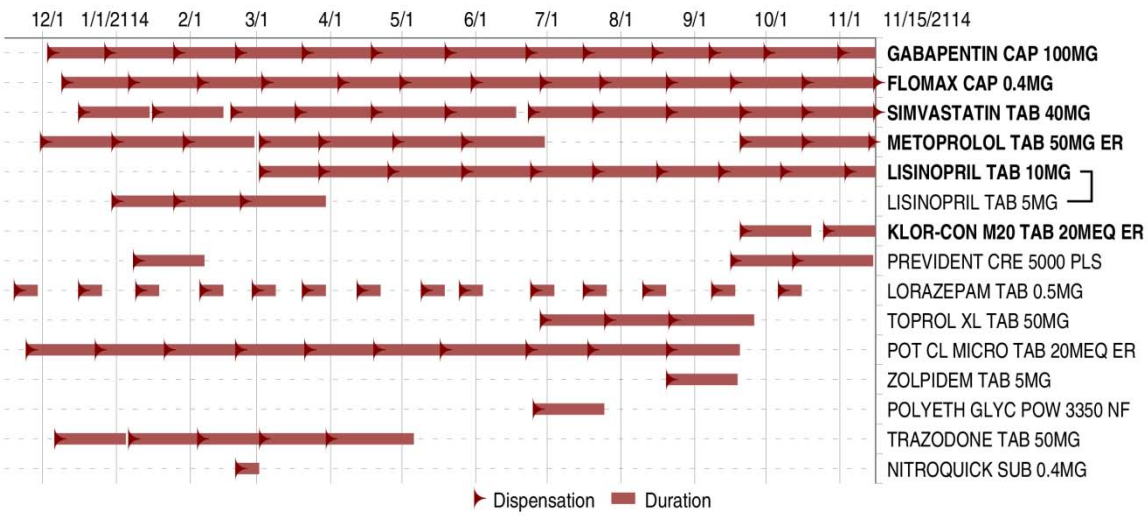


Figure 1. Patient Prescription History Data on a Timeline Chart

Methods

Surescripts' feed provides two types of HL7 message information: An RDS, pharmacy dispense message, and an ORU, unsolicited transmission of observation message.[5] Our system processes the observation message to create the graphical timeline chart. In the input, each dispensation is represented in three subsequent OBX HL7 segments (see Figure 2). These segments contain tabular information about drug names, dosage forms, dispense dates, dispense units, prescription durations, prescribers, dispensing pharmacies, data entry types and information sources.

```
OBX|45|TX|MEDS^DRUG||IBUPROFEN 800 MG TABLET (IBUPROFEN)||||||F
OBX|46|TX|MEDS^DRUG||Fill Date    Qty         Days   Prescriber  Pharmacy   Type    Source||||||F
OBX|47|TX|MEDS^DRUG||01/01/2020   12          3                  7          Claim    CMX||||||F
```

Figure 2. A Small Portion of De-identified Observation Message Representing a Single Dispensation

In addition to this detailed dispensation information, we receive a summary textual report that includes patient identifiers, including the name, the zip code, the date of birth, and the gender of the patient. This text also includes providers and pharmacy lists related to the dispensations.

We parse both sets of information using Perl scripts and compute the "drug use" intervals by adding the prescription duration days to the dispensation date of the drug. We truncate all intervals of recently dispensed drugs to the current date if they pass beyond the current date of information. We order the list of the drugs according to the "last drug use date," i.e. the end date of the most recent interval for each drug. This way all medications currently in use come at the top of the list. The secondary ordering among equals is the total duration of use. If the secondary ordering does not break the tie for some, we order them alphabetically.

We also group medications with the same name but in different dosages (see LISINOPRIL TAB in Figure 1) and connect them together with a bracket emphasizing their relation to each other. Although the medications listed at the top are the current ones, we further emphasize the current medication by printing their labels in boldface.

After all information is parsed from input and the necessary computations are made, our Perl script translates this information into Postscript. Since Postscript is both a representation language of graphical information and a specialized programming language, it is capable of computing and adjusting a number of graphical parameters automatically. We first calculate the maximum width of the right-most column, where we list the medications, and then divide the remaining at the left into 365 (day) logical units. Postscript scales the width of the chart as well as all other visual elements accordingly.

Unfortunately, the input is not always clean or simple. For example, a refill usually does not occur on the same day that the patient is out of the current supply of medication. It is expected that the patient gets his/her refills a few days

before the end of use day. However, what we frequently observe is multiple refills of the same drug within a couple of days. There may be different reasons for this type of noise, but apparently, the same dispensation information is passed to Surescripts from different sources, for example both by the pharmacy and by the pharmacy benefit managers, separately.

If the same label in the same form and dosage had been refilled within a couple of days with the same amount, and the duration of the medication is not short (e.g., 15-day or longer), we could assume that the second refill was a duplicate and draw the chart accordingly. Although we believe that our heuristic was quite good, we could not validate it for an extended period due to the lack of resources. In order not to take the associated liability, we decided not to continue with our heuristic plan. Instead, we decided to develop a visual representation of the data that reflects the incoming information as closely as possible.

Our problem was how to present to the provider overlapping dispensation information with clarity. We came up with a new version of the chart (see Figure 3). In this chart, each drug has a band allocated for printing the "drug use interval" lines. Now, we draw these intervals as a thin line through the mid-height of this band unless two or more intervals overlap. In that case, we shift the overlapping interval lines up and down by a few pixels in an alternating up and down pattern until the bound of the band is reached. Our implementation allows up to 9 concurrent intervals. When that level of multiplicity is reached, we put a footnote above the line, indicating that there are too many overlapping dispensations to plot. This way, each dispensation interval has a distinct beginning and end point and all overlaps are clearly visible.

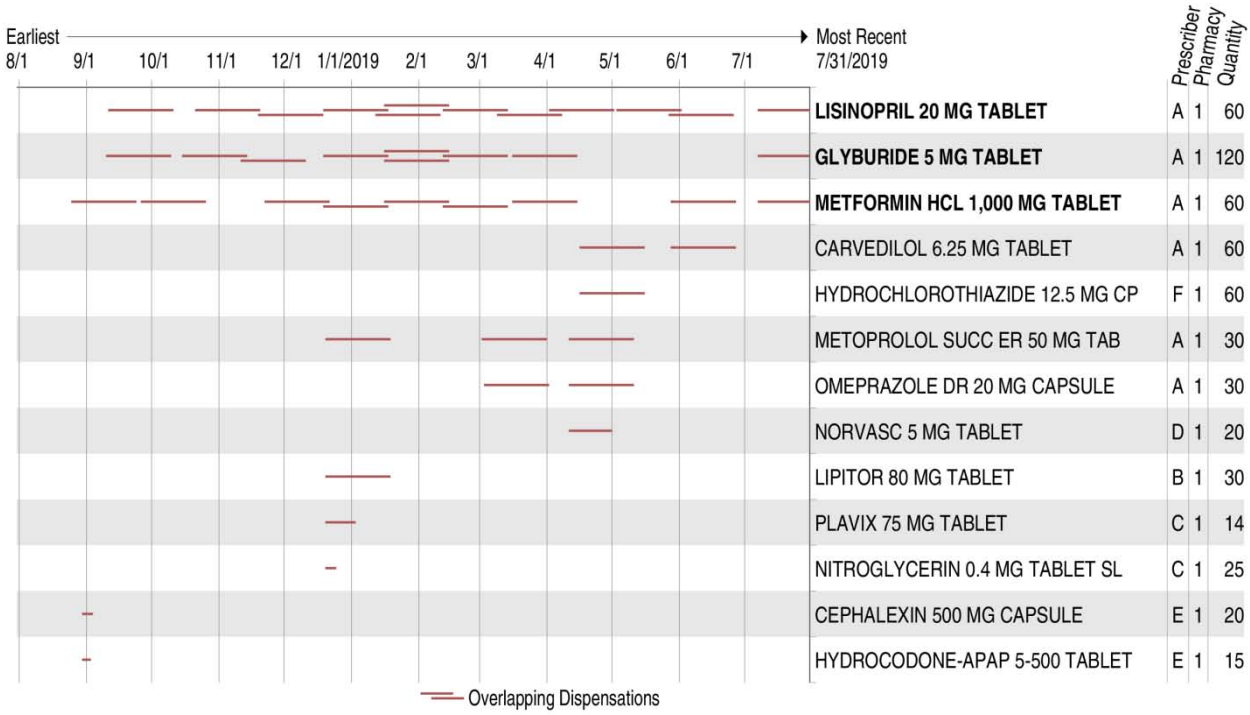


Figure 3. Patient History Data with Overlapping Dispensation Information on a Timeline Chart

In this version, we also decided to include the prescriber and pharmacy information when available along with the dispensed quantity in the chart: The last three columns on the right in Figure 3 above. The letters and numbers in the prescriber and pharmacy column refer to the lists printed later in the report. In the course of a one-year history, the same drug may be prescribed by different physicians, dispensed by different pharmacies, and/or prescribed in different amounts. The columns located on the far right display only the most recent information about that particular prescription and dispensation.

Results

By using Perl and Postscript languages, we created information-rich visual timeline displays, which can be absorbed by a clinician almost at a glance saving time and facilitating care. A sample copy of a de-identified printout that the providers receive is displayed in its entirety in Figure 4.

On the top of this page, we provide the timestamp of the information, patient name (redacted with Xs in Figure 4) and page number followed by a table containing patient identifiers, including name, address, gender, and date of birth. Before printing the standard disclaimer text, we provide a medication history date range. In this example, it ranges from 08/01/2018 to 07/31/2019. Due to the de-identification, the surrogate dates of this example lie in the future.

Starting about one-third of the way down the page, we draw a timeline corresponding to the date range. The direction of time from left to right is denoted by an arrow. At the tip of the arrow, the current full date (i.e., the date when the information is acquired) is provided. The timeline is divided into monthly portions and each tick mark is labeled with the first day of the corresponding month (Month/Day). Only the label of the New Year day tick mark contains the year information (i.e., 1/1/2019) to aid the provider for a quick temporal orientation in the chart.

Every other row is highlighted with a band for an easy read. To the right of medication labels, we inserted three columns. Both the prescriber and pharmacy columns are narrow, holding only one letter key to denote the prescriber reference or one digit key to denote the pharmacy reference. The last column is wider, to hold up to 3 digits denoting the most recently dispensed quantity. The pharmacy and prescriber tables with proper cross-reference keys are located below the chart.

The disclaimer at the top of the page could be moved to the end of the report since it does not convey any new information for the experienced provider; however, mostly for legal reasons, we had to place it where it currently is. We placed another set of standard remarks at the bottom of the page.

Discussion

Prescription data is an essential part of the medical history of the patient, but it can be quite voluminous. In emergency care, when rapid response is necessary, providers may not have time to sort through 15-page long prescription fulfillment report. The problem would be exacerbated during a disaster, as emergency departments would at that time become triage centers as well. Our project started with this kind of scenario in mind and we adhered to that perspective in all design decisions.

In order to provide all necessary information with utmost clarity, we attempted to follow best practices in visual display design.[6] We tried to avoid any superfluous text and visual design elements. We strived to strike a balance between the proportions designated to the visual timeline and to the textual information, i.e. drug label, dose, quantity, prescriber, and pharmacy information. Our top-down order of drugs has a narrative quality by itself, capturing part of a medical story of the patient.[6]

These design principles guided us in both versions of the timeline chart. We believe the first version is most suitable for emergency care providers, since it does not contain any unnecessary information or noise such as duplications. On the other hand, the second version seems very useful for pharmacists who need to conduct medication reconciliation and need every bit of information about the dispensation detail.

The three columns of prescriber, pharmacy, and quantity cross-references also helped us to eliminate all other textual data, since per consultation with the pharmacists at the trauma center, the current layout constitutes the necessary and sufficient information needed for medication reconciliation.

In addition to its valuable use in routine emergency care and in medication reconciliation, the visualization of the prescription history data help healthcare providers to quickly spot drug abuse and drug seeking behavior. We also believe that due to simple physical proximities of the currently used drugs, providers have a better chance of spotting pairs of medications with adverse drug event potentials. Note however that the efficacy and utility of our design and its effect on the patient outcome need to be measured and evaluated by future studies.

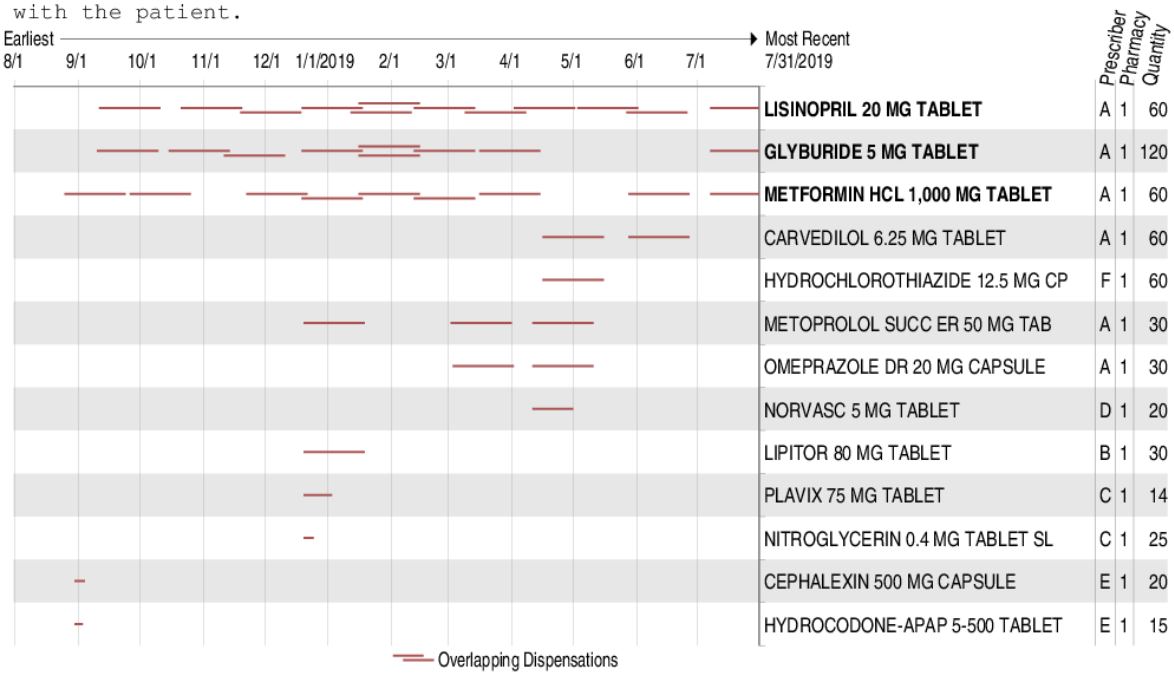
As one of our reviewers rightly pointed out, our design would benefit from a mapping between brand name drugs and their generic counterparts. If such a mapping were in place, we could move up Toprol XL tab 50mg (brand) and clamp it with its generic version Metorprolol tab 50mg ER (generic) in Figure 1. This way the clinician could immediately see that there was no discontinuity of the drug use.

RxHub Patient Medication History

Patient Id: XXXXXXX
Name: XXXXXXXXXXX
Address: X
XXXXXX, XX XXXXX
DOB: XX/XX/XXXX Gender:Male

Medication History Date Range: 08/01/2018 - 07/31/2019

DISCLAIMER: Certain information may not be available or accurate in this report, including over-the-counter medications, low cost prescriptions, prescriptions paid for by the patient or non-participating sources, or errors in insurance claims information.
Some information may be duplicated because of multiple data sources.
The provider should independently verify medication history with the patient.



Key	Source Pharmacy	Pharmacy Phone
1	ACME PHARMACY	555-555-5555

Key	Prescriber	Key	Prescriber
A	William Osler MD	B	Virginia Apgar DDS
C	C. Everett Koop MD	D	Florence R. Sabin MD
E	Louis Pasteur MD	F	Asclepius Hippocrates MD

Figure 4. Patient Prescription History Report

Our approach can be extended to a visual interactive display on tablets or monitors. In such an interactive environment, one could allow for drilling down and zooming in for more detailed information, allow for selectively narrowing the report to certain drug classes to focus on critical elements in the report. Even, automatic drug interaction warnings could be implemented.

Funding

This work was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

References

1. Committee on identifying and preventing medication errors. Preventing medication errors. Institute of Medicine of the National Academies. National Academic Press. 2007.
2. The Joint Commission. Using medication reconciliation to prevent errors, Issue 35.January 25, 2006.
3. Bethesda Hospitals' Emergency Preparedness Partnership. URL http://www.bethesdahospitalsemergencypartnership.org. Accessed on 3/1/2014.
4. Payers & PBMs connected to Surescripts. URL: http://surescripts.com/network-connections/mns/payers-and-pbms. Accessed on 3/1/2014.
5. Health Level Seven International. HL7 glossary of terms. 2012.
6. Tufte ER. The visual display of quantitative information.1983. Graphics Press, CT.

Locating Relevant Patient Information in Electronic Health Record Data Using Representations of Clinical Concepts and Database Structures

Xuequn Pan, PhD[1, 2], James J. Cimino, MD[1, 2]
[1]Lister Hill National Center for Biomedical Communications, National Library of Medicine;
[2]Laboratory for Informatics Development, NIH Clinical Center; Bethesda, MD

Abstract

Clinicians and clinical researchers often seek information in electronic health records (EHRs) that are relevant to some concept of interest, such as a disease or finding. The heterogeneous nature of EHRs can complicate retrieval, risking incomplete results. We frame this problem as the presence of two gaps: 1) a gap between clinical concepts and their representations in EHR data and 2) a gap between data representations and their locations within EHR data structures. We bridge these gaps with a knowledge structure that comprises relationships among clinical concepts (including concepts of interest and concepts that may be instantiated in EHR data) and relationships between clinical concepts and the database structures. We make use of available knowledge resources to develop a reproducible, scalable process for creating a knowledge base that can support automated query expansion from a clinical concept to all relevant EHR data.

Introduction

The retrieval of clinical data from electronic health records (EHRs) for whatever reason, including care of a particular patient or broader inquires to support research, can be thought of as starting with some concept of interest and then expanding to include the concepts, and their representations, that actually appear in EHR data. For example, a clinician looking for evidence that a patient has had a particular disease or a researcher seeking records of all patients with a particular disease will naturally search record documents for codes and text that correspond to the disease in question. However, in order to improve recall, they will likely need to expand their inquiries to include such things as test results or medication orders that imply the presence of the disease. [1]

Once the set of concepts has been selected, there still remains the challenge of finding them in the EHR documents. Diagnoses might appear with some unobvious code (e.g., "Diagnosis not Elsewhere Classified") or synonym. Relevant laboratory results might be easy to find, based on test names, but they might require interpretation of complex results, or they might be buried in a laboratory test comment or even an image of a document obtained from an outside laboratory.

We are developing a set of methods to help overcome these challenges by the creation of an ontology that includes representations of concepts of interest, related concepts that appear in EHR data, and concepts corresponding to EHR database structures. Using Lyme disease as an example concept of interest, we demonstrate the construction and use this ontology through four steps: 1) identify terms related to Lyme disease in available knowledge resources, 2) add these concepts and their relationships to a dictionary of coded terms found in EHR data, 3) connect relevant EHR data concepts identified in the previous step with concepts that correspond to EHR data structures, and 4) identify pathways through the ontology that ultimately relate concepts of interest to database structures to support automated, comprehensive data queries.

Background

Lyme Disease

We selected Lyme disease, the most common tickborne infectious disease in the US, as our concept of interest. Lyme disease is diagnosed based on its typical symptoms (e.g., erythema migrans) and detailed medical history of possible exposure to infected ticks. Serologic testing is important for the diagnosis by looking for evidence of antibodies (IgG or IgM) to the causative agent, Borrelia (B.) burgdorferi. To support the diagnosis of Lyme disease, the Centers for Disease Control and Prevention (CDC) recommends a two-step process: a screening test that uses sensitive enzyme immunoassay (EIA)/enzyme-linked immunosorbent assay (ELISA) or indirect immunofluorescence assay (IFA); after a positive result, a specific immunoblot (Western blot) test should be conducted. [2, 3] The diagnosis is considered confirmed if there is a positive result on both tests.

The Systematized Nomenclature of Medicine – Clinical Terms (SNOMED-CT)

SNOMED-CT is a clinical terminology that represents a wide variety of clinical information that appears in EHRs. SNOMED-CT is composed of concepts, descriptions, and relationships. Each concept contains a unique fully specified name (FSN), a preferred term (PT), and synonyms that are additional terms and phrases about the concept. SNOMED-CT contains 344,000 concepts and a rich set of hierarchical or non-hierarchical relationships. Non-hierarchical semantic relationships connect concepts to relevant concepts from other domains and semantic categories to represent definitional information about the concept. [4]

The Biomedical Translational Research Information System (BTRIS)

BTRIS is a clinical research data repository at the National Institutes of Health (NIH) that collects data from over 35 NIH sources with 477,000 human subjects since 1976. BTRIS's major data sets include patient demographics, research study enrollment information, laboratory results, medication orders and administration records, and notes and reports from EHR systems. [5] All data in BTRIS are coded with the NIH's Research Entities Dictionary (RED), a terminology resource that comprises the controlled terminologies used by systems that supply data to BTRIS. Each source term is associated with a specific RED concept; knowledge about the terms is represented as literal-valued properties or through semantic relationships between concepts, including hierarchical relationships organized in a directed acyclic graph. The RED includes concepts that correspond to terms actually found in BTRIS ("data concepts"), database structure concepts that correspond to database structure ("database concepts"), and general biomedical concepts that are not actually present in EHR data ("knowledge concepts").

Methods

The ultimate goal of the project is to facilitate the matching of patient data in EHRs to a user's concepts of interest. For example, if a clinician is interested in finding evidence that a particular patient has Lyme disease, or a researcher is interested in identifying all records of patients with Lyme disease, comprehensive retrieval would at least need to examine problem lists, clinician notes and diagnostic test results. We therefore developed a method to expand the RED to include the knowledge necessary to support queries of these disparate data sources in BTRIS, based on an initial concept of interest. For this initial work, we focused on the knowledge related to Lyme disease.

General Knowledge Relating Concepts of Interest to Clinical Data

We reviewed clinical information about Lyme disease from authoritative sources and identified relevant key terms. This approach applied general knowledge to identify ways in which terms interrelate, such as "diseases are caused by organisms", "organisms produce antigens", "antibodies are produced in response to antigens", "laboratory tests detect organisms and substances", etc.

Identifying Specific Concept Relationships in Knowledge Resources

We examined the ways in which SNOMED CT and the RED represent general knowledge about diseases and their related concepts. We then explored and identified specific concepts and relationships in knowledge resources, such as Lyme disease and its associated terms. We generated a list of concepts based on parent-child relationships among the retrieved concepts. We then expanded our list to include concepts from other domains related through non-hierarchical relationships, such as "Borrelia (organism) is the causative agent of Lyme disease (disease)".

Adding the Concepts and their Relationships to RED

After review by domain experts, we added concepts and relationships discovered in the previous step into the RED through the usual maintenance process. We selected the concepts to be added and linked the concept of interest to EHR data concepts through their associations defined.

Linking EHR Data Concepts to BTRIS Database Structures

Finally, we linked EHR data concepts to database concepts that correspond to the BTRIS database structures, based on the BTRIS data.[5] For example, EHR data concept, Lyme disease (RED Code C114654) is linked to the specific database concept (C2179248) that represents the Observation_Value_Text column in the Observation_General table where diagnosis information was stored.

Results

General Knowledge Relating Concepts of Interest to Clinical Data

We reviewed clinical information about Lyme disease from CDC and NIH Web sites. The terms identified included "Lyme disease", "Borrrelia burgdorferi", "erythema migrans", and various antigens of and antibodies to B. burgdorferi, as well as terms for polymerase chain reaction (PCR) tests for B. burgdorferi DNA, serologic tests, antibody screens, and Western blot to measure those antibodies. Based on these findings, we therefore generated a general knowledge structure (Figure 1) in which the concept of interest is connected to terms from clinical domains of disease, organism, substances, and laboratory tests.

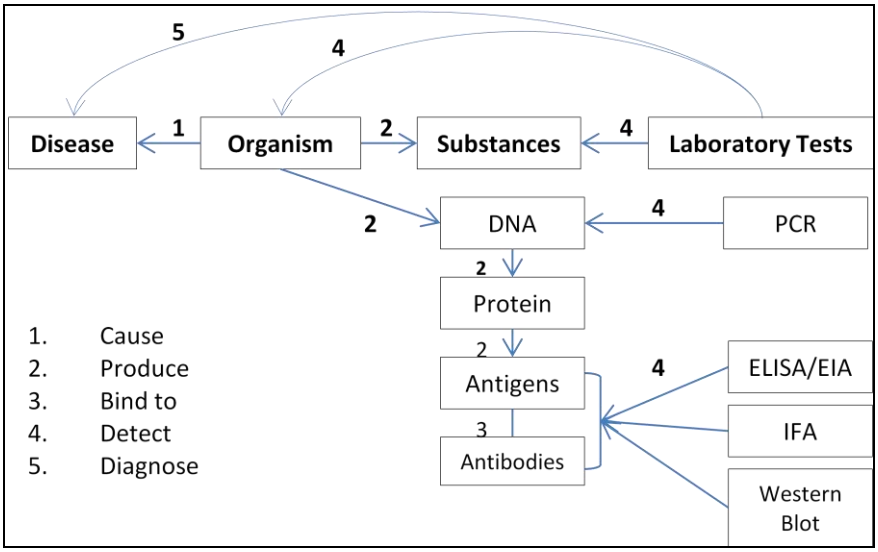


Figure 1. General knowledge structure for Lyme disease

Identifying Specific Concept Relationships in Knowledge Resources

In SNOMED CT, we retrieved a total of 64 SNOMED CT concepts related to Lyme disease including terms for 27 procedures, 21 substances, disorders, and 1 organism. The three non-hierarchical relationships we found are shown in Table 1. Figure 2 depicts for examples of concepts related to Lyme disease in SNOMED CT.

Table 1. Non-hierarchical relationships between Lyme disease and other terms in SNOMED CT

Non-hierarchical Relationship	Examples
disorder|causative agent|organism	Lyme disease|Caustive Agent|Borrelia
procedure|component|substance	Measurement of Borrelia burgdorferi antibody | Component | Borrelia burgdorferi antibody Measurement of Borrelia burgdorferi 29 kDa antibody | Component | Borrelia burgdorferi 29 kDa antibody Borrelia burgdorferi band pattern dectected|Component| Borrelia burgdorferi antibody Borrelia burgdorferi DNA Assary|Component| Borrelia burgdorferi DNA
finding|interpret|procedure	Lyme ELISA equivocal|Interpret|Lyme ELISA test

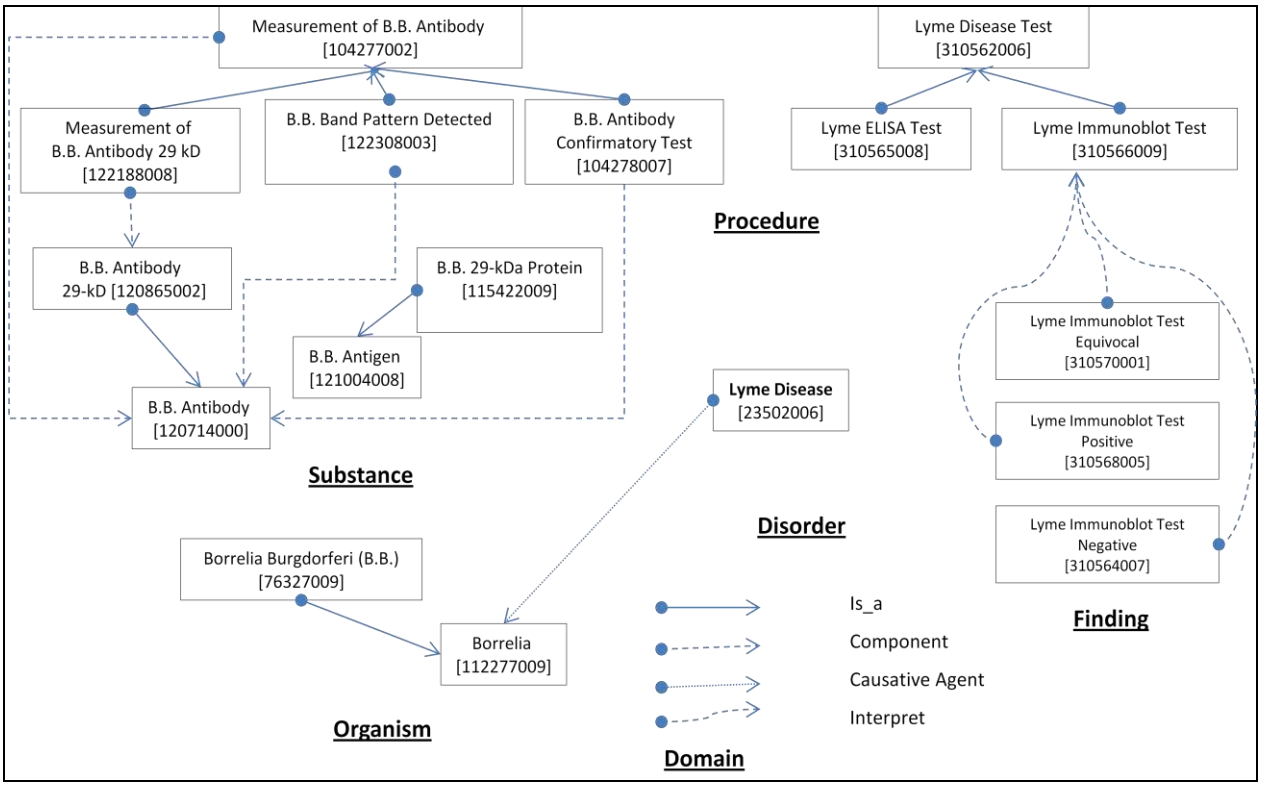


Figure 2. Examples of concepts related to Lyme disease in SNOMED CT. Numbers in brackets are SNOMED CT identifiers.

We retrieved one diagnosis concept from the RED (Lyme disease), 101 laboratory procedure concepts (tests and panels), and one organism concept (Borrelia burgdorferi). A total of 57 test and panel concepts are actually used to code EHR data. RED relationships found in these concepts included "procedure has component", "diagnostic procedure has targeted infectious organism", and "procedure has measured gene product". Figure 3 illustrates the example of the Lyme Disease Cerebrospinal Fluid (CSF) Antibody Panel (C1163227) and its two components: an antibody screen test (C116326) and a confirmation panel (C119063). The confirmation panel, in turn, includes Western blot IgG and IgM tests. The Lyme Disease CSF Antibody Panel targeted the bacterium B. burgdorferi and measured the antibodies.

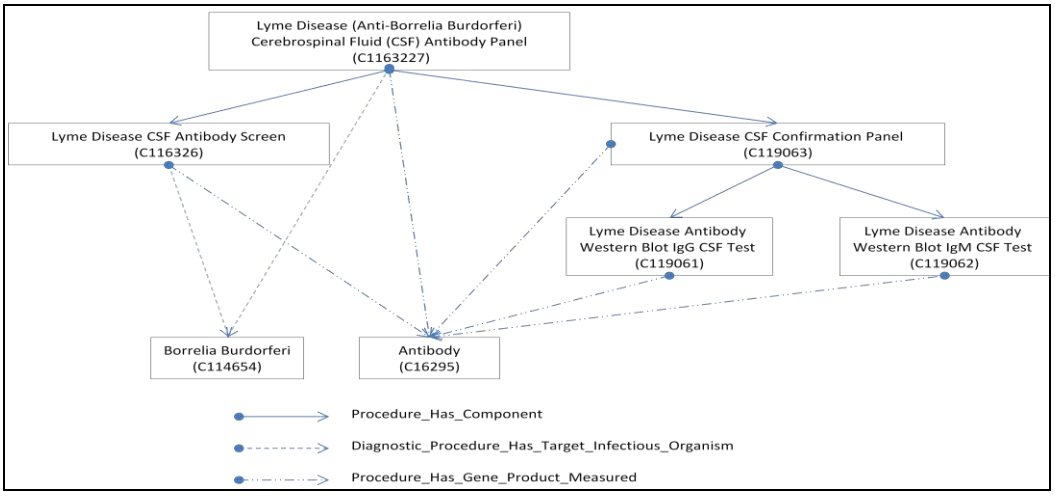


Figure 3. Examples of concepts related to Lyme disease in the RED. Numbers in brackets are RED codes.

Adding Concepts and Relationships to the RED

The RED substance concepts were expanded to include B. burgdorferi gene products, antigens, antibodies, IgG and IgM, and IgG and IgM to specific antigens. We did not add additional laboratory terms because the RED laboratory terms already cover the available EHR data. Terms for findings were not included in the current stage of the project.

Instances of the "Substance measured" relationship were added between the substance and laboratory test terms. Figure 4 illustrates examples of added concepts and links. A "has causative agent" relationship was created and an instance of that relationship was added between Lyme disease and Borrelia burgdorferi.

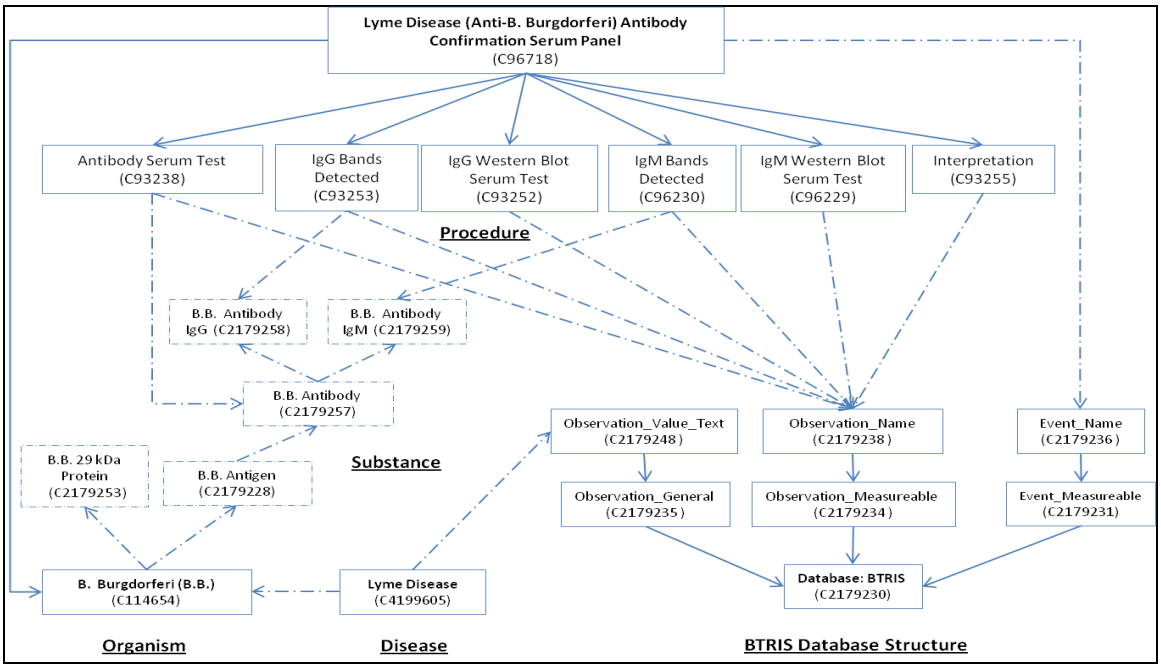


Figure 4. Concepts and relationships added to RED to represent EHR database concepts and relate them to EHR data concepts.

Linking EHR Data Concepts to BTRIS Database Structures

We related domains of EHR data concepts to specific database concept as follows. Diagnostic terms were related to the Observation_General table. Laboratory test order terms (panels) were related to the Event_Measurable table. The terms that were both tests and orders (basically, panels containing only one test) were related to both Event_Measurable and Observation_Measurable tables.

We defined the database column associated with laboratory test results (not shown in Figure 4). A specific organism's RED code appears in the Observation_Value_CONCEPT column, and its name appears in the Observation_Name. Table rows associated with laboratory tests of particular interest included those for which the RED concepts were "Specimen" and "Culture Comment" in the Observation_Name column.

In BTRIS, laboratory finding values are in Observation_Value_Text or Observation_Value_Numeric columns, depending on the format of the value. When a RED concept was available (such as for B. burgdorferi), the code was recorded in the Observation_Value_Concept column. Additional textual information is stored in Observation_Note and Observation_Value_Name columns.

Identify Pathways that ultimately Relate concepts of interest to Data Structure

The general knowledge structure supports the semantic expansion of the concept of interest to EHR data concepts that connected with database concepts. The expansion allows us to follow the concept of interest to its associated concepts from different data domains to identify patients. In the following example, we demonstrate how starting from Lyme disease, to generate SQL queries and retrieve patients who had diagnostic (confirmatory) laboratory tests. We looked up Lyme disease and related EHR data concepts which led to B. burgdorferi, the causative

organism, and then B. burgdorferi Antibody IgG and IgM, the measureable substances (step 1). We set concepts association value as where the concepts are these two substances and the association is substance measured (step 2). We used this list of B. burgdorferi Western blot tests to obtain patient data from BTRIS (step 3).

Step 2: To obtain the list of concepts representing tests measured B. burgdorferi Antibody IgG and IgM

```
"SELECT Distinct [Concept_ID] , [Database_Concept_Column], [Database_Concept_Table]
  FROM [BTRIS].[RED_Knowledge_Structure]
  WHERE Substance_Measured in (' B. burgdorferi Antibody IgG ',' B. burgdorferi Antibody IgM')
```

The list of tests includes C92569, C93238, C93253, C96230, C93416, C96770, C96713, C119061, C119062, and they are associated with the Observation_Measurable table, and the Observation_Name_CONCEPT column.

Step 3: To identify patients with laboratory results of B. burgdorferi Antibody Western blot IgM and IgG tests

```
"SELECT *
  FROM [BTRIS].[Observation_Measurable][Database_Concept_Table]
  WHERE Observation_Name_CONCEPT [Database_Concept_Column] in (<List of tests>)
```

Discussion

In this study, we added knowledge of an external clinical terminology and relational database structures to a repository terminology for semantically enhanced EHR data retrieval. Although we only focused on Lyme disease; our specific approach can be applied to any infectious diseases (looking for organisms that cause disease) and this general approach of identifying concept relationships should be applicable to any domain. The knowledge sources used, SNOMED CT and RED may currently be incomplete, but they are expandable and growing. We also consider other knowledge sources such as LOINC or co-associations of terms in PubMed citations [6, 7].

Our approach for knowledge extraction is based on established and validated methods in ontology development. For efficient query evaluation, we can measure patient data retrieval performance. We need to work with human raters to construct a gold standard from our EHR data sets. We can then test the retrieval results based on queries generated against the gold standard to calculate recall and precision. The actual validation work is beyond the scope of this paper.

All the work was done manually. Automated steps can be considered in the process of knowledge extraction from knowledge resources and mapping database structure concepts. Since the locations of concepts in tables and columns have been specified, SQL statements can be generated programmatically [8]. Currently our knowledge model does not handle the assessment of results (for example, to determine if a test result is positive or negative). Additional steps in the SQL query can assist with this or results could be processed through some other means if this is desirable. A richer knowledge model can open more fields in the EHR systems and support dynamic SQL for users.

This study only focuses on coded EHR data. The queries can be expanded to text searches. For example, some lab tests' names are too general without specific clinical meaning, such as "Other Laboratory Test" so that a further linking process is needed to connect text data stored in the results or notes. The names of all related concepts and their synonyms are the key terms to identify patient information from textual clinical data. Natural language processing (NLP) techniques can be applied and help the connection of the concept of interest to relevant clinical data hidden in the text.

This knowledge model can be applied to other EHR system. We consider the use of HL7 RIM as the intermediary, with the FHIR mapping [9] as the key to connect our knowledge model with local coding systems used in other system. Through mappings, the knowledge model can be replicable, and relevant concepts related to concepts of interest will be connected to their local EHR data concepts. The work left is to add and link database structure concepts of their own databases.

Conclusion

We constructed a simple general knowledge structure and used it to relate a concept of interest, Lyme disease, to EHR data concepts and their database concepts with specified values of identifiers (table and column) corresponding to BTRIS data structures. This method could be applied to any EHR database in which coded data are stored in specific structures based on their classification.

Acknowledgements

This research was supported by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM) and Lister Hill National Center for Biomedical Communications (LHNCBC). This research was also supported in part by an appointment to the NLM Research Participation Program, administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the US Department of Energy (DoE) and the NLM.

Disclaimer

The views and opinions of the authors expressed herein do not necessarily state or reflect those of the National Library of Medicine, National Institutes of Health or the US Department of Health and Human Services.

Competing Interests

None

References

1. Pathak J, Kho AN, Denny JC.Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. J Am Med Inform Assoc. 2013 Dec; 20(e2):e206-11.
2. The Centers for Diseases and Prevention. Lyme disease. Available from: http://www.cdc.gov/lyme/
3. The National Institutes of Health. Understanding Lyme disease. Available from: http://www.niaid.nih.gov/topics/lymeDisease/Pages/lymeDisease.as
4. International Health Terminology Standards Development Organisation. SNOMED CT® User Guide July 2013 International Release. Available from: http://ihtsdo.org/fileadmin/user_upload/doc/download/doc_UserGuide_Current-en-US_INT_20130731.pdf
5. Cimino JJ, Ayres EJ. The clinical research data repository of the US National Institutes of Health. Stud Health Technol Inform. 2010;160(Pt 2):1299-303.
6. Vreeman DJ, McDonald CJ, Huff SM. LOINC® - a universal catalog of individual clinical observations and uniform representation of enumerated collections. Int J Funct Inform Personal Med. 2010;3(4):273-291.
7. Rindflesch TC, Kilicoglu H, Fiszman M, Rosemblat G, Shin D. Semantic MEDLINE: An advanced information management application for biomedicine. Information Services and Use, 2011, 31(1): 15-21.
8. Post AR, Kurc T, Cholleti S, Gao J, Lin X, Bornstein W, Cantrell D, Levine D, Hohmann S, Saltz JH. The analytic information warehouse (AIW): a platform for analytics using electronic health record data.J Biomed Inform. 2013 Jun;46(3):410-24.
9. FHIR: Fast healthcare interoperability resources. Available from: http://hl7.org/implement/standards/fhir/

Does Query Expansion Limit Our Learning? A Comparison of Social-Based Expansion to Content-Based Expansion for Medical Queries on the Internet

Christopher Pentoney[1], Jeff Harwell[1], Gondy Leroy PhD[1,2]
[1]Claremont Graduate University, Claremont, CA; [2]University of Arizona, Tucson, AZ

Abstract

Searching for medical information online is a common activity. While it has been shown that forming good queries is difficult, Google's query suggestion tool, a type of query expansion, aims to facilitate query formation. However, it is unknown how this expansion, which is based on what others searched for, affects the information gathering of the online community. To measure the impact of social-based query expansion, this study compared it with content-based expansion, i.e., what is really in the text. We used 138,906 medical queries from the AOL User Session Collection and expanded them using Google's Autocomplete method (social-based) and the content of the Google Web Corpus (content-based). We evaluated the specificity and ambiguity of the expansion terms for trigram queries. We also looked at the impact on the actual results using domain diversity and expansion edit distance. Results showed that the social-based method provided more precise expansion terms as well as terms that were less ambiguous. Expanded queries do not differ significantly in diversity when expanded using the social-based method (6.72 different domains returned in the first ten results, on average) vs. content-based method (6.73 different domains, on average).

Introduction

Searching for medical and healthcare information is the third most popular online activity[1]. However, it poses several difficulties to online health information consumers: composing good queries, identifying objective information, and understanding the content. The second and third problems have been the topic of much research. Evaluating the quality of websites automatically and manually has been shown not only to be possible, but also effective[2], and there exist tools and frameworks for evaluating health information online specifically[3,4]. Improving understanding has relied on readability formulas, and more recently, evidence-based corpus-driven simplification. The readability formulas are applied to a variety of text, e.g., for patient education materials[5,6], bereavement materials[7], informed consent forms[8],and even survey instruments[9]. The formulas are recommended to help evaluate text[10] suggested by government instances and in IRB training courses. More recent corpus-based approaches use semi-automated simplification and have demonstrated that using more familiar terms, grammar structures, or shorter noun phrases is associated with both perceived and actual text difficulty[11,12].

The first problem is less frequently addressed or placed as the goal of information systems development in the medical field. This is an issue because knowing what type of information is best to present to a user can be difficult. Especially with complex medical terms and conceptual relationships, naïve users may not have enough knowledge to even form an effective search query. A current, popular solution of generic search engines, e.g., Google, is to automatically offer additional terms to a user that might help better define their search. Relevant terms are typically appended to the user's search based upon popularity of similar searches conducted by other users[13]. While query formation has also been subject of considerable research, it only became a common tool when Google provided their automatic query completion.

This study compares Google's method of query expansion to another method that focuses on the information within the documents being searched. Our goal is to evaluate the effect of using the different query expansion terms on the query and the query results. Since the social-based queries rely on other people's queries, they may make searching less diverse. At the population level, this may lead to less information being read and digested. If this limited information happens to be biased or untrue, the longitudinal effects may affect the general population. To evaluate social-based expansion, we compare it with an objective, text-based query expansion baseline that uses the existing text on the Internet as the basis for query expansion. In order to explore information, it may be more useful to expand based on what exists in the documents rather than what other people are searching. If searchers are already naïve, using their queries would seem to be a somewhat misguided way to finding reliable health information online.

Background and Significant Information

Online Medical Search

With 72% of internet users in the U.S. responding that they have searched online for health information in 2013[14], there is a clear need to deliver this knowledge in a way that presents internet users with reliable information. Furthermore, 77% of people who searched online for information about health said that they began the search with a common search engine (i.e. Google, Bing, Yahoo), and 52% of smartphone owners say that they have even used their phones to search for medical information[14]. People are constantly looking for health information, even on-the-go, and it needs to not only be reliable, but accessible. It is important for search engines to return medical knowledge that is useful and trustworthy. One way to make this information accessible is to guide people to useful information through the use of query expansion. By helping them form their searches, a search engine can provide relevant information.

Query Expansion

Returning relevant and reliable information can be difficult. Natural language itself is often very ambiguous, and health language is no exception; it is further complicated by medical jargon. Query expansion helps guide a user in generating a query by suggesting search terms that may be relevant to their search. Historically, these additional terms have been generated through methods such as relevance feedback, co-occurrence, and ontologies[15]. This is not a trivial problem since words can have multiple meanings and different words can share similar meanings. Many expansion techniques were evaluated at the Text Retrieval Conferences (http://trec.nist.gov/) with varying results. However, regardless of technique, most expansion methods showed improved search results. Today, query expansion as offered by Google is based on overall popularity of search queries already submitted by others[13]. While it has been shown that successive querying (more queries) leads to looking for increasingly serious conditions[16], the effect of query expansion (more terms) is unknown. For the individual, seeing alternate query options may lead to more diverse searching or a different search than originally intended. For the population of all searching individuals, reusing queries may lead to less diverse search and even more limited medical domain knowledge.

While other research has explored different methods to expanding queries in medical text, none has used the actual document text to suggest additional terms. Knowledge-based types of query expansion that leveraged the Unified Medical Language System (UMLS) as a related information source for expanding queries led to increases in precision-recall[17]. Expansion methods that employ the UMLS Metathesaurus for expansion with synonyms or hierarchically related terms[18] showed similar improvements. Using these content-based or knowledge based methods of query expansion have advantages over the social-based methods, and they should be compared.

Research Interests

Our overall interest was to determine the differences between two different methods of expanding medical queries. It is possible that the social-based expansions will not offer diverse enough results, which may result in a misleading picture of the results. This particular project addressed a comparison between the status quo (Google's social-based expansion) and a content-based expansion method that can sometimes present results more clearly. Given these two types of expansion methods, we were able to conduct a study that explored two questions: (1) how do the expansion terms for medical queries differ based on expansion type, and (2) how do results differ based upon expansion type? Google searches were conducted using each of the methods on 1% of the queries that were identified as being medically relevant.

Using existing knowledge, as found in text, to expand medical queries may provide some benefit over socially popular expansion types. Current social-based methods may help to spread common misunderstandings or misinformation, while content-based methods will show term relationships that exist in text already. Furthermore, as popularity of a term increases, it may become easier for people to come across information that is only popularly searched, and much more difficult to find new knowledge. This study investigates both content-based and social-based expansion types.

Social and Content-based Query Expansion

Nearly a quarter (24.2%) of all queries in the AOL User Search Collection was identified as trigrams, whereas only 19.9% were bigrams and 17.4% were unigrams. Since trigrams were so frequent, we decided to limit the study to all three keyword searches.

Social-based expansion i.e., Google's Auto-complete function, involves appending the search with terms that have been popularly used by other individuals who have searched with similar queries. In order to do this type of expansion, we programmatically submitted queries to Google. Google's suggestions were obtained using Google's Autocomplete API[19] and a Python 2.7 script that sent each search with its expansions to the Google Search (URL: https://ajax.googleapis.com/ajax/services/search/web?v=1.0&rsz=8&q=). The returned set contains 10 queries. We use all of these top 10 suggestions to expand upon, although it is unclear what the order is based on. Each suggested term is added to the initial query to provide a more precise search.

Content-based expansion depends upon the text in online documents themselves. We used the 2006 Google Web Corpus[20], which reflects the index of all sites available to Google. Since we used trigram queries, the four-gram data contained in the corpus (approximately 1.3 million) was needed to suggest the additional terms: the original trigram and one extra term. Punctuation and symbols were filtered out before matching the queries to the Web Corpus in order to find text matches.

Table 1 below shows how the two types of expansion yielded different searches. Typically, the content-based expansion led to broader expansion terms, which have more Google results overall. Even though terms may have been related between the expansion types (i.e. "seroconversion" and "labor"; "figures" and "info"), the content-based ones were often times more broad and returned a larger number of results from Google.

Table 1. Examples of social-based and content-based query expansions with the number of results returned by Google.

Original Search	Social-Based Expansion	Content-Based Expansion
HIV testing during	HIV testing during seroconversion (122,000 results)	HIV testing during labor (2,200,000 results)
Obesity and blood pressure	Obesity and blood pressure cuff (1,050,000 results)	Obesity and blood pressure baby (15,900,000 results)
Diabetes Facts and	Diabetes facts and figures (1,690,000 results)	Diabetes facts and info (21,900,000 results)

Study Design

Stimuli

Medical Queries: Queries were obtained from the AOL 500k User Session Collection, which is a set of approximately 20M queries from 650K AOL users gathered during a three-month period in 2006. We selected queries that matched terms from the UMLS that were associated with the following UMLS semantic types: antibiotic, disease or syndrome, injury or poisoning, or body substance. We choose these semantic types because they are related to top medical concerns as described on Health.gov. The fourth gram of a four-gram was used to expand the base three-gram queries for content-based For each query, we collected up to ten expansion terms using both methods. This results in a total of 346,145 expanded queries. However, since each query is submitted to Google for the results evaluation and to complete the project in a reasonable time frame, we extracted a random 1% sample

for the analysis. This provided 1,731 expanded searches for social-based expansions, and 1,731 expanded searches for corpus-based expansions.

Variables

There was one independent variable used in the study: query expansion type. All medical queries were expanded in two fashions: using social query expansion and using content query expansion.

We evaluated the effect of expansion on the query itself using *specificity* and *ambiguity* and on the results by using *domain diversity* and *expansion edit distance*. These metrics were chosen because they provide an objective, if not complete, evaluation of queries and because they can be used for large sets of queries in an automated fashion. The first two dependent variables (ambiguity and specificity) help evaluate the effect of expansion on the queries. They are of interest because they measure how easily understood a word might be. Words that are more ambiguous have more possible meanings and would be expected to lead to more diverse results. Words that are more specific are further down a conceptual hierarchy (i.e. German Shepard is more specific than dog) and would be expected to result in a more focused results set (even though that set of results can be large). Both are calculated using WordNet 3.0 through the NLTK for Python 2.7.

Ambiguity was calculated as the number of synsets for a specific term, or how many different concepts that word can have. For example, the word "dog" has eight synsets or cognitive synonyms in WordNet, while the word "beagle" only has 1. This is an indication of how many different concepts each word might have, suggesting that "dog" is more ambiguous than "beagle."

Specificity was as the depth in the hypernym path of the first synset for each word. In this case, "dog" is found at the ninth level in WordNet's tree (top term is simply: "entity"), while "beagle" has 12, indicating that "beagle" is a more specific term conceptually than "dog."

The effect of expansion on the results themselves was measured by the *diversity* of domains in the results. Similarly, this measure was chosen because it is a straightforward way to evaluate the number of different domains returned by a search. This was measured using the Google Search API, which returns results as a JSON object given a query. Results returned show the base URL of each result as seen in Table 2.

Table 2. Example of the top three domains returned by a content expanded query and social expanded query.

Query	Top Three Domains
HIV testing during seroconversion (Social-based expansion)	www.healthline.com www.aidsmap.com www.jhsph.edu
HIV testing during labor (Content-based expansion)	www.aids.gov www.cdc.gov www.jiasociety.org

Finally, the *expansion edit distance* was calculated to compare the effect of an expanded query to the original (unexpanded) query. This was done in both a ranked and unranked manner. In the unranked method, the order of results was not taken into account, only the domains themselves. Searches that returned one different result when the expansion was included compared to when it wasn't would have a distance of 1. For example, in the results above, if the query "HIV testing during seroconversion" had returned "www.thebody.com" instead of "www.jhsph.edu," but the other two returned domains had been the same, the unranked edit distance would be 1. For the ranked edit distance, order was taken into account. For example, if the query had this time returned had returned

"www.thebody.com" instead of "www.jhsph.edu," and "www.healthline.com" and "www.aidsmap.com" had been switched in order, the edit distance would be 2.

Results

Searching the 3462 expanded queries required approximately 38.2 hours to complete using a Python script with the Google Search API previously described.

Table 3 shows the details of expansion type on ambiguity and specificity of queries. We provided the numbers for the entire set, and for the subset that were used to evaluate the results. Since the original query did not change, only the expansion terms, these numbers were averages of the expansion terms only.

Overall, the social-based expansion terms were almost an entire level more specific in the WordNet 3.0 hierarchy. The difference in ambiguity was larger. On average, there was an extra meaning (sense) for content-based expansion terms versus social-based expansion terms. A similar trend is seen for the 1% sample, although the differences are smaller.

Table 3. Query Evaluation: Average specificity and ambiguity for different query expansion types.

Complete Set (N = 346,145)	Specificity	Ambiguity
Social-based (N = 277,319)	7.89	6.37
Content-based (N = 68,826)	7.01	7.68
1% Sample (N = 3462)		
Social-based (N = 1731)	7.16	7.37
Content-based (N = 1731)	7.10	7.50

This means that for each individual using this query expansion method, the social-based method adds a term that narrows the search term to a subcategory of the original search. These words also tend to have fewer different meanings.

Table 3 also shows the results of social-based and content-based expansion the results of the search. The number of unique domains did not vary depending upon the type of expansion used. For content-based expansion, an average of 6.73 (SD = 1.53) unique domains were returned, for social-based expansions an average of 6.72 (SD = 1.55). A Welch's t-test showed no significant difference in domain diversity for the two expansion types ($t = .04$, $p = 0.96$, $d < 0.01$).

Table 4 shows the expansion edit distance (steps needed to go from unexpanded to expanded result set) for the two expansion types. There was a statistically significant difference between ranked social-based and content-based expansion. The social-based expansion had a slightly higher edit distance (M = 6.67, SD = 1.57) than content-based expansion (M = 6.44, SD = 1.73) $t(3,460) = 4.07$, $p < .001$. This means that, on average, social-based expansion resulted in more different results before and after expansion than the results from the content-based expansion. It is important to note, however, that the mean difference was small (0.23), and may not be a meaningful difference although it is statistically significant. The difference between the unranked edit distance was also significant in the same direction, such that social-based queries had a slightly higher unranked edit distance (M = 4.61, SD = 2.00) than content-based queries (M = 4.38, SD = 2.20), $t(3,460) = 3.22$, $p < .01$. In other words, query expansions matters: expanding the queries by a single term caused four of the ten results to change, on average.

Table 4. Average edit distances for top ten results from content-based and social-based expanded queries.

	Ranked	Unranked
Content-based (n = 1731)	M=6.44 SD=1.73	M=4.38 SD=2.20
Social-based (n = 1731)	M=6.67 SD=1.59	M=4.61 SD=2.00

Discussion

Two main findings stand out from this study. First, there is an interesting difference in specificity and ambiguity in the content-based expansion and social-based expansion. Terms generated through social-based expansion tended to be less ambiguous (have fewer different meanings) and more specific (further down the concept hierarchy) than terms created through content-based expansion. Since social-based expansion takes into account the popularity of searched phrases, it is likely that people use much more specific terms when they are searching information than actually exist in the text online. Information that exists on the web is written to express a certain point or relationship an author wants to express.

Secondly, although the findings above show that the two types of expansions resulted in different types of terms on average, there was almost zero difference in diversity of domains returned. However, since the starting queries in the 1% were less different based on expansion method, the results have most likely been affected. There is the possibility that Google purposely places a similar number of domains for every query in order to keep results diverse. In addition, a limitation of our work is that we could not directly compare the domains returned by social-based to content-based queries. On an individual query basis there was no good way to pair queries for a matched comparison. That is, a social-based expansion "hiv during seroconversion" is asking for different, more specific information than "hiv during pregnancy," and would obviously return different domains. Because of this, we only were able to compare the *number* of different domains returned by each expansion type, on average.

Edit distances provided some interesting information. Although the small difference in edit distance between the two expansion types was statistically significant, the raw difference was small. It is worth noting, however, that the addition of the terms from their base query did change results by quite a bit. Not only was the order of existing results changed, but nearly half of the results completely changed depending upon whether the query was expanded or not. At the very least, it can be argued that expanding does matter, but the method of expansion may not matter as much. It is unclear how much the expansion would affect user behavior, and subjective evaluation from human raters is necessary to further investigate the results here. Furthermore, there may be some impact of Google's AdWords on results in a real user setting. Certain search results will actually be from paid advertisements that are a function of the keywords searched. These were not collected in this study, but would likely have some influence on actual search behavior nonetheless. It may be hypothesized that users form subsequent queries based on the text of initial results. As such, paid advertisements within the results may influence what is being searched.

Overall, the small size of these effects is surprising, but there is the possibility that there may still be an impact on the diversity of information people read. More studies are needed evaluate the effects of the current query expansion on health information consumption. A further limitation is that this study represents only the effects on one query followed by an expanded query. The effect may be much more pronounced when 3 or 4 searches on one topic (as would be common for someone searching online) are combined. An engaged user is very likely to have interest in a topic beyond a single query. The effect could be such that results end up converging, or diverge even further. The effect could be such that search results converge or diverge even further. The actual impact of the human behavioral process of searching is a complex question that requires further study. Current directions in this work involve exploring how an individual user forms multiple queries upon the same search topic.

A final limitation of this study was the sole use of objective measures, so there was no direct human interaction with the search process. Subjective evaluations from human raters will be added to fully evaluate the results and look at the quality of information and the ability of readers to form a comprehensive picture.

Conclusion

Queries can be expanded using content from the documents being searched to provide queries that are broader on average than terms used by popular social-based expansion methods. Both social-based and content-based methods of expansion do not lead to differences in diversity of information returned on the internet, but whether or not the results differ in other ways is unknown. It is possible that one of the two methods returns *better* results depending on different criteria. Further research is needed to explore this possibility.

Future work will include subjective ratings of results and expansion terms in addition to analysis of temporal trends and human interaction studies with concrete search tasks. Specifically, the effects of multiple queries on a search topic need to be investigated to provide a more complete picture of how different types of search expansion affect the results returned. It is unlikely that an interested individual would conduct a search on a topic using a single query. In addition to research on the process of searching, evaluation of the search results needs to be refined and measured using broader criteria, such as subjective ratings of the quality and diversity of webpages. Ultimately, human raters and searchers will be a crucial part of next steps in this research, including the formation of queries and multiple search process, as well as the subjective evaluation of search results. Subjective ratings will give insight into human understanding and initial response to results returned by the different expansion types. Exploring results from search engines (i.e. Bing or DuckDuckGo) may also provide different results, but it is unclear what direction those results might go. Ultimately, more interaction from human raters is needed to draw strong conclusions about the different expansion types.

References

1. Fox, S. Health topics. Pew Research Internet Project. February 2011. [online] Available: http://www.pewinternet.org/2011/02/01/health-topics-2/
2. Hasan L, Abuelrub E. Assessing the quality of web sites. Applied Computing and Informatics. 2011; 9.1 11-29.
3. Weitzel L, Quaresma P, de Oliveira JPM. Evaluating quality of health information sources. Advanced Information Networking and Applications (AINA), 2012.
4. Wang Y, Zhenkai L. Automatic detecting indicators for quality of health information on the Web. International Journal of Medical Informatics. 2007; 76.8: 575-582.
5. Maples P, Franks A, Ray S, Stevens A, Wallace L. Development and validation of a low-literacy Chronic Obstructive Pulmonary Disease knowledge Questionnaire (COPD-Q). Patient Education and Counseling 2009; 81:19-22.
6. Weiss BD. Health literacy and patient aafety: Help patients understand (manual for clinicians): American Medical Assocation; 2007.
7. Leroy G, Endicott JE. Term familiarity to indicate perceived and actual difficulty of text in medical digital libraries. International Conference on Asia-Pacific Digital Libraries (ICADL 2011). 2011.
8. Leroy G, Endicott JE. Combining NLP with evidence-based methods to find text metrics related to perceived and actual text difficulty. In: 2nd ACM SIGHIT International Health Informatics Symposium (ACM IHI 2012); 2012.
9. Leroy G, Endicott JE, Mouradi O, Kauchak D, Just M. Improving perceived and actual text difficulty for health information consumers using semi-automated methods. In: American Medical Informatics Association (AMIA) Fall Symposium. 2012.
10. Leroy G, Endicott JE, Kauchak D, Mouradi O, Just M. User evaluation of the effects of a text simplification algorithm using term familiarity on perception, understanding, learning and information retention. Journal of Medical Internet Research (JMIR) 2013;15:e144.
11. Mouradi O, Leroy G, Kauchak D, Endicott JE. Influence of text and participant characteristics on perceived and actual text difficulty. Hawaii International Conference on System Sciences. 2013.
12. Leroy G, Kauchak D, Mouradi O. A user-study measuring the effects of lexical simplification and coherence enhancement on perceived and actual text difficulty. Int J Med Inform 2013; 82:717-30.
13. Sullivan D. How Google's instant autocomplete suggestions work. Search Engine Land. April 2011. [online]. Available: http://searchengineland.com/how-google-instant-autocomplete-suggestions-work-62592
14. Health Fact Sheet. Pew Research Internet Project. December 2013. [online]. Available: http://www.pewinternet.org/fact-sheets/health-fact-sheet/
15. Bhogal J, Macfarlane A, Smith P. A review of ontology based query expansion. Information Processing and Management. 2007; 43:866-886.
16. White RW, Horvitz E. Cyberchondria: studies of the escalation of medical concerns in web search. ACM Transactions on Information Systems (TOIS). 2008; 27: 23.
17. Liu Z, Chu WW. Knowledge-based query expansion to support scenario-specific retrieval of medical free text. Information Retrieval. 2007; 10:173-202.
18. Hersh W, Price S, Donohoe L. Assessing thesaurus-based query expansion using the UMLS metathesaurus. Proceedings of AMIA. 2000.
19. Chand S. Google autocomplete API. Sheryas Chand. January 2013. http://shreyaschand.com/blog/2013/01/03/google-autocomplete-api/
20. Franz A, Brants T, Norvig P. All our n-gram are belong to you | research blog. Research blog: the latest news from research at Google. August 2006; [online]. Available: http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html

ARX - A Comprehensive Tool for Anonymizing Biomedical Data

Fabian Prasser[12], Florian Kohlmayer[12], Ronald Lautenschläger[1], Klaus A. Kuhn[1]
[1] Technische Universität München, München, Germany, [2] Equal contributors

Abstract

Collaboration and data sharing have become core elements of biomedical research. Especially when sensitive data from distributed sources are linked, privacy threats have to be considered. Statistical disclosure control allows the protection of sensitive data by introducing fuzziness. Reduction of data quality, however, needs to be balanced against gains in protection. Therefore, tools are needed which provide a good overview of the anonymization process to those responsible for data sharing. These tools require graphical interfaces and the use of intuitive and replicable methods. In addition, extensive testing, documentation and openness to reviews by the community are important. Existing publicly available software is limited in functionality, and often active support is lacking. We present ARX, an anonymization tool that i) implements a wide variety of privacy methods in a highly efficient manner, ii) provides an intuitive cross-platform graphical interface, iii) offers a programming interface for integration into other software systems, and iv) is well documented and actively supported.

Introduction

Collaboration and data sharing have become core elements of biomedical research; see, e.g., the joint statement on sharing research data[1] and the guidelines for access to research data of the *Organization for Economic Co-operation and Development* (OECD)[2]. For clinical data, and increasingly for genetic data as well as for combinations of these data, there is a growing understanding of related privacy threats. Disclosure of data may lead to harm for individuals, especially when different data sources are available for linkage[3]. From the legal perspective, national laws, e.g., the *Health Insurance Portability and Accountability Act* (HIPAA)[4] *Privacy Rule*, and international regulations, e.g., the *European Directive on Data Protection*[5], mandate stringent protection of personal data.

The HIPAA Privacy Rule[6] defines two basic methods for de-identifying datasets. The first requires the removal of a pre-defined set of attributes from the dataset. This procedure significantly reduces re-identification risks[7], but it can *1)* obstruct data use if the involved attributes are essential[8], and *2)* under certain circumstances not prevent re-identification[9]. The second method is "expert determination": A professional "determines that the risk is very small that the information could be used […] to identify an individual"[6]. In this context, *statistical disclosure control* (SDC) allows balancing privacy risks with data quality[8]. Examples include methods focusing on data extracts, such as *differential privacy*[10], and methods for microdata release, such as *k-anonymity*[11].

In the biomedical domain, methods for microdata release are currently preferred[12]. The primary reason for this is that they can be implemented with non-pertubative methods that preserve the truthfulness of data[12]. They have been included in guidelines of best-practices for de-identifying health data[13], and they have been successfully applied to research data[14]. Moreover, many approaches for microdata release have been developed specifically for the biomedical domain[15,16,17].

Objectives

Although anonymization is an important method for privacy protection, there is a lack of tools which are both comprehensive and readily available to informatics researchers and also to non-IT experts, e.g., researchers responsible for the sharing of data. As protection has to be balanced against losses in data utility, responsible researchers should be able to keep an overview of the anonymization process and the trade-offs chosen. This requires powerful but easy to use tools which can be integrated in research workflows. *Graphical user interfaces* (GUIs) and the option of using a wide variety of intuitive and replicable methods are needed. Tools have to offer interfaces allowing their integration into pipelines comprising further data processing modules. Moreover, extensive testing, documentation and openness to reviews by the community are of high importance. Informatics researchers who want to use or evaluate existing anonymization methods or to develop novel methods will benefit from well-documented, open-source software libraries. The lack of such a framework is illustrated by the fact that, although data anonymization has been researched for a long period of time already, only recently efforts have started to systematically evaluate and compare existing methods[18,19].

The landscape of existing tools is heterogeneous. *PARAT*[20] is the leading commercial de-identification software. It is a closed-source tool for which only limited information is available to the public. We will focus on

non-commercial tools in the remainder of this article. The *UTD Anonymization Toolbox*[21] and the *Cornell Anonymization Toolkit* (CAT)[22] are research prototypes that have mainly been developed for demonstration purposes. Problems with these tools include scalability issues when handling large datasets, complex configuration requiring IT-expertise, and incomplete support of privacy criteria and methods of data transformation. *sdcMirco*[23] is a package for the *R* statistics software, which implements many primitives required for data anonymization but offers only a limited support for using them to find data transformations that are suitable for a specific context. *μ-Argus*[24] is a closed-source application that implements a broad spectrum of techniques, but it is no longer under active development. A comparison of our work with these tools is presented in the *Discussion* section.

To overcome these limitations we present ARX, a comprehensive open-source data anonymization framework that implements a simple three-step process. It provides support for all common privacy criteria, as well as for arbitrary combinations. It utilizes a well-known and highly efficient anonymization algorithm. Moreover, it implements a carefully chosen set of techniques that can handle a broad spectrum of data anonymization tasks, while being efficient, intuitive and easy to understand. Our tool features a cross-platform user interface that is oriented towards non-IT experts. Additionally, it provides a stand-alone software library with an easy-to-use public *application programming interface* (API) for integration into other systems. Our code base is extensible, well-tested and extensively documented. As such, it provides a solid basis for developing novel privacy methods.

Background and Terminology

In the context of SDC, data is organized in a tabular structure, where each *record* represents the data about one individual. Basically, *identifiers* within a dataset are modified in such a way that linkage is prevented[25]: Firstly, highly distinguishing attributes that can be used for re-identification and that are not required for analyses are removed. Examples for such *direct identifiers* include *Social Security numbers* or names. Secondly, *quasi-identifiers*, i.e., attributes which are required for analyses while being potentially identifying are *recoded* to make sure that the data fulfills well-known *privacy criteria*[25]. Examples include the age or sex of data subjects. Data recoding is typically performed with *generalization hierarchies*[11], which can be built for categorical and continuous attributes (see Figure 1). To increase the utility of resulting datasets, this method is often combined with *tuple suppression*: data records that violate the privacy criteria (so called *outliers*) are automatically removed from the dataset, while the total number of suppressed records is kept under a given threshold[11]. As a result, less generalization is required to ensure that the remaining records fulfill the privacy criteria.

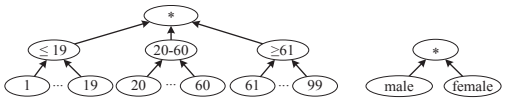


Figure 1. Hierarchies for attributes "*age*" and "*sex*"

k-Anonymity[11] is the most wide-spread privacy criterion. It ensures that each data record cannot be distinguished from at least k-1 other data records regarding the quasi-identifiers. It aims at protecting datasets against *identity disclosure*, i.e., from linking an individual to a specific data record by an attacker[26]. *ℓ-Diversity*[27] and *t-closeness*[28] aim at protecting datasets against *attribute disclosure*, where an attacker can infer information about an individual without necessarily linking it to a specific record in the dataset[26]. As an example, linkage to a set of records allows inferring information if all records share a certain attribute value. The attributes in which an attacker might be interested and which, if disclosed, could cause harm to the data subject are called *sensitive attributes*. Different variants exist for both criteria, which offer different privacy guarantees by enforcing different properties on the distributions of sensitive attribute values within a set of indistinguishable data records. *δ-Presence*[29] aims at protecting datasets against *membership disclosure*, which means that an attacker is able to determine whether or not data about an individual is contained in a dataset[26]. For an overview of further privacy criteria we refer to the work by Fung et al.[25].

Different types of algorithms can be used to transform datasets so that they fulfill a given set of privacy criteria. For the biomedical domain, the use of *globally-optimal full-domain anonymization* algorithms using *multi-dimensional global recoding* has been recommended[30]. These algorithms construct a search space, which consists of all possible combinations of generalization levels for all quasi-identifying attributes. This space of possible generalizations is then traversed to find a transformation that fulfills all privacy criteria while resulting in optimal data quality. To this end, utility is measured with *metrics* for information loss[25].

Methods

Our work aims at making data anonymization available to a wide variety of end users. We therefore decided to implement a type of algorithm that is intuitive to non-IT experts: a globally-optimal full-domain anonymization algorithm that uses multi-dimensional global recoding[17]. On the upside, such algorithms implement anonymization

procedures that can easily be configured by users, e.g., by altering generalization hierarchies or choosing a suitable transformation from the solution space, and result in datasets that are well suited for biomedical analyses[30]. On the downside, the underlying coding model is strict and potentially results in low data utility. To attenuate this, our framework combines the method with local tuple suppression. It increases data quality, but can also significantly increase execution times. Consequently, our tool is also able to approximate the result in much less time. Approximated results are guaranteed to fulfill the given criteria but might not be optimal in terms of data quality, because only some transformations in the solution space are classified with absolute certainty.

In order to cover a broad spectrum of privacy problems, our tool comes with implementations of all commonly used privacy methods: k-anonymity, all variants of ℓ-diversity, two variants of t-closeness, and δ-presence. A querying interface allows selecting a research subset, which is a subset of the data records that are to be included in the final anonymized dataset. Moreover, δ-presence can be enforced on the subset, i.e., attackers that know the overall dataset can be prevented from determining whether a specific tuple is included in the subset. In contrast to previous approaches our tool is the first to support classifying the complete solution space for arbitrary combinations of privacy criteria while using generalization *and* suppression. For assessing data utility, we included a large set of metrics for information loss, including simple methods, such as *height* and *precision* as well as more sophisticated approaches, such as *discernibility*[31] and *non-uniform entropy*[30].

Data and generalization hierarchies can be imported from many different types of sources in order to provide compatibility with a wide range of data processing tools. ARX currently features interfaces for importing data from *character-separated values* (CSV) files, *MS Excel* spreadsheets and *relational database management systems* (RDBMSs), such as *PostgreSQL* or *MySQL*. Data imported into ARX is immutable and cannot be changed. ARX implements several methods that can be used for identifying data quality issues: *1)* data can be sorted, compared and analyzed regarding statistical properties, and *2)* the query interface can be used to search for records with quality issues. For handling such issues, ARX supports the automated and manual removal of records. If more complex data cleansing tasks must be performed, data can be exported to other software systems. ARX safely handles missing values by treating them similar to all other values (i.e., missing values match other missing values)[32]. This scheme allows for truthful and non-truthful handling of missing values, depending on how they are generalized, without introducing privacy issues[32].

ARX offers methods for manual and semi-automatic creation of generalization hierarchies. Semi-automatic creation is supported for all common types of attributes, such as numerical (discrete or continuous) and categorical variables. Hierarchies can be generated by grouping values based on a natural or user-defined ordering, by mapping them into user-defined or automatically created intervals or by data redaction. Hierarchies are represented in a tabular manner, which is an intuitive representation that enables compatibility with third-party applications, such as spreadsheet programs.

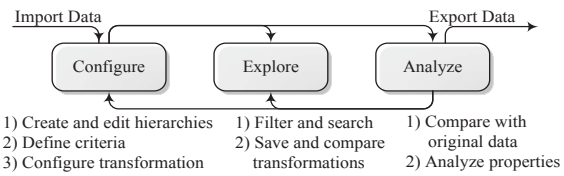


Figure 2. Implemented de-identification workflow

ARX supports all major aspects of data anonymization. These different aspects are combined in a multi-step process that allows users to iteratively adjust parameters, until the result matches their requirements. As depicted in Figure 2, the basic steps consist of *1)* configuring the anonymization process, *2)* exploring the solution space and *3)* analyzing the transformed data. In the configuration phase, the input data is loaded, generalization hierarchies are imported or created and all further parameters, such as privacy criteria, are specified. When the solution space has been characterized by executing the anonymization algorithm, the exploration phase allows searching the solution space for privacy-preserving data transformations that fulfill the user's requirements. To assess suitability, the analysis phase allows comparing transformed datasets to the original input dataset. Based on this analysis, further solution candidates might be considered and analyzed, or the configuration of the anonymization process might be altered. Our tool features a cross-platform graphical interface for non-IT experts. All methods implemented in ARX are accessible via the API and the GUI.

An important goal of our efforts is to make the anonymization framework, consisting of the graphical application and the software library, available to software developers and informatics researchers. To this end, we chose to implement it in the *Java* ecosystem, which offers one of the most popular cross-platform development environments. We chose to implement the GUI with the *Standard Widget Toolkit* (SWT) and the *JFace* library, which provide support for the native *look and feels* of the three most common platforms: *OS X*, *Windows* and

Linux/GTK. We spent extensive time on documenting our code as well as the public API and on adding a large set of examples to our code base. Our project has a high test coverage, featuring hundreds of unit tests with different configurations and a broad spectrum of input datasets, including a large set of tests for which the results were validated manually. Our website provides users with background information and extensive documentation[33].

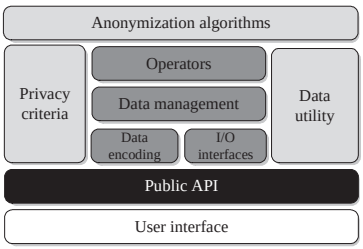


Figure 3. High-level architecture

A high-level overview of the architecture of ARX is shown in Figure 3. We carefully designed the framework, to prevent tight coupling of subsystems and ensure extensibility. The core modules (dark gray), which normally need not be modified, provide the basis for our framework. The *I/O module* provides methods for reading data from and writing data to external storage, while the *data encoding module* transforms the data into the format and memory layout required by the framework. The *data management module* deals with this internal representation and implements several optimizations (e.g., caching). Problem-specific *operators* are built on top of this representation and allow grouping of data records and computing of frequency distributions over sensitive attribute values. This provides the basis for extensible modules (light gray), which implement *privacy criteria* and metrics for measuring *data utility*. Analogously, anonymization algorithms can be plugged into the framework. Currently our tool features several variants of the *Flash* algorithm but the framework can be used to implement a large set of methods[18]. The public API is based on both the extensible and the core modules. It is also used by the graphical interface, which is completely decoupled from the internals of the framework.

Our three-step process poses considerable challenges in terms of efficiency. Firstly, ARX automatically classifies the complete solution space to support users in finding a transformation suitable for their application scenario. Secondly, the iterative character of the process potentially requires this classification to be performed repeatedly. It is thus very important that classification can be carried out in near real-time. For this purpose, our framework features a highly efficient algorithm[17]. Moreover, instead of using existing database systems, we implemented a runtime environment that is tailored to the problem domain. In previous work we have shown that our method significantly outperforms comparable algorithms within our highly optimized framework[18].

Results

In this section, we present an overview of the graphical interface for end-users which illustrates ARX's functionality. We then address the public API for researchers and software developers. Finally, we shortly analyze the scalability of our tool in terms of execution time and memory requirements.

Importing data and configuring the de-identification process

The graphical interface of ARX is divided into three perspectives that follow the workflow outlined in the previous section. In the first perspective, a dataset can be imported and the anonymization process can be configured.

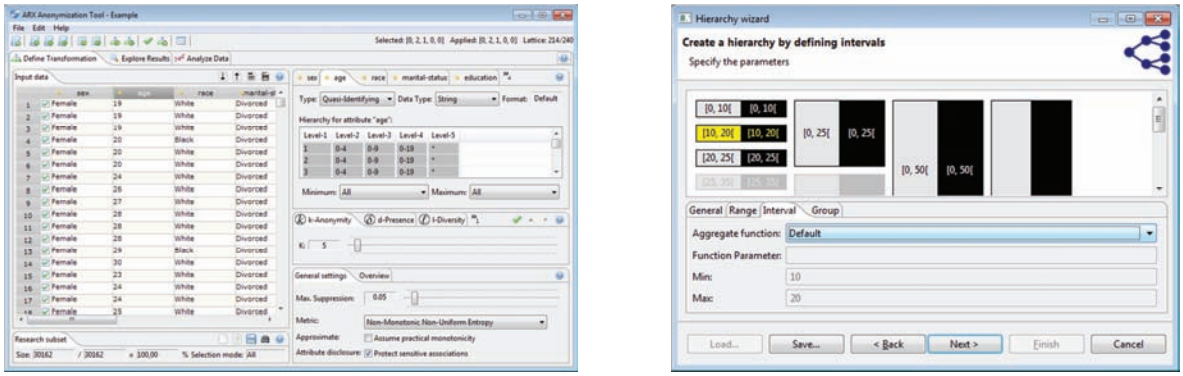


Figure 4. Interfaces for configuring the de-identification process and creating generalization hierarchies

A screenshot is presented in Figure 4. On the left-hand side, the imported dataset is displayed. Individual records can be included or excluded via checkboxes to define a research subset. An overview of the current subset is presented in the lower left area of the perspective. The header of the table showing the dataset also indicates the attribute type associated to each column in terms of a color scheme. These attribute types can be defined in the upper right area of the perspective: directly identifying attributes are removed from the dataset, quasi-identifiers are transformed by

applying the provided hierarchies, while sensitive attributes are not transformed but can be used to derive t-close or ℓ-diverse transformations. *Insensitive* attributes are simply kept unchanged. The area also displays a tabular representation of the generalization hierarchies associated to the attributes. Hierarchies can be edited manually, e.g., by adding and removing generalization levels or altering labels. The lower right area of the perspective allows for defining privacy criteria and for configuring the transformation and classification process. Important aspects include selecting the metric that is to be used for assessing data utility and defining an upper bound on the number of records that can be suppressed. When all parameters have been configured, the solution space can be classified.

ARX offers several wizards for semi-automatically creating generalization hierarchies. The wizard for interval-based hierarchies is shown on the right side of Figure 4. Intervals are a natural means of generalization for variables with a ratio scale, such as integers, decimals or dates and timestamps. Each level of the hierarchy is represented by one column, with the lowest level being defined by a sequence of intervals (leftmost column). Subsequent levels can be added by grouping any given number of elements from the previous level. Any sequence of intervals or groups is automatically repeated to cover a predefined range. For values outside of this range, *top and bottom coding* can be applied. Automatically created repetitions of intervals and groups are indicated visually.

Exploring the solution space

When the solution space has been classified, the exploration perspective allows users to browse the set of all possible transformations. The aim of the perspective, a screenshot of which is presented on the left-hand side in Figure 5, is to select a set of interesting transformations for analysis.

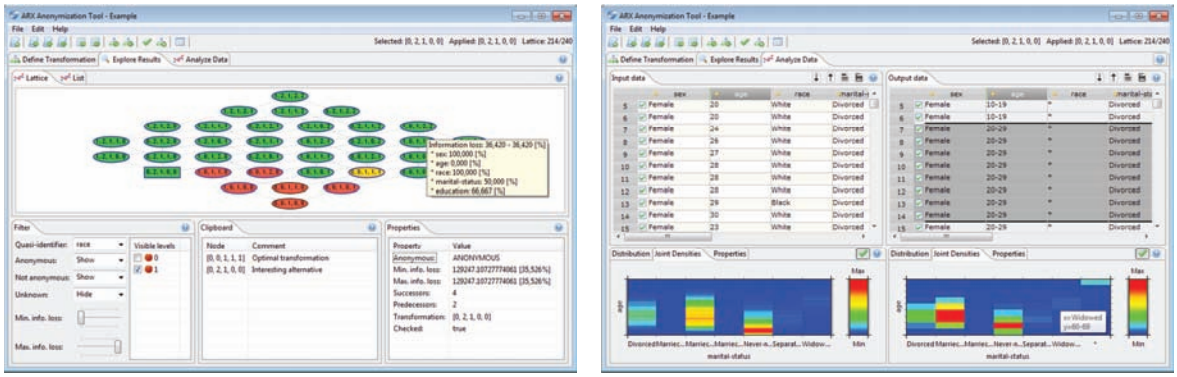


Figure 5. Interfaces for exploring the solution space and analyzing the results of data transformations

In the center of the screen, the view displays a subset of the solution space. Each node represents one transformation, which is identified by the generalization levels that it defines for the quasi-identifiers in the input dataset. Transformations are characterized by four different background colors: green denotes that the transformation results in an anonymous dataset, red denotes that the transformation does not result in an anonymous dataset, orange denotes that the transformation is the global optimum and gray denotes that the anonymity property of the transformations is unknown (only applies when approximating the result). The view supports zooming and moving the visualization of the solution space. Interesting transformations can be added to a clipboard, where they can be organized. Applying a transformation to the dataset allows exporting the resulting dataset or analyzing it in the view presented in the next section. A filter allows selecting a subset of the solution space by defining that only transformations with certain generalization levels, certain anonymity properties or with their information loss being within a defined interval should be visible. In case of an approximated result, this also includes transformations which are probably anonymous or probably non-anonymous. If such a transformation is applied, its actual anonymity property will be computed in the background and the state of the solution space will be updated.

Analyzing transformed data

A given privacy problem can often be solved with several different transformations. Although ARX is able to automatically find a solution which is optimal regarding the selected metric for data utility, automatically choosing an appropriate solution for a given usage scenario is often difficult. The aim of this perspective, which is shown on the right-hand side in Figure 5, therefore is to support users in assessing the utility of a transformed dataset for a specific application scenario. For this purpose, it allows comparing transformed data to the original dataset.

The perspective displays the input dataset on the left and the output dataset on the right. The tables are synchronized when scrolling, allowing an easy comparison of different parts of the data. The data can be sorted according to selected attributes. It is possible to toggle between showing the whole dataset or only the research subset. Moreover, the resulting equivalence classes can be highlighted. For comparing the statistical properties of two data representations, the view also shows the frequency distributions of values of a selected attribute in both tables. Moreover, a graphical representation of contingency tables for two selected attributes is included. This feature allows for visually comparing the combined occurrence of values from two different attributes. In the example, the number of values of the attribute "marital-status" remains the same, while the number of values of the attribute "age" is reduced from 100 to 10. Additionally, some values of the attribute "age" and exactly one value of the attribute "marital-status" have been suppressed. This is reflected by the heat map on the right-hand side, which shows a slightly shifted generalization of the contingency from the input dataset.

Programmatic access via the API

All features that are accessible via the graphical interface are also accessible via the public API. However, the aim of the programming interface is to provide de-identification methods to other software systems and we note that interaction with the software library provided by ARX will often be simpler than interaction with the graphical tool. Programmatic access will usually rely on ARX's ability to automatically determine a solution to a privacy problem.

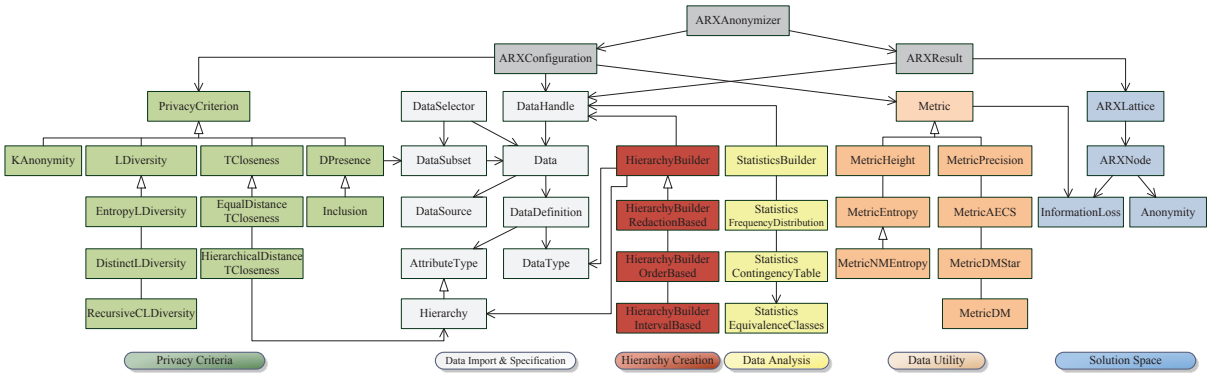


Figure 6. UML diagram of the most important classes in the public API

A *Unified Modeling Language* (UML) diagram of the most important classes of our API is shown in Figure 6. It can be seen that the API consists of a set of packages for *1) data import and specification*, *2) hierarchy generation*, *3) privacy criteria*, *4)* measuring *data utility*, *5) data analysis*, and *6)* representing the *solution space*. The classes `ARXConfiguration`, `ARXAnonymizer` and `ARXResult` provide the main interfaces to ARX's functionalities.

Data and attribute types are provided as static variables from the `DataType` and `AttributeType` classes respectively. The class `DataHandle` allows users to interact with the data (read-only), by performing operations such as sorting, swapping rows or reading cell values. Handles can be obtained for input data, output data and research subsets of such data. Data handles representing input and derived output data are linked with each other, meaning that operations performed on one representation are transparently performed on all other representations as well. For example, sorting the input data sorts the output data analogously. The class `ARXLattice` offers several methods for exploring the solution space and for obtaining information about the properties of transformations

```
/* Load data from SQLite database*/
DataSource source = DataSource.createJDBCSource("jdbc:sqlite:test.db", "test");
source.addColumn("zipcode", DataType.STRING);
source.addColumn("age", DataType.INTEGER);
source.addColumn("diagnosis", DataType.STRING);
Data data = Data.create(source);

/* Create hierarchy with redaction*/
data.getDefinition().setAttributeType("zipcode", HierarchyBuilderRedactionBased.create(Order.RIGHT_TO_LEFT, Order.RIGHT_TO_LEFT, ' ', '*'));

/* Load hierarchies*/
data.getDefinition().setAttributeType("age", Hierarchy.create("age.csv", ';'));

/* Define sensitive attribute*/
data.getDefinition().setAttributeType("diagnosis", AttributeType.SENSITIVE_ATTRIBUTE);
```

Figure 7. Importing data and creating hierarchies with the API

(represented by the class `ARXNode`). The class `Inclusion` implements a dummy criterion that can be used to exclude tuples from the input dataset by defining a research subset.

In the remainder of this section, we will present an example for deriving an optimally anonymized transformation from a dataset loaded from an RDBMS. The process of importing data and of loading as well as creating generalization hierarchies is outlined in Figure 7. Firstly, a `DataSource` is created, which encapsulates all information required to access a database as well as the schematic properties of the data to be imported. The dataset is loaded by creating an instance of the class `Data`. Secondly, a generalization hierarchy for the attribute "zipcode" is created automatically by applying redaction. Finally, attribute types are specified. Quasi-identifiers are defined by associating a hierarchy, which is loaded from a CSV file for the attribute "age".

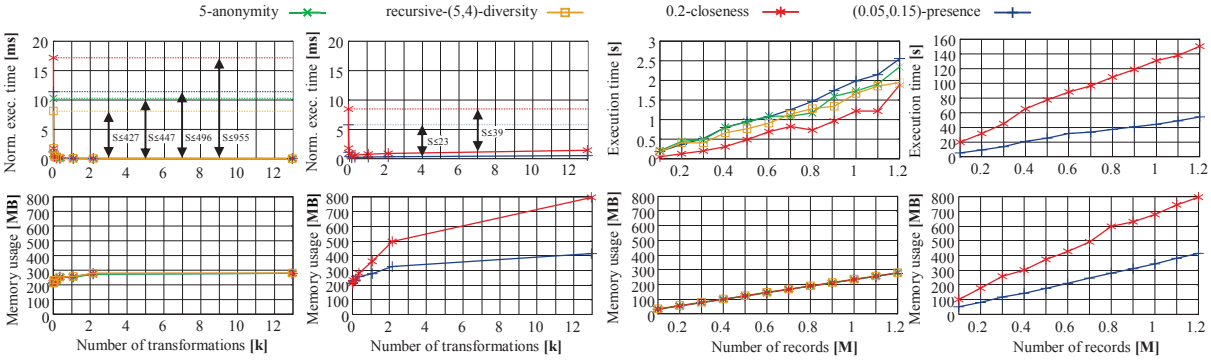
```
/* Configure the anonymization process*/
ARXConfiguration config = ARXConfiguration.create();
config.addCriterion(new KAnonymity(5));
config.addCriterion(new HierarchicalDistanceTCloseness("diagnosis", 0.6d, Hierarchy.create("diagnosis.csv", ';')));
config.setMaxOutliers(0d);
config.setMetric(Metric.createEntropyMetric());

/* Perform classification of the solution space*/
ARXAnonymizer anonymizer = new ARXAnonymizer();
ARXResult result = anonymizer.anonymize(data, config);

/* Write result of applying the optimal transformation*/
result.getOutput(result.getGlobalOptimum()).save("output.csv", ';');
```

Figure 8. Anonymizing and exporting data with the API

In Figure 8, the process of defining privacy requirements, configuring the transformation process, classifying the solution space, applying the globally-optimal transformation to the dataset and writing the result to a CSV file is sketched. The example configuration features 5-anonymity and 0.6-closeness for the sensitive attribute "diagnosis". The distance between distributions of the sensitive attribute is computed using a generalization hierarchy[28]. Tuple suppression is disabled and information loss is measured with non-uniform entropy[30].

Scalability of ARX

As already highlighted in the previous section, it is crucial that the process of classifying the search space is highly efficient. Most optimal anonymization algorithms are built on the assumption of monotonicity, meaning that generalizations of anonymous datasets are also anonymous and that specializations of non-anonymous datasets are also non-anonymous. This is, e.g., not true for ℓ-diversity and t-closeness combined with tuple suppression. ARX is the first tool to support such configurations, too, thus offering full support for tuple suppression. Our algorithm dynamically leverages any form of monotonicity. It can already prune parts of the search space if only some privacy criteria from a combination of several criteria are monotonic or if only the metric for assessing data utility is monotonic. In the context of utility metrics, monotonicity means that data quality decreases monotonically with generalization. This, again, is not always the case when tuple suppression is utilized.



(a) Scalability for monotonic configurations depending on the size of the search space

(b) Scalability for non-monotonic configurations depending on the size of the search space

(c) Scalability for monotonic configurations depending on the number of records

(d) Scalability for non-monotonic configurations depending on the number of records

Figure 9. Scalability of ARX for the *Integrated Health Interview Series* (IHIS) dataset

For our experiments we used a dataset from the *Integrated Health Interview Series* (IHIS)[34]. We have chosen this dataset, as it is publicly available and large, containing records about nearly 1.2 million individuals. From its

attributes, we have chosen eight quasi-identifiers (QIs) and one sensitive attribute[18], resulting in a solution space consisting of 12,960 transformations. To assess scalability, we evaluated our tool with two different types of privacy problems: scenarios with 0% suppression in which criteria and metrics are monotonic (leading to best-case performance) and scenarios with 5% suppression in which neither criteria nor metrics are monotonic (resulting in worst-case performance). Additionally, we scaled the size of the dataset and the size of the solution space. The latter has been realized by incrementally increasing the number of QIs. In this case it has to be considered that adding a QI to a dataset multiplies the size of the solution space by the number of levels in the associated hierarchy and leads to the need to transform one additional column. To exclude this effect and derive comparable numbers for different scales, we report a *normalized execution time*, which is defined as the execution time divided by the number of transformations in the solution space and divided by the number of QIs.

The results are summarized in Figure 9. The normalized execution times for the monotonic case show that the optimizations implemented in ARX (including pruning based on monotonicity) lead to a speedup of up to a factor of 955 compared to the baseline execution time. The baseline is defined by the configuration with only one QI, as, in this case, almost no optimizations can be leveraged. In the non-monotonic scenario, the entire solution space must be searched. Still, ARX is able to leverage its optimized runtime environment to achieve speedups of up to a factor of 39 compared to its baseline performance. The factor between best-case execution times (~2.5s) and worst-case execution times (~150s) is roughly 60. The results of our experiments with increasing dataset size and a non-monotonic configuration show that our tool can evaluate 12,960 transformations in about 150s. This means that in an average of roughly 12ms, ARX is able *1)* to generalize and suppress values in a dataset consisting of 1.2 million records, with eight QIs each, *2)* to group the transformed records according to the QIs, *3)* to compute frequency distributions for the sensitive attribute values in each group, *4)* to check whether each frequency distribution fulfills t-closeness measured with the *Earth Mover's Distance* (EMD) based on a generalization hierarchy, and *5)* to compute the overall utility of the dataset in terms of non-uniform entropy. We note that our experiments were performed on cheap commodity hardware[18]. Performance increases further when no frequency distributions are required, e.g., for k-anonymity or δ-presence. In terms of memory requirements, our tool uses up to 300MB for monotonic and up to 800MB for non-monotonic configurations. We note that ARX implements a space-time trade-off that allows reducing its memory consumption, if required[17].

Discussion

Principal Results

We have presented a comprehensive open-source data anonymization framework. It is under active development, well-tested, and available for many platforms. We have given an overview of the core design of our system, the graphical user interface, and the public API for external software. Our tool implements a three-step data anonymization process, and it supports arbitrary combinations of privacy criteria and the use of tuple suppression for finding optimal transformations regarding t-closeness, ℓ-diversity or δ-presence.

ARX is still under development and constantly being updated with new features; we have just released version 2.2.0 which fully supports the anonymization workflow described in this paper. While this workflow and the implemented methods are motivated by usage scenarios of the biomedical domain, the tool can also handle other types of data. Through feedback from users and researchers we have learned that there is indeed a strong demand for data anonymization tools such as ARX. We constantly update our online documentation to provide answers to common questions and extend our tool to cover functionalities required by our users.

Tests and Experiences

We have successfully tested and evaluated the current version of ARX with multiple real-world benchmark datasets[18], including publicly available biomedical data, such as IHIS[34]. We plan to add a formal study with real-world use cases. From a scientific perspective, we have used our framework as a basis for several informatics research projects[17,18,35,36]. The hypothesis that our efforts have made data anonymization technologies available to a broader audience is supported by access statistics to our project website. In one year (2013/7-2014/7), our website had more than 2000 unique visitors, hundreds of whom have downloaded our tool. Visitors from India, the USA, Germany and Japan account for over 50% of all traffic.

Comparison to Prior Work

The UTD Anonymization Toolbox[21] supports three different privacy criteria (k-anonymity, ℓ-diversity and t-closeness) and uses a *SQLite* database backend. In our experiments, we encountered problems with larger datasets.

It further lacks a graphical interface and requires configuration to be performed via an XML file. It does not support combining tuple suppression with ℓ-diversity or t-closeness, which can lead to low data quality.

Table 1. Comparison to previous approaches

		UTD-AT	CAT	sdcMicro	μ-Argus	ARX
Developer Support	Open source	Yes	Yes	Yes	No	Yes
	Active	No	No	Yes	No	Yes
	Public API	No	No	Yes	No	Yes
	Extensibility	Low	Low	Low	No	High
	Cross-platform	Yes	Yes	Yes	No	Yes
	Prog. Language	Java	C++	R	C++	Java
Usability	GUI coverage	None	Full	Partial	Full	Full
	Hierarchy creation	No	No	Yes	No	Yes
	Visualization	No	Data, Risks	Data, Risks	Risks	Data, solution space
	Data sources	CSV	Proprietary	CSV, Various	CSV, Various	CSV, Excel, DBMS
	Hierarchy format	Proprietary	Proprietary	Proprietary	Proprietary	CSV
	Standalone	No	Yes	No	Yes	Yes
Anonymity Methods	Automatic solution	Yes	Yes	Partial	No	Yes
	Privacy criteria	k, ℓ, t	ℓ, t	k, ℓ	None	k, ℓ, t, δ
	Generalization	Yes	Yes	Yes	Yes	Yes
	Tuple suppression	Partial	No	Yes	Yes	Yes
	Risk assessment	No	Limited	Yes	Yes	Limited

The Cornell Anonymization Toolkit (CAT)[22] supports ℓ-diversity and t-closeness. It was developed for demonstration purposes and is no longer under active development. It lacks support for tuple suppression and requires input data to be in a tool-specific format. sdcMicro[23] is a package for the R statistics software and as such not meant to be a standalone application. It provides a graphical user interface, but only implements limited methods for automatically solving privacy problems or classifying the solution space. It supports k-anonymity and ℓ-diversity and is still being actively developed. μ-Argus[24] is a project which is no longer under active development. It is a closed-source Windows application that is not intended to act as a software library. It provides a broad spectrum of recoding techniques, including global recoding with local suppression as well as top and bottom coding and multiple methods for risk estimation. De-identification must be performed manually. SECRETA[19] is a tool that allows comparing different anonymization algorithms for relational and transactional data, but it is not available to the public. Many related tools include methods for evaluating re-identification risks, e.g., sdcMicro or μ-Argus. A simple but often used measure for the *prosecutor re-identification risk*[37] is the minimum, maximum and average size of equivalence classes, which is also available in our tool. If users need to use further risk assessment methods, data can be exported into other applications. The results of our comparison have been summarized in Table 1.

Future Work

In future work, we plan to enhance ARX with several additional features. We already implemented multiple risk estimators, e.g., the approach by Dankar et al.[38], but integrating the results into our tool will require further work. Additionally, we plan to combine our method with less restrictive coding models (e.g., local recoding) and to provide a set of typically needed hierarchies. We are also actively working on support for transactional attributes as well as methods for secure continuous data publishing. Currently, ARX focusses on methods for privacy-preserving release of microdata. In future work, we plan to integrate non-interactive variants of differential-privacy, which provide provable privacy guarantees that are independent of an attacker's background knowledge[12].

References

1. Wellcome Trust. Sharing research data to improve public health. 2013. Available from: http://wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Public-health-and-epidemiology/WTDV030690.htm
2. OECD. Principles and guidelines for access to research data from public funding. 2006. Available from: www.oecd.org/sti/sci-tech/38500813.pdf.
3. Heeney C, Hawkins N, de Vries J, Boddington P, Kaye J. Assessing the privacy risks of data sharing in genomics. Public Health Genomics. 2011;14(1):17-25.
4. United States Congress. Health insurance portability and accountability act of 1996. Public Law. 1996:1-349.
5. Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data. Off J Eur Communities. 1995;38(L. 281).
6. U.S. Department of Health and Human Services. Office for Civil Rights. HIPAA Administrative Simplification Regulation, 45 CFR Parts 160, 162, and 164; 2013.

7. Benitez K, Malin B. Evaluating re-identification risks with respect to the HIPAA Privacy Rule. J Am Med Inform Assoc. 2010;17(2):169-77.
8. Malin B, Benitez K, Masys D. Never too old for anonymity: a statistical standard for demographic data sharing via the HIPAA Privacy Rule. J Am Med Inform Assoc. 2011;18(1):3-10.
9. Loukides G, Denny J, Malin B. The disclosure of diagnosis codes can breach research participants' privacy. J Am Med Inform Assoc. 2010;31:1-31.
10. Dwork C. Differential privacy. Proc Int Coll Automata, Languages and Programming. 2006;1-12.
11. Samarati P, Sweeney L. Protecting respondents identities in microdata release. IEEE Trans Knowl Data Eng. 2001;13(6):1010-1027.
12. Dankar F et al. Practicing differential privacy in health care: a review. Trans Data Priv. 2013;5:35-67.
13. Health System Use Technical Advisory Committee Data De-Identification Working Group. Best practice guidelines for managing the disclosure of de-identified health information. 2010. Available from: http://www.ehealthinformation.ca/documents/de-idguidelines.pdf.
14. El Emam K, Paton D, Dankar F, Koru G. De-identifying a public use microdata file from the Canadian national discharge abstract database. BMC Med Inform Decis Mak. 2011;11(1):53.
15. Benitez K, Loukides G, Malin B. Beyond safe harbor. Proc Int Conf Health Inform. 2010;163-172.
16. Ye H, Chen ES. Attribute utility motivated k-anonymization of datasets to support the heterogeneous needs of biomedical researchers. AMIA Annu Symp Proc. 2011;1573-82.
17. Kohlmayer F, Prasser F, Eckert C, Kemper A, Kuhn KA. Flash: efficient, stable and optimal k-anonymity. Proc Int Conf Privacy, Secur Risk Trust. 2012:708-717.
18. Prasser F, Kohlmayer F, Kuhn KA. A benchmark of globally-optimal anonymization methods for biomedical data. Proc Int Symp Computer-Based Medical Systems. 2014;66-71.
19. Poulis G, Aris GD, Grigorios L, Spiros S, Christos T. SECRETA: a system for evaluating and comparing relational and transaction anonymization algorithms. Proc Int Conf Ext Database Technology. 2014;620-623.
20. About PARAT De-Identification Software [cited 04 Aug 2014]. Privacy Analytics Inc. Available from: http://www.privacyanalytics.ca/software/parat/
21. UTD Anonymization Toolbox [cited 04 Aug 2014]. UT Dallas Data Security and Privacy Lab. Available from: http://cs.utdallas.edu/dspl/cgi-bin/toolbox/
22. Cornell Anonymization Toolkit [cited 04 Aug 2014]. Cornell Database Group. Available from: http://sourceforge.net/p/anony-toolkit/
23. sdcMicro [cited 04 Aug 2014]. Data-Analysis. Available from: http://cran.r-project.org/web/packages/sdcMicro/
24. μ-Argus manual. Available from: neon.vb.cbs.nl/casc/Software/MuManual4.2.pdf
25. Fung BCM, Wang K, Fu AWC, Yu PS. Introduction to privacy-preserving data publishing: concepts and techniques. 1st ed. Chapman and Hall/CRC; 2011:376.
26. Li T, Li N, Zhang J, Molloy I. Slicing: a new approach for privacy preserving data publishing. IEEE Trans Knowl Data Eng. 2012;24(3):561-574.
27. Machanavajjhala A, Kifer D, Gehrke J. l-Diversity: privacy beyond k-anonymity. Trans Knowl Discov from Data. 2007;1(1):3.
28. Li N, Li T, Venkatasubramanian S. t-Closeness: privacy beyond k-anonymity and l-diversity. Proc Int Conf Data Eng. 2007:106-115.
29. Nergiz ME, Atzori M, Clifton C. Hiding the presence of individuals from shared databases. Proc ACM SIGMOD Int Conf Manag data. 2007:665-676
30. Emam K El, Dankar F, Issa R, Jonker E, D. A globally optimal k-anonymity method for the de-identification of health data. J Am Med Inform Assoc. 2009;16(5):670-682.
31. Bayardo RJ, Agrawal R. Data privacy through optimal k-anonymization. Proc Int Conf Data Eng. 2005:217-228.
32. Ciglic M, Eder J, Koncilia C. k-anonymity of microdata with null values. Proc Int Conf Database and Expert Sys Appl. 2014.
33. ARX - Powerful Data Anonymization [cited 04 Aug 2014]. TUM. Available from: http://arx.deidentifier.org.
34. Integrated Health Interview Series [cited 04 Aug 2014]. NHIS. Available from: http://www.ihis.us.
35. Kohlmayer F, Prasser F, Eckert C, Kemper A, Kuhn KA. Highly efficient optimal k-anonymity for biomedical datasets. Proc Int Symp Computer-Based Medical Systems. 2012:1-6.
36. Kohlmayer F, Prasser F, Eckert C, Kuhn KA. A flexible approach to distributed data anonymization. J Biomed Inform. 2014;50:62-76.
37. El Emam K, Dankar FK. Protecting privacy using k-anonymity. J Am Med Inform Assoc. 2008;15(5):627-637.
38. Dankar FK, El Emam K, Neisa A, Roffey T. Estimating the re-identification risk of clinical data sets. BMC Med Inform Decis Mak. 2012;12(1):66.

Extending the HL7/LOINC Document Ontology Settings of Care

Sripriya Rajamani, MBBS, PhD, MPH[1, 2], Elizabeth S. Chen, PhD[4, 5, 6], Yan Wang, MS[2], Genevieve B. Melton, MD, MA[2, 3]

[1]Public Health Informatics Program, [2]Institute for Health Informatics, [3]Department of Surgery, University of Minnesota, Minneapolis, MN; [4]Center for Clinical & Translational Science, [5]Department of Medicine, [6]Department of Computer Science, University of Vermont, Burlington, VT

Abstract

Given federal mandates recommending document standards, increasing numbers of electronic clinical documents being created, and local initiatives/projects using clinical documents, there is a growing need to better represent clinical document metadata. The HL7/LOINC Document Ontology (DO) was developed to provide a standard representation of clinical document attributes with a multi-axis structure. Prior studies have demonstrated the need for extension of DO axes values and proposed new values for some axes, but significant gaps remain for representing the DO "Setting" axis. This study aimed to extend the "Setting" axis by combining the current values in the DO with values from 5 other sources. Evaluation and refinement by subject matter experts over a series of four iterative sessions resulted in a reorganized hierarchy with 254 additional values from a baseline of 20. Incorporating a comprehensive set of "Settings" in DO provides better representation of clinical information across the healthcare ecosystem.

Introduction

Federal initiatives like the Centers for Medicare & Medicaid Services (CMS) Electronic Health Record (EHR) Incentive Program along with standards and certification efforts led by the Office of the National Coordinator for Health Information Technology (ONC) have propelled the adoption and use of EHR systems[1]. These efforts have also led to dramatic increases in usage of electronic clinical documents, some of which are associated with requirements to use document standards for exchange of particular types of clinical information and for certain public health reporting use cases. With continuous increasing volumes of clinical documents, there is a need for better document metadata representation to promote standardization and reduce local customization. Availability of a comprehensive representation schema is critical to fit the needs of providers and the full set of inter-professional partners and sites in health care and also to facilitate their future retrieval and use for clinical care and research.

The HL7/LOINC Document Ontology (DO) is an ontology for the representation and standardization of clinical documents in a hierarchical structure comprising of five axes: *Kind of Document (KOD)*, *Type of Service* (*TOS*), *Setting*, *Subject Matter Domain* (*SMD*) and *Role*. The purpose of the DO is to support exchange of clinical documents across organizations, systems, and other stakeholders and to facilitate retrieval and reuse of documents for research and other secondary uses. Each DO axis is comprised of a set of values. A valid document specification with the DO consists of specification of the KOD axis and at least one additional DO axis. A select combination of valid DO specifications that have been precoordinated have assigned LOINC codes.

Several researchers have explored the application of DO and have proposed extensions to value sets, which have been incorporated by the LOINC Committee. A recent study that attempted mapping of data items from a large clinical repository to DO concluded that document attributes do not always link consistently with DO axes and recommended that additional values for certain axes, particularly for *Setting* and *Role* are needed[2]. This study addresses that recommendation and focuses on one axis of DO – the *Setting* axis.

The objective of this study was to create a comprehensive representation of care settings that combines the current values in DO with values from other representative sources. As emphasis shifts from traditional care in hospitals and office visits to care coordination, care management, and care within home and other nontraditional healthcare settings, healthcare is being offered in places that have been typically outside the scope of traditional health delivery which likely requires better representation in DO. As an additional potential benefit, this study is poised to benefit a wide range of secondary uses of documents for research, quality initiatives and population health management.

Background

Meaningful Use recommendations for document standards specify the HL7 Clinical Document Architecture (CDA) and HL7 Consolidated CDA (C-CDA)[3]. Both structure and semantics of a clinical document are guided by CDA, which is an XML-based standard[4]. C-CDA is a consolidation of various CDA templates to facilitate implementation[3].

The implementation guide for C-CDA that contains a library of CDA templates is available from HL7[5]. A companion guide was created by the ONC Standards and Interoperability Framework Initiative to offer guidance on implementation of certification criteria which required HL7 C-CDA[6].

The HL7 CDA header contains the context of a clinical document and facilitates exchange across and within institutions as well as retrieval for various purposes. The CDA body holds the actual facts pertaining to the document. The CDA header comprises of document information, encounter data, service actors and service targets. Document information includes a unique identifier that is drawn from the realm of clinical LOINC codes, which are guided by the HL7/LOINC Document Ontology[7]. Figure 1 includes a sample XML showing a portion of a C-CDA header that includes the LOINC code "34133-9" for "summarization of episode note" (the value for the LOINC System axis is "{Setting}" indicating that a particular setting is not designated).

Figure 1: Sample XML for C-CDA with Select Header Portion and Document Code[8]

```
<code
    codeSystem="2.16.840.1.113883.6.1"
    codeSystemName="LOINC"
    code="34133-9"
    displayName="Summarization of Episode Note"/>
```

The current model of DO was started by the HL7 Document Ontology Task Force (DOTF) and later continued through a joint effort with the LOINC Committee. The current DO is available as part of the LOINC User's Guide, which is updated biannually[9]. LOINC codes in the Document Ontology class (DOC.CLINRPT) and values from various axes can be downloaded as part of the LOINC Document Ontology file, which has been publicly available since the LOINC 2.44 release[10]. An implementation guide for HL7/LOINC Clinical Document Ontology has been released as of September 2013 as a Draft Standard for Trial Use (DSTU)[8]. As mentioned, this 5 axis model uses multiple classes under each axis requiring specification of the *KOD* axis and at least one of the other four axes. Though *Setting* is not a required component for a valid document specification, it is often included in clinical document names to help distinguish between important classes of documents. Initial development of DO with relevant background is described by Frazier *et al.*[11].

In related work, detailed analysis of the *Subject Matter Domain* (*SMD*) axis was conducted by Shapiro *et al.*[12] who mapped a set of inpatient document titles and showed that 56% of documents were classified as "not specified" with respect to values on the DO SMD list. This study created a new polyhierarchical *SMD* structure by combining the values from the DO SMD axis with values from American Board of Medical Specialties (ABMS), thereby increasing the coverage of SMD significantly. Veterans Administration (VA) has also contributed to expansion of the *Type of Service (TOS)* axis, which characterizes the kind of service or activity provided to the patient or subject of the service where sub-classes under this axis include evaluations, consultations, and summaries. More specifically, the VA work expanded the Compensation and Pension Examination value set that is under Disability Examination and part of the Evaluation and Management sub-class, and these are now available as part of DO[9].

Other pertinent studies have focused on mapping of document names from inpatient settings to DO and attempted to understand DO in depth and precoordination of its axes with LOINC. Document names from two large institutions were mapped to DO, and LOINC codes were identified based on mapping by Chen *et al.*[13]. This study showed that a majority of document names could be assigned a LOINC code, but granularity was often lost. The work highlighted strengths and limitations of DO and LOINC codes for representation of documents.

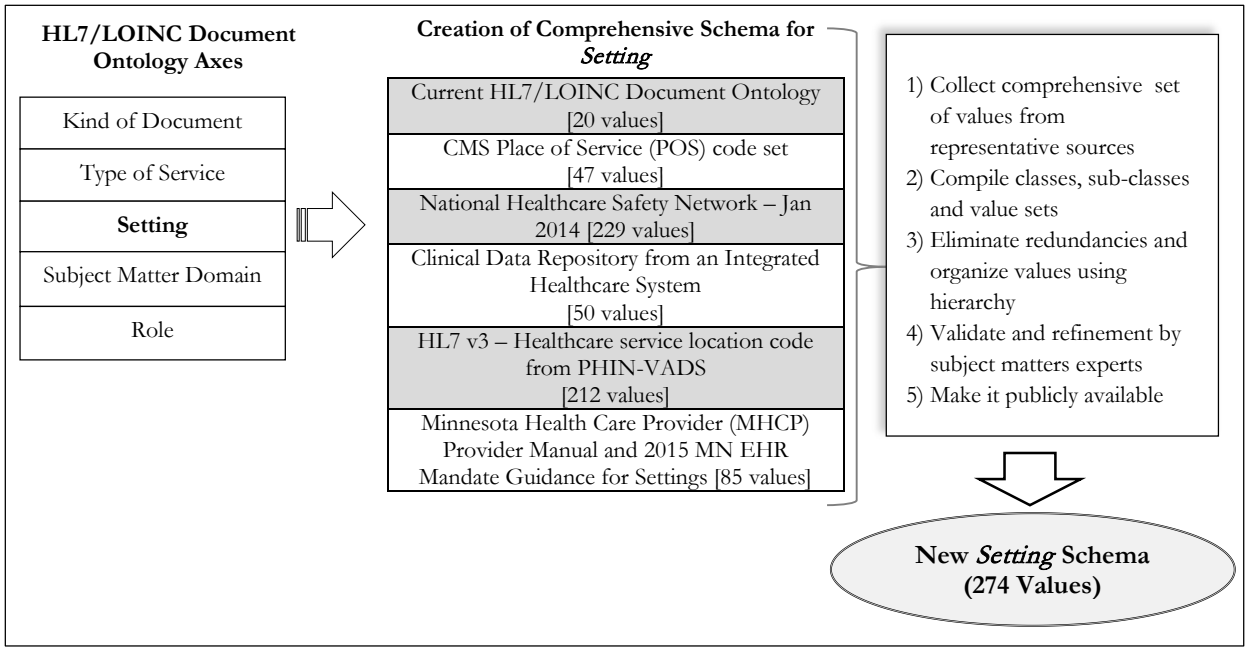
Comparison of the three versions of DO in their adequacy to represent document names in an inpatient setting was done by Hyun *et al.*[14]. Dugas *et al.*[15] examined 86 document types from an inpatient information system and their coverage using LOINC codes. Usage of LOINC codes for document exchange across inpatient settings of two campuses of a hospital was explored and coverage of DO on clinical document titles was evaluated as part of process by Li *et al.*[16]. Hyun and Bakken[17] extracted section headings from nursing documents and then identified DO components and mapped them to LOINC as a potential model for representing this class of texts. Most recently, Wang *et al.*[2] evaluated the adequacy of DO for representing clinical documents in a data repository from an integrated health delivery system which contained data from legacy and current EHR systems and is used for multiple purposes including research. A common theme across these studies was the need for extension of attribute values of the LOINC

semantic model, more specific LOINC codes and limitations with precoordination, and the recommendation to expand the value sets for certain DO axes, particularly for *Setting* and *Role*.

Methods

This study involved creation of a comprehensive representation of the *Setting* axis in the HL7/LOINC Document Ontology (DO). The methodology comprises of five main components: (1) collection of a comprehensive set of values from representative sources; (2) compilation of classes, sub-classes, and value sets for the *Setting* axis; (3) elimination of redundancies and organization of values into an extended hierarchy; (4) validation and refinement by subject matters experts; and (5) a resulting publicly available hierarchy for sharing the resultant value set with the broader community. Figure 2 details the methodology for the proposed expansion.

Figure 2: Methodology for Creation of Comprehensive Schema for *Setting* in Document Ontology



The current value set for *Setting* in DO[9] is comprised of 20 values that are organized into 12 classes with 8 sub-classes for two of the classes, leading to 20 overall values for document representation. The first step in the process was to examine the value sets for *Setting* in the secure local research-oriented clinical data repository, including large numbers of clinical documents, which was established in a partnership between the University of Minnesota and its affiliated integrated healthcare system, Fairview Health Services. Prior work of the authors examined the representation of documents in this repository and demonstrated inadequacy of the *Setting* axis in DO[2]. Next, other valid sources were identified that could contribute to representation of *Setting* values. The Place of Service (POS) code set[18] from CMS was identified and selected, as CMS is the largest insurer and provider of healthcare services in the US. In order to find corresponding value sets for a Medicaid source, values were obtained from the Minnesota Health Care Plans (MHCP) provider manual[19] and the 2015 Minnesota Electronic Health Record mandate guidance[20] on settings. Other sources incorporated in this study were the HL7 version 3 (v3) code set[21] for representation of settings, which was retrieved from the Public Health Information Network (PHIN) Vocabulary Access and Distribution System (PHIN-VADS). Since the HL7 codes were based off of the National Healthcare Safety Network (NHSN), the current version of NHSN codes (Jan 2014 release)[22] was also incorporated. Compilation of all these values resulted in classes and sub-classes of 327 values.

Once all value sets from the six sources including DO were obtained, redundancies were identified and eliminated. Table 1 shows some examples of values across sources and consensus arrived by subject matter experts as part of an iterative review process.

Table 1. Mapping Examples of Different Classes Across Sources

Current HL7/LOINC Document Ontology (LOINC User's Guide – Dec 2013)	CMS Place of Service (POS) code set – Nov 2012 version	Clinical Data Repository from an Integrated Healthcare System – June 2013	National Healthcare Safety Network (NHSN) – Jan 2014 release	HL7 v3 – Healthcare service location code set from PHIN-VADS (2010 version)	Minnesota Health Care Provider (MHCP) Provider Manual and 2015 MN EHR Mandate Guidance for Settings (2013)	Subject Matter Expert (SME) Consensus
Office	Office	Office				Office
			Physician's Office	Physician's Office		
Hospital						Inpatient setting
	Inpatient Hospital	Inpatient Hospital			Inpatient Hospital	
			Inpatient location			
Ambulatory Surgical Center	Ambulatory Surgical Center	Ambulatory Surgical Center	Ambulatory Surgical Center	Ambulatory Surgical Center	Ambulatory Surgical Center	Ambulatory Surgical Center
Urgent Care Center			Urgent Care Center	Urgent Care Center		Urgent Care Center
	Urgent Care Facility	Urgent Care Facility				
					Urgent Care Clinic	
						Public Health - Public Health Clinic - State Agency: Health - Prison Infirmary - Mass Immunization Center
	Public Health Clinic	Public Health Clinic				
					State Agency – Health	
			School or Prison Infirmary	School or Prison Infirmary		
	Mass Immunization Center	Mass Immunization Center				

As shown in Table 1, this process also included identifying synonyms and the need for "preferred terms." For example, "office" is used currently in the DO, CMS POS, and MHCP; however, "physician's office" is used in HL7 v3 and NHSN, which were meant to represent the same setting. From this, a consensus decision was made by the subject matter experts to keep the current "office" value and eliminate "physician's office" in light of the variety of inter-professional office settings, including non-physician provider offices. Also, some settings in NHSN were deemed irrelevant to this study which focuses on document representation. This phase of selection and redundant value harmonization reduced the total number of values to 274, which were then organized into a hierarchy.

A series of iterative sessions with 4 subject matter experts (2 providers, 1 public health provider, and 4 informaticians) was held where the hierarchy was constructed, refined, and validated. In particular, three of experts were experienced with the DO and biomedical standards evaluation. The reviewers looked at the following aspects of the sources and value sets in placing them into the hierarchy, including: (1) relative granularity of values, (2) representation of current content, and (3) need for additional layers in the hierarchy. Experts reviewed the value sets independently and also together to compare proposed revisions. This process was repeated iteratively four times until there was consensus amongst the experts. As a final step, the proposed final list was made available publicly for use by the larger community.

Results

The processes resulted in the creation of a comprehensive *Setting* value set with compilation of values from relevant and representative sources which included: current HL7/LOINC Document Ontology [20 values], CMS Place of Service (POS) code set [47 values], National Healthcare Safety Network [229 values], clinical data repository setting value set from an integrated healthcare system [50 values], HL7 v3 – healthcare service location code from PHIN-VADS [212 values] and Minnesota Health Care Provider (MHCP) provider manual and 2015 MN EHR mandate guidance for settings [85 values]. Using the methodology described, this resulted in the creation of a final proposed list of 274 values. Table 2 provides a comparison of the current hierarchy in DO for *Setting* with proposed extended hierarchy.

Table 2: Comparison of Values across Current and Proposed HL7/LOINC Document Ontology for Settings

Current HL7/LOINC Document Ontology Values [includes all classes and sub-classes]	***Proposed Extended Hierarchy for Settings of Care [displays main classes only]***
Ambulance	Inpatient setting
Birthing center	Critical care unit
Emergency department	Step down unit
Hospital	Mixed acuity unit
Intensive care unit	Operating room
Long term care facility	Ward
Custodial care facility	Inpatient psychiatric facility
Nursing facility	Hospice
Skilled nursing facility	24 hour observation area
Unskilled nursing facility	Hyperbaric oxygen center
Outpatient	Emergency department
Ambulatory surgical center	Outpatient setting
Office	Acute setting
Outpatient hospital	Clinic (non-acute) setting
Urgent care center	Other outpatient setting
Patient's home	Long term care facility
Pharmacy	Skilled nursing facility
Rehabilitation hospital	Unskilled nursing facility/Custodial care facility
Telehealth	Intermediate care facility
Telephone encounter	Behavioral health facility
	Rehabilitation center
	Birthing center
	Telehealth
	Telephone encounter
	Mobile
	Transport
	Pharmacy
	Laboratory
	Public Health
	Community
	Home

The current DO axis has 12 main classes and 8 sub-classes, whereas the study's proposed axis has 14 main classes and 17 sub-classes and also includes additional value sets. The representation of 274 values with the extended hierarchy comprises of Inpatient setting [n=97], Emergency department [n=2], Outpatient setting [n=93], Long term care facility [n=17], Behavioral health facility [n=5], Rehabilitation center [n=4], Birthing center [n=3], Telehealth [n=1], Telephone encounter [n=1], Mobile [n=8], Pharmacy [n=1], Laboratory [n=11], Public Health [n=21] and Community [n=10].

Table 3 offers a snapshot of the proposed values and their hierarchical organization for three of the 14 main classes: Public Health, Community, and Mobile. The entire structure comprising of 274 values is available publicly (http://www.bmhi.umn.edu/ihi/research/nlpie/) for dissemination to the broader community (e.g., for sharing with the LOINC Committee and other interested stakeholders).

Table 3: Snapshot of Proposed Extensions to the HL7/LOINC Document Ontology for *Setting*

Proposed Values in Select Classes		
Public Health	**Community**	**Mobile**
City/county agencies: Jails	**Home**	Mobile unit
City/county agencies: Human services	Home-based hospice	Mobile blood collection center
City/county agencies: Local health departments	Home hemodialysis	Mobile MRI/CT
Community mental health center	Group home	**Transport service**
Federally qualified health center (FQHC)	Assisted living facility	Mobile emergency services/EMS
Indian health service free-standing facility	Place of employment - worksite	Ambulance – air or water
Indian health service provider-based facility	Homeless shelter	Ambulance – land
Mass immunization center	Temporary lodging	
Migrant health clinic	Other place of service	
Military treatment facility		
Prison – Correctional facility		
Prison Infirmary		
Public health clinic		
State agency – corrections		
State agency – health		
State agency – human services		
State or local health clinic		
Tribal 638 free-standing facility		
Tribal 638 provider-based facility		
Veterans Affairs (VA)		

Discussion

In this study, we have described an effort to construct a comprehensive value set for representation of *Setting* for the HL7/LOINC Document Ontology. The methodology involved identifying additional representative and valid sources for *Setting* values, integrating these value sets to eliminate redundancies and combine synonyms, extending the existing DO hierarchy for *Setting* and organizing the values into this hierarchy, validating with subject matter experts in multiple sessions, and disseminating the results. One of the challenges encountered in this work was the effort and decisions around optimal hierarchical representation of value sets. The sessions with subject matter experts revealed the complexity of this task due to numerous places in which care is delivered across inpatient, outpatient, home, and community locations, as some of the major settings of care, and the variation in care delivery by organizational structure. Also, there were numerous redundancies and synonyms to deal with across the value sets from various sources. Expert review revealed that there are a multitude of ways to organize the setting values, depending on the specification of higher classes and the granularity of terms. One consensus was that the *Setting* axis in the current HL7/LOINC Document Ontology needs inclusion of additional settings of care to truly reflect the changing healthcare delivery landscape to accurately capture information on settings of care and facilitate representation and reuse of documents for research.

Guidelines were also developed for systematic review and organization of settings in a hierarchy. Criteria which were discussed included: (1) inclusion of specialty locations separately in a particular setting or roll them up as part of over-arching settings of care such as inpatient and outpatient, (2) potential need for poly-hierarchy with a class to be included as a sub-class of various categories, (3) validity of sources for inclusion, (4) addition of other reliable sources, and (5) identification of main classes in settings of care. Multiple rounds of expert review, including both independent and combined sessions revealed the complexity in organization of the setting concept. These assessments also revealed inconsistencies in representation across sources, redundancies, and synonyms, all of which had to be addressed.

One of the limitations of this work is the difficulty in obtaining data on internal validity. Although we did have multiple rounds and exposed the data hierarchy to multiple expert stakeholders, because of the large number of values and need to leverage group wisdom about care settings, separate and independent assessments of the value sets was found to be difficult for applying inter-rater reliability measures, which are often applied to terminology coding tasks. An ongoing challenge is balancing the comprehensiveness of information in the hierarchy and associated granularity with the

practical aspects of implementation and mapping burden. This study proposes an over 10-fold extension of the *Setting* axis values from a current list of 20 to a proposed set of 274 values. More coverage in the Setting axis is needed, but further efforts are needed to determine the optimal set of values and their organization for uses ranging from clinical care to research.

Next steps for this study would be submission of the value sets to the LOINC Committee for a formal review as outlined in the LOINC User's guide[9]. This is a critical next step in the generalizability of this work and its applicability to *Setting* axis extension. Other ongoing work involves studying the representation of *Role* in DO as a recent study recommended the need for additional values in the *Role* axis[2]. Ultimately, these findings need to be promoted and value sets shared as a form of external validation by other interested stakeholders, including those from the standards community. Both informal feedback and also the formal LOINC submission and review process would decide the extent of *Setting* axis expansion. As this study group and others continue to promote better representation of clinical information across providers and partners in healthcare ecosystem, the extension of DO value sets is needed to incorporate a comprehensive and reflective set of values for settings of care.

Acknowledgement

The authors would like to thank grant support from National Library of Medicine 1R01LM011364-01 (EC/GM), Agency for Healthcare Research and Quality 1R01HS022085-01 (GM) and the University of Minnesota Clinical and Translational Science Award 8UL1TR000114-02.

References

1. Centers for Medicare and Medicaid Services (CMS). EHR Incentive Programs. 2014. Available from: http://www.cms.gov/ehrincentiveprograms. Last accessed March 11, 2014.
2. Wang Y, Pakhomov S, Dale J, Chen ES, Melton GB. Application of HL7/LOINC document ontology to a university-affiliated integrated health system research clinical data repository. AMIA Summit proceedings/2014 Joint Summits on Translational Science. 2014.
3. Office of the National Coordinator for Health Information Technology (ONC). Standards Hub. 2014. Available from: http://www.healthit.gov/policy-researchers-implementers/meaningful-use-stage-2-0/standards-hub. Last accessed July 31, 2014.
4. Dolin RH, Alschuler L, Boyer S, Beebe C, Behlen FM, Biron PV, et al. HL7 Clinical Document Architecture, Release 2. Journal of the American Medical Informatics Association : JAMIA. 2006;13(1):30-9. Epub 2005/10/14.
5. Health Level Seven (HL7). HL7 Implementation Guide for CDA® Release 2: IHE Health Story Consolidation, Release 1.1 - US Realm. 2012. Available from: http://www.hl7.org/implement/standards/product_brief.cfm?product_id=258. Last accessed July 30, 2014.
6. ONC Standards and Interoperability Framework. Companion Guide to HL7 Consolidated CDA for Meaningful Use Stage 2. 2012. Available from: http://wiki.siframework.org/Companion+Guide+to+Consolidated+CDA+for+MU2. Last accessed July 30, 2014.
7. Dolin RH, Alschuler L, Beebe C, Biron PV, Boyer SL, Essin D, et al. The HL7 Clinical Document Architecture. Journal of the American Medical Informatics Association : JAMIA. 2001;8(6):552-69. Epub 2001/11/01.
8. Health Level Seven (HL7). HL7/LOINC Clinical Document Ontology Implementation Guide (US Realm), Draft Standard for Trial Use. 2013. Available from: http://wiki.hl7.org/index.php?title=LOINC_Clinical_Document_Ontology. Last accessed July 30, 2014.
9. McDonald C, Huff S, Deckard J, Holck K, Vreeman D. LOINC User's Guide - December 2013 Available from: http://loinc.org/downloads/files/LOINCManual.pdf. Last accessed March 11, 2014.
10. The Regenstrief Institute. LOINC Document Ontology File. 2014. Available from: http://loinc.org/downloads/accessory-files. Last accessed July 30, 2014.
11. Frazier P, Rossi-Mori A, Dolin RH, Alschuler L, Huff SM. The creation of an ontology of clinical document names. Studies in health technology and informatics. 2001;84(Pt 1):94-8.
12. Shapiro JS, Bakken S, Hyun S, Melton GB, Schlegel C, Johnson SB. Document ontology: supporting narrative documents in electronic health records. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2005:684-8. Epub 2006/06/17.
13. Chen ES, Melton GB, Engelstad ME, Sarkar IN. Standardizing Clinical Document Names Using the HL7/LOINC Document Ontology and LOINC Codes. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2010;2010:101-5. Epub 2011/02/25.

14. Hyun S, Shapiro JS, Melton G, Schlegel C, Stetson PD, Johnson SB, et al. Iterative evaluation of the Health Level 7--Logical Observation Identifiers Names and Codes Clinical Document Ontology for representing clinical document names: a case report. Journal of the American Medical Informatics Association : JAMIA. 2009;16(3):395-9. Epub 2009/03/06.
15. Dugas M, Thun S, Frankewitsch T, Heitmann KU. LOINC codes for hospital information systems documents: a case study. Journal of the American Medical Informatics Association : JAMIA. 2009;16(3):400-3. Epub 2009/03/06.
16. Li L, Morrey CP, Baorto D. Cross-mapping clinical notes between hospitals: an application of the LOINC Document Ontology. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2011;2011:777-83. Epub 2011/12/24.
17. Hyun S, Bakken S. Toward the creation of an ontology for nursing document sections: mapping section names to the LOINC semantic model. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2006:364-8. Epub 2007/01/24.
18. Centers for Medicare and Medicaid Services (CMS). Place of Service Code Set. 2012. Available from: http://www.cms.gov/Medicare/Coding/place-of-service-codes/Place_of_Service_Code_Set.html. Last accessed March 11, 2014.
19. Minnesota Department of Human Services. Minnesota Health Care Programs (MHCP) Provider Manual. 2014. Available from: http://www.dhs.state.mn.us/main/idcplg?IdcService=GET_DYNAMIC_CONVERSION&RevisionSelectionMethod=LatestReleased&dDocName=dhs16_157386#. Last accessed March 11, 2014.
20. MDH Office of Health Information Technology. Guidance for Understanding the Minnesota 2015 Interoperable EHR Mandate 2013. Available from: http://www.health.state.mn.us/e-health/hitimp/2015mandateguidance.pdf. Last accessed March 11, 2014.
21. HL7. Healthcare Service Location (HL7) Code System. 2010. Available from: https://phinvads.cdc.gov/vads/SearchCodeSystems_search.action?searchOptions.searchText=location. Last accessed March 11, 2014.
22. National Healthcare Safety Network (NHSN), Centers for Disease Control and Prevention (CDC). CDC Locations and Descriptions. 2014. Available from: http://www.cdc.gov/nhsn/pdfs/psc/mappingpatientcarelocations.pdf. Last accessed March 11, 2014.

Differences in Nationwide Cohorts of Acupuncture Users Identified Using Structured and Free Text Medical Records

Doug Redd, MS[1,2], Jinqiu Kuang, MS[1], Qing Zeng-Treitler, PhD[1,2]

[1]VA Salt Lake City Health Care System; [2]Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah

Abstract

Integrative medicine including complementary and alternative medicine (CAM) has become more available through mainstream health providers. Acupuncture is one of the most widely used CAM therapies, though its efficacy for treating various conditions requires further investigation. To assist with such investigations, we set out to identify acupuncture patient cohorts using a nationwide clinical data repository. Acupuncture patients were identified using both structured data and unstructured free text notes: 44,960 acupuncture patients were identified using structured data consisting of CPT codes;. Using unstructured free text clinical notes, we trained a support vector classifier with 86% accuracy and was able to identify an additional 101,628 acupuncture patients not identified through structured data (a 226% increase). In addition, characteristics of the patients identified through structured and unstructured data were compared, which show differences in geographic locations and medical service usage patterns. Patients identified with structured data displayed a consistently higher use of the Veterans Health Administration (VHA) medical system.

Introduction

Over the past decade, integrative medicine has gained increasing attention from providers and researchers. Compared to traditional healthcare, integrative medicine's emphasis on a partnership between patients and clinicians takes a holistic view of patients' health and well being, and incorporates complementary and alternative medicine (CAM) approaches such as acupuncture and massage into treatment options. Many large hospitals now provide some form of integrative health services to their patients.

At the same time, the safety and effectiveness of many CAM treatments are not sufficiently understood. For instance, acupuncture is widely practiced to relieve pain and treat certain health problems, but debate on its effectiveness continues in the literature. Witt et al evaluated clinical and economical effectiveness of acupuncture on chronic low back pain in a large randomized controlled trial (RCT).(1) They demonstrated that acupuncture in addition to routine care considerably improved clinical outcomes and was relatively cost-effective. A systematic review of RCTs looking at acupuncture for pain was published by Linde et.al. It included thirteen trials (3,025 patients) with a variety of pain conditions and found a small analgesic effect from acupuncture, hardly distinguishable from bias.(2) Another systematic review of 23 RCTs on the effectiveness of acupuncture for nonspecific lower back pain by Yuan et al showed moderate evidence that acupuncture is more effective than no treatment, and strong evidence of no significant difference between acupuncture and sham acupuncture, for short-term pain relief.(3) This review concluded that acupuncture can be a useful supplement to other forms of conventional therapy for nonspecific lower back pain, but the effectiveness of acupuncture compared with conventional therapies requires further investigation. Considering acupuncture is one of the most studied CAM modalities, these uncertain results indicate that more research is needed to ascertain the efficacy of CAM practices.

Secondary analysis of electronic medical records (EMR) is a powerful approach to study treatment safety and effectiveness. At the Veterans Health Administration (VHA), we have begun leveraging its nationwide EMR repository to study the use of acupuncture to manage pain and control other symptoms like nausea. A critical step in EMR secondary analysis is cohort identification.

In this paper, we describe our effort to identify a cohort of patients who had undergone acupuncture treatments while receiving care from the VHA. Both structured data and unstructured data were used. To understand the impact of data source on the resultant cohorts, cohorts identified from the two methods were compared in terms of size of patient characteristics.

Materials and Methods

Data Source

Data for this study was procured through the Veterans Informatics and Computing Infrastructure (VINCI), VHA. The VHA comprises 152 medical facilities in addition to 1,400 clinics that are community-based and tailored to serve individuals on an outpatient basis, Vet Centers, community living centers, and Domiciles. In total, these facilities employ over 53,000 healthcare professionals who provide their services to over 8.3 million veterans on an annual basis. VINCI is a collaboration between the Office of Research and Development and the Office of Information and Technology in the U.S. Department of Veterans Affairs (VA), providing data and infrastructure needs of the VHA research community. VINCI provides access to structured and unstructured health information originating from the VISTA electronic health record system, and includes data for over 17 million patients. We identified patients receiving acupuncture treatments through structured as well as unstructured data using the process outlined in Figure 1.

Cohort Identification Using Structured Data

VHA offers many forms of CAM treatments from acupuncture to sweat lodge. Patients receiving specific treatments within the VHA system can be identified through Current Procedural Terminology (CPT) codes identifying specific patient procedures. Acupuncture treatments are represented by CPT codes 97780, 97781, 97810, 97811, 97813, and 97814.

Many non-standard treatments can be identified through the locations of patient visits. In the VHA, clinic "Stop Codes" are included in the outpatient visit records to indicate the clinic or work group providing specific services. We were, however, only able to identify a single location for acupuncture services using the "Stop Code". Since acupuncture services are widespread in the VHA system, we resorted to CPT codes for their identification.

Cohort Identification Using Free Text Data

Structured data has been shown to be insufficient for cohort identification in many cases (4). Some patients receiving acupuncture will not have corresponding CPT codes assigned for various reasons. For example, many patients obtain treatment from non-VHA providers, particularly when VHA clinics offering a specific therapy are not available in the geographic area of the patient. In some cases, their VHA clinicians do not prescribe or authorize the treatments. Although they may not be recorded by CPT codes, many VHA healthcare providers do ask Veterans about the non-VHA treatments they are receiving and document them in narrative clinical notes. Thus, we searched unstructured, free text clinical notes for mentions of acupuncture.

Free text clinical notes have been shown to be rich in medical information that can be accessed using natural language processing techniques (5-7). Searching of the unstructured notes was accomplished using the Voogo search engine, which was developed specifically for searching structured and unstructured data within VINCI. Using Voogo, patients with clinical documents containing the string "acupuncture" were identified. Snippets of text containing acupuncture, including surrounding context, were extracted and manually annotated to identify if the snippets were positive, negative, or prescribed (if the snippet described a recommendation) for use of acupuncture treatment by the patient. A support vector machine (SVM) was trained for automated acupuncture text classification. Using text classification results, patients were classified as positive for acupuncture treatment use if they had at least one positive snippet; prescribed if they had no positive snippets but at least one prescribed snippet; or negative if they had only negative snippets.
Patients with a positive history of acupuncture use identified through unstructured data is referred to as UD.

Comparing Cohorts from Structured and Free Text Data

We compared the two cohorts (UD and SD) to determine the distribution of patients identifiable only from SD, only from UD, or both. The UD and SD patients were then compared and contrasted for geographic location, gender, age, and most frequent medical procedures, diagnoses, and prescriptions.

A list of the 25 most common procedures was constructed by combining the 21 most frequent Current Procedural Terminology (CPT) codes from UD patients and the 21 most frequent CPT codes from SD patients (the number of codes from UD and SD patients was chosen by trial and error to obtain a combined number of 25). Similarly, a list of the 25 most common diagnoses was constructed by combining the 23 most frequent International Classification of Diseases version 9 (ICD-9) codes from UD and SD patients. And again with prescriptions, the 24 most frequent drug names from UD and SD patients were determined, for a combined set of 25 unique drug names. We then determined the proportions of UD and SD patients receiving these most frequent procedures, diagnoses, and prescriptions..

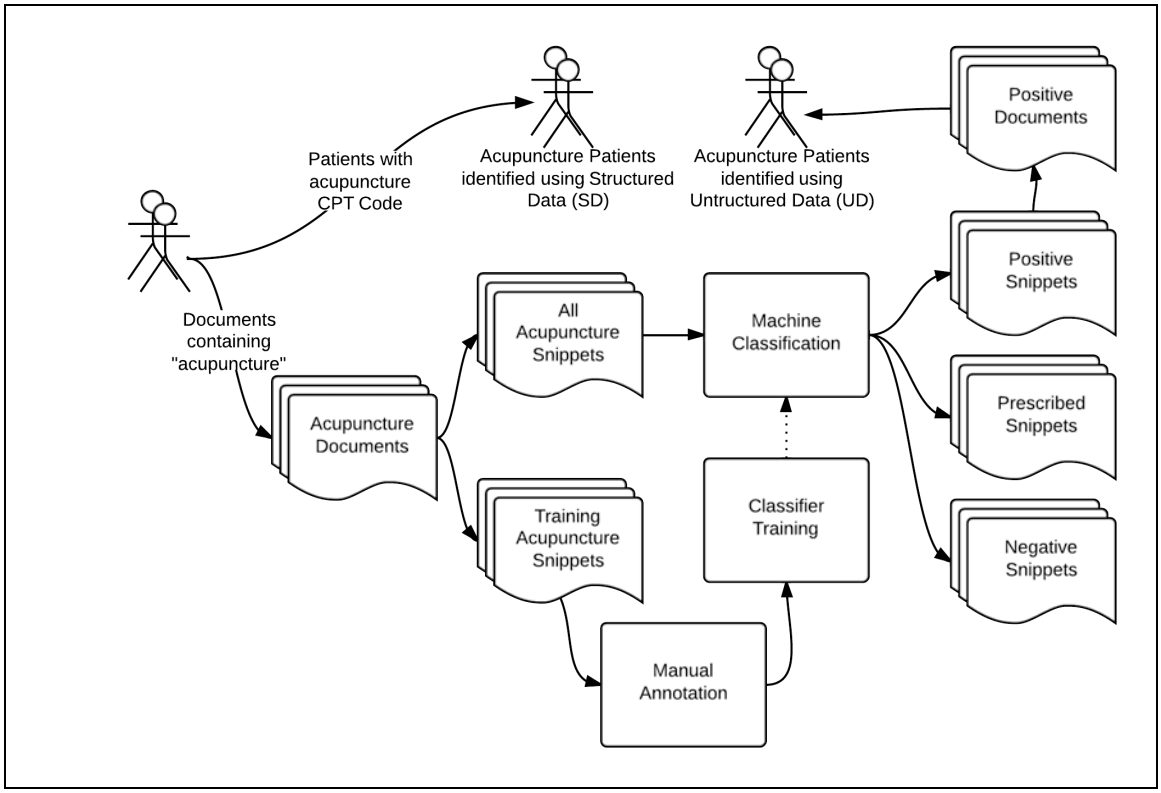


Figure 1. Acupuncture Patient Cohort Identification from Structured Data (SD) and Unstructured Data (UD)

Results

Using CPT codes, 44,960 patients were identified as receiving acupuncture treatment using structured For identification of acupuncture using unstructured data (UD), 1,245,753 documents mentioning identified representing 400,350 patients. 297 snippets were classified as positive, prescribed, or annotators with an inter-rater reliability kappa score of 0.74. Since the kappa is relatively low, reached through discussion to create the reference standard. A support vector machine (SVM) using these snippets and validated with 10-fold cross validation. This resulted in the ability to identify text with an overall accuracy of 0.862 (precision 0.883, recall 0.743, and f_1-measure 0.785) (Table 1). Using the SVM classification model, 140,525 patients were identified as positive for acupuncture use. SD and UD identified patients were compared to determine an intersection of 38,897 patients, so that an additional 101,628 (226%) patients were identified using UD that were not identifiable using SD (Figure 2).

Table 1. Confusion matrix for acupuncture classifier.

		Reference Standard			
		Yes	No		
Acupuncture Classifier	Yes	77	13	Precision	88.3%
	No	24	183	Recall	74.3%
				F1 Measure	78.5%
				Accuracy	86.2%

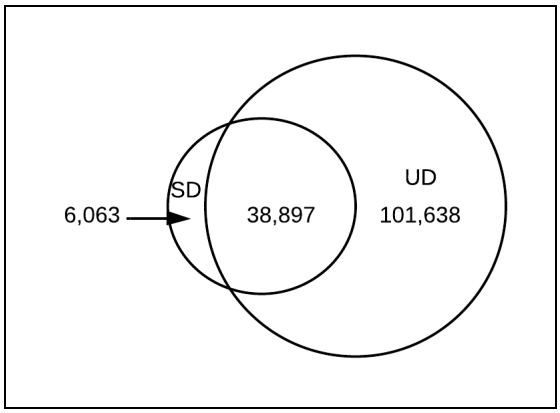


Figure 2. Distribution of patients between groups identifiable by structured data (SD), unstructured data (UD), or both (SD + UD).

We compared the geographic locations of UD and SD patients. Overall, patients congregated around major population centers. There were some differences in the distributions, however. There was a much higher proportion of UD patients in the northwest region of Oregon, and a higher proportion of SD patients in the New York City metropolitan region. (Figure 3).

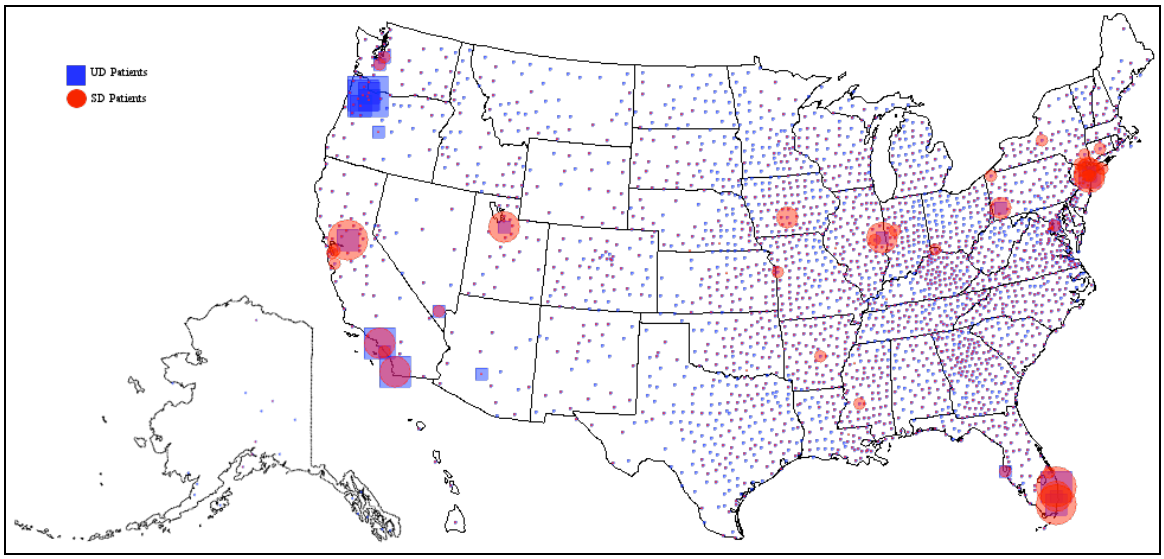


Figure 3. Geographic distribution of acupuncture patients identified from structured data (SD) and unstructured data (UD).

The age distribution of UD and SD patients were essentially similar (Figure 4). The mean age was 56.4 (stdev. 15.2) for UD patients and 56.1 (stdev. 14.6) for SD patients. The difference was statistically significant ($p < 0.00$ by student t-test) due to the large sample size, however this small difference is not clinically meaningful.

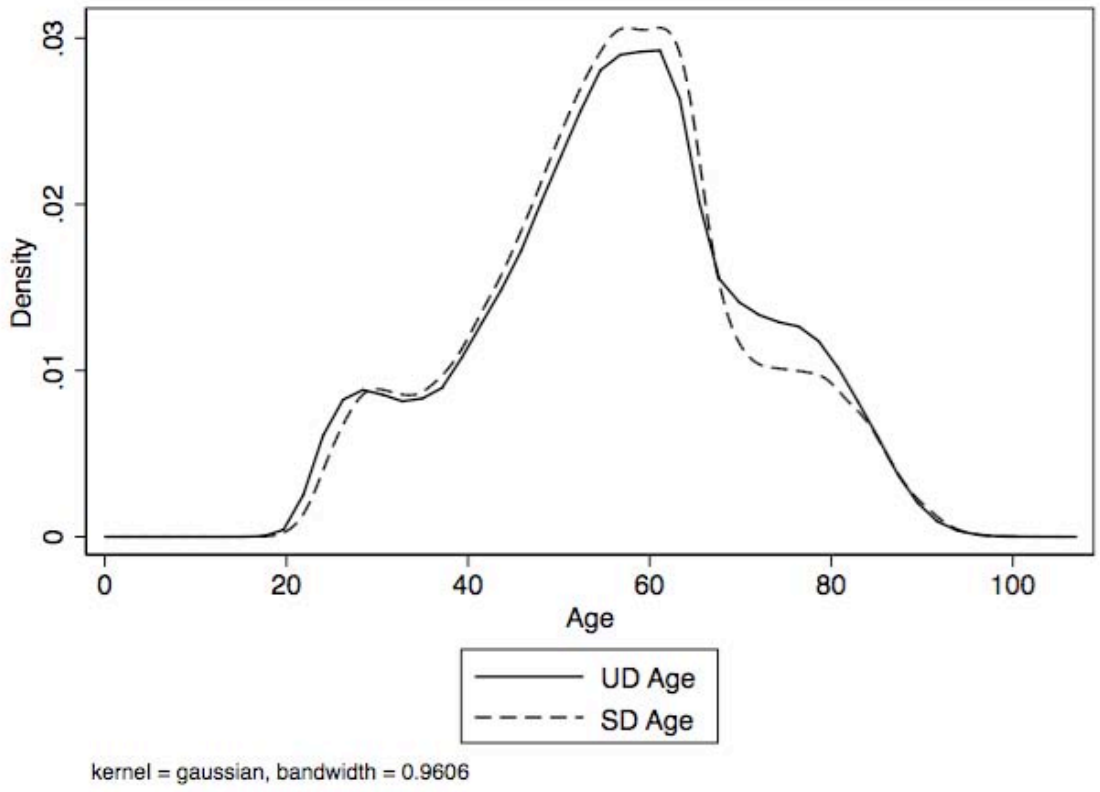


Figure 4. Density of age distribution for acupuncture patients identified from unstructured data (UD) and structured data (SD).

We compared the percent of UD and SD patients receiving the most common procedures, diagnoses, and prescriptions (Figure 5). Although the percentages are somewhat similar, overall a higher percent of SD patients received the measured procedures, diagnoses, and prescriptions. There are some procedures where SD patients show a higher percent that is more pronounced, i.e. metabolic panel total calcium, patient evaluation, therapeutic exercises, and comprehensive metabolic panels. Alternatively, UD patients show a higher percent of assays for quantitative blood glucose, alanine aminotransferase, and assay of urea nitrogen. For diagnoses SD patients also have a higher percentage in most cases, exceptions including unspecified reason for consultation and unspecified tobacco use disorder. Prescriptions continue the trend of higher SD percentages, with SD having higher percentages in all cases. We also examined the per-patient average number of all procedures, diagnoses, prescriptions, and visits between the two groups. This analysis confirmed that SD patients had a higher rate of use in all cases (Table 2).

Overall, both diagnoses and prescriptions indicate the presence of pain and pain management. Diagnoses of lumbago (low back pain), unspecified back pain, and cervicalgia (neck pain) are frequent, as are pain medications such as hydrocodone, gabapentin, cyclobenzaprine, naproxen, tramadol, oxycodone, codeine, etc.

We also compared the gender distribution (Figure 5). There was a higher percent of females in SD patients (14%) as opposed to UD patients (12%), both of which reflect the expected minority of females in the veteran population.

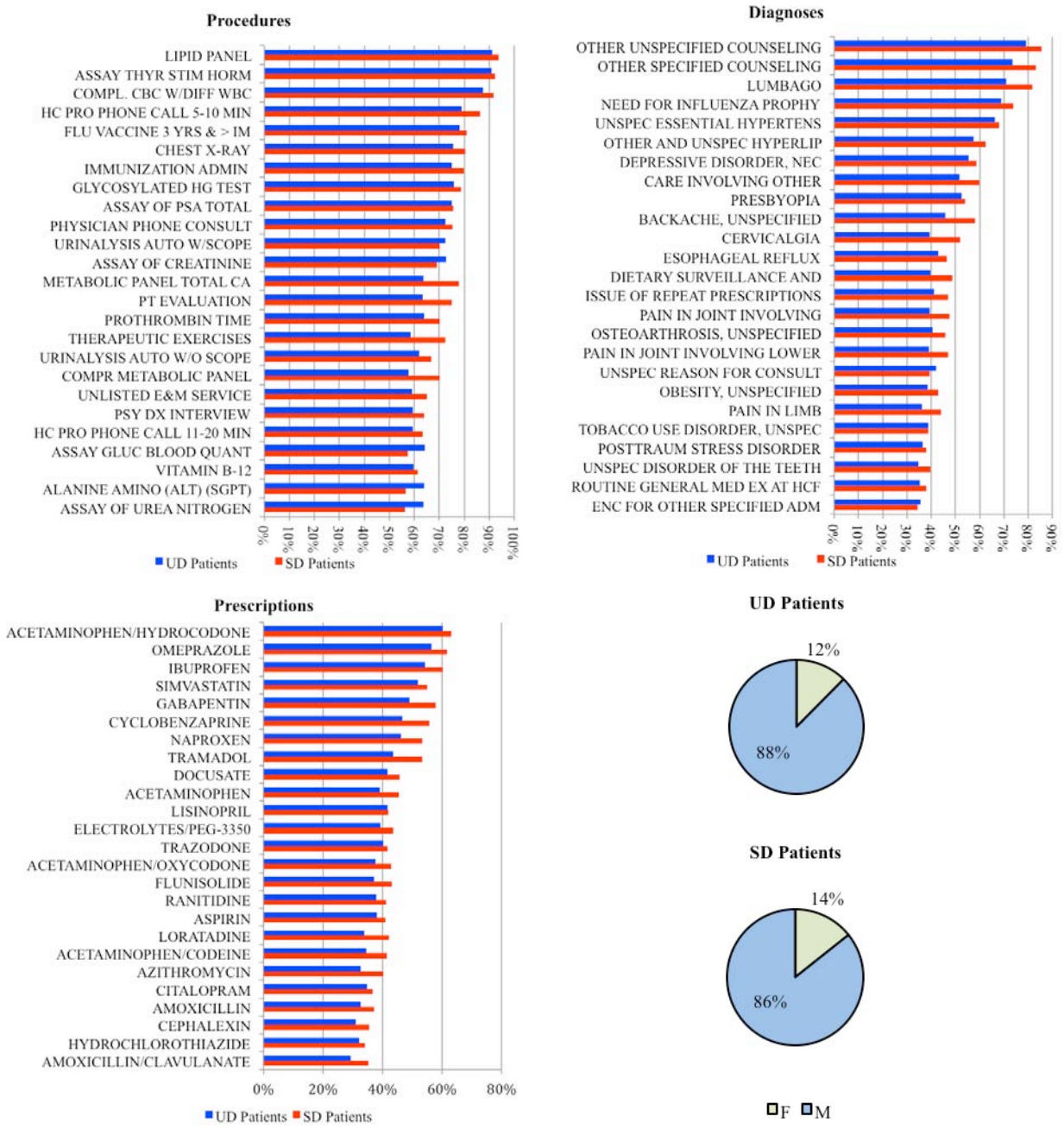


Figure 5. Percent of UD and SD patients with the most frequent procedures (by CPT code), diagnoses (by ICD9 code), and prescriptions, and gender distribution of UD and SD patients.

Table 2. Average per-patient procedure, diagnosis, prescription, and outpatient visit rates for UD and SD patients.

	Procedures per Patient	Diagnoses per Patient	Prescriptions per Patient	Visits per Patient
UD Patients	634	473	202	483
SD Patients	724	568	232	562

Discussion

In this study we identified patients in the VHA system being treated with acupuncture. We identified the cohorts using structured data and unstructured full-text data. We used CPT codes to identify patients for the structured data cohort, and SVM classification of unstructured free-text clinical notes to identify patients in the unstructured data cohort. There was a large overlap in the two sets, with only 13% of structured data patients not also being present in the unstructured data set. However, 72% of the unstructured data patients were not present in the structured data set, demonstrating the ability to significantly enlarge the set of identified acupuncture patients by using unstructured data. Our study shows that while it is feasible to identify acupuncture cohorts through structured and unstructured data independently, combining the two approaches can maximize the cohort size. This finding is consistent with findings reported by prior studies (8-11), but we show a more dramatic increase due to this medical domain not being traditionally included in electronic health records.

Aside from increasing the cohort size, combining structured and unstructured data can lead to a more representative patient population. Some prior studies compared sensitivity and specificity of different cohort identification methods, while we compared the cohorts. In comparing the cohort characteristics, we found a high degree of similarity but also some meaningful differences. Geographically, large acupuncture patient populations tend to locate in or near large metropolitan centers. The unstructured data cohort had a much higher proportion in the northwest region of Oregon, and those in the structured data cohort were proportionally more highly represented in the New York City metropolitan area. Some large metropolitan centers showed low acupuncture populations from either method. This suggests a variance in practice and/or documentation, although there are many other possibilities that will require further study to identify.

The distribution of ages in the two cohorts showed no significant difference, with the mean patient age at the time of treatment being about 56 years old in both groups. The gender representation in the two cohorts was also very similar. The rankings of the most frequent medical procedures, diagnoses, and prescriptions were very similar between the two cohorts, however there were consistently higher percentages of patients in the structured data cohort that received each procedure, diagnosis, and prescription. A frequent application of acupuncture treatment is for pain management (12), which is reflected in the frequent use of pain management prescriptions and diagnoses related to pain conditions in both cohorts.

Our data suggest that the patients in the structured data cohort had consistently higher rates of procedures, diagnoses, and prescriptions in general, not only in the most frequent sets. Patients in the structured data set also had a higher average outpatient visit rate. This indicates a difference in medical resource usage pattern between the two cohorts, with those in the structured data cohort consistently displaying higher use of VHA resources. This may indicate that patients identified only through unstructured data are relatively healthy or relying less on VHA as the sole provider.

Acknowledgements

This work is funded by VA grants CHIR HIR 08-374 and VINCI HIR-08-204.

References

1. Witt CM, Jena S, Selim D, Brinkhaus B, Reinhold T, Wruck K, et al. Pragmatic randomized trial evaluating the clinical and economic effectiveness of acupuncture for chronic low back pain. American journal of epidemiology. 2006;164(5):487-96.
2. Linde K, Vested Madsen M, Gøtzsche PC, Hrobjartsson A. Acupuncture Treatment for Pain: Systematic Review of Randomised Clinical Trials with Acupuncture, Placebo Acupuncture, and no Acupuncture Groups. Deutsche Zeitschrift für Akupunktur. 2010;53(2):40-1.
3. Yuan J, Purepong N, Kerr DP, Park J, Bradbury I, McDonough S. Effectiveness of acupuncture for low back pain: a systematic review. Spine. 2008;33(23):E887-E900.
4. Jacobson BC, Gerson LB. The inaccuracy of ICD-9-CM Code 530.2 for identifying patients with Barrett's esophagus. Diseases of the esophagus : official journal of the International Society for Diseases of the Esophagus / ISDE. 2008;21(5):452-6.

5. Lin J, Jiao T, Biskupiak JE, McAdam-Marx C. Application of electronic medical record data for health outcomes research: a review of recent literature. Expert review of pharmacoeconomics & outcomes research. 2013;13(2):191-200.
6. Friedlin J, McDonald CJ. A natural language processing system to extract and code concepts relating to congestive heart failure from chest radiology reports. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2006:269-73.
7. Friedlin J, Grannis S, Overhage JM. Using natural language processing to improve accuracy of automated notifiable disease reporting. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2008:207-11.
8. Liao KP, Cai T, Gainer V, Goryachev S, Zeng-treitler Q, Raychaudhuri S, et al. Electronic medical records for discovery research in rheumatoid arthritis. Arthritis care & research. 2010;62(8):1120-7.
9. Li L, Chase HS, Patel CO, Friedman C, Weng C. Comparing ICD9-encoded diagnoses and NLP-processed discharge summaries for clinical trials pre-screening: a case study. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2008:404-8.
10. Jeff Freidlin D, Marc Overhage MD P, Mohammed A Al-Haddad M, Joshua A Waters M, J. Juan R Aguilar-Saavedra M, Joe Kesterson M, et al., editors. Comparing Methods for Identifying Pancreatic Cancer Patients Using Electronic Data Sources. AMIA 2010 Symposium; 2010.
11. Elkin PL, Froehling D, Wahner-Roedler D, Trusko B, Welsh G, Ma H, et al. NLP-based identification of pneumonia cases from free-text radiological reports. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2008:172-6.
12. Vickers AJ, Cronin AM, Maschino AC, Lewith G, MacPherson H, Foster NE, et al. Acupuncture for chronic pain: individual patient data meta-analysis. Archives of internal medicine. 2012;172(19):1444-53.

Development, Implementation and Use of Electronic Surveillance for Ventilator-Associated Events (VAE) in Adults

Ervina Resetar, MIM, PMP[1,3], Kathleen M. McMullen, MPH, CIC[2], Anthony J. Russo MPH[2], Joshua A. Doherty, BS[3], Kathleen A. Gase, MPH, CIC[3], Keith F. Woeltje, MD, PhD[1,3]

[1]Washington University School of Medicine, Saint Louis, MO
[2]Barnes Jewish Hospital, Saint Louis, MO
[3]BJC HealthCare, Saint Louis, MO

Abstract

Mechanical ventilation provides an important, life-saving therapy for severely ill patients, but ventilated patients are at an increased risk for complications, poor outcomes, and death during hospitalization.[1] The timely measurement of negative outcomes is important in order to identify potential issues and to minimize the risk to patients. The Centers for Disease Control and Prevention (CDC) created an algorithm for identifying Ventilator-Associated Events (VAE) in adult patients for reporting to the National Healthcare Safety Network (NHSN). Currently, the primarily manual surveillance tools require a significant amount of time from hospital infection prevention (IP) staff to apply and interpret. This paper describes the implementation of an electronic VAE tool using an internal clinical data repository and an internally developed electronic surveillance system that resulted in a reduction of labor efforts involved in identifying VAE at Barnes Jewish Hospital (BJH).

Introduction

The use of mechanical ventilation can be an important, life-saving treatment for severely ill patients. The Centers for Disease Control and Prevention (CDC) estimate as many as 300,000 patients receive mechanical ventilation each year, but also notes that this can lead to additional complications, poor outcomes, and death during hospitalization.[1] Because of the increased risk to patients, it is important that hospitals monitor for adverse events associated with the use of mechanical ventilation.

Prior to 2013, the BJC HealthCare (BJC) Infection Prevention (IP) staff performed surveillance for ventilator-associated pneumonia (VAP). VAP is among the most common hospital-acquired infections, but accurate surveillance for VAP was difficult because of the lack of universally accepted objective definitions. VAP surveillance was time consuming, potentially less accurate than clinical/microbiologic criteria, and the use of quantitative lower respiratory tract cultures for the establishment of VAP is not universally performed.[2-3]

The CDC's National Healthcare Safety Network (NHSN) Working Group developed a new and more objective approach for surveillance that focuses on Ventilator-Associated Events (VAE). These include Ventilator-Associated Conditions (VAC), Infection-related Ventilator-Associated Complications (IVAC), possible VAP and probable VAP.[1] The new methodology was developed with the goals of: limiting the VAP definition's subjectivity and inaccuracy; using readily available and objective clinical data; making inter-facility comparisons more meaningful; and encouraging broader prevention strategies. It was also designed to allow electronic data collection, given the availability of a hospital-wide electronic health record (EHR) system with exportable data. The new definition still requires daily monitoring of patients, and would be time intensive to complete manually in large institutions. In order to decrease the burden of manual surveillance for VAE, BJC developed an electronic surveillance process for gathering relevant data elements. The approach used a combination of new interfaces and the incorporation of a new electronic algorithm with an existing, intranet-based surveillance system called Surveillance Assistant (SA) that was internally developed and deployed at BJC in 2009.

Hospital Setting

BJC is a large, nonprofit healthcare organization affiliated with Washington University School of Medicine (WUSM) that delivers services to the greater Saint Louis metropolitan region. The 12 BJC hospitals provide adult and pediatric care at locations ranging from rural and community hospitals to large, academic institutions.

Barnes Jewish Hospital (BJH) is a 1250 bed, tertiary care academic facility associated with WUSM. There are six intensive care units (ICUs) at BJH with 121 total beds, including medical, surgical, cardiac, cardiothoracic and neurological specialties. Ventilated patients can also be cared for on a bone marrow transplant unit, which is staffed to have up to 8 ventilated patients at a time, a step-down long term vent unit with 10 beds and an advanced heart failure unit with 6 beds. On average, there are a little over 1,500 ventilator days each month at BJH.

Methods

The first step in developing the electronic algorithm for VAE surveillance was to analyze the NHSN specification. Based on CDC's NHSN VAE protocol, a VAC represents an episode of sustained respiratory deterioration, caused by both infectious and non-infectious conditions and complications occurring in mechanically-ventilated patients. A VAC is defined by a sustained period of worsening oxygenation that immediately follows a baseline period of stability or improvement on the ventilator. To meet the VAC definition, a mechanically-ventilated patient must have at least two calendar days of stable or decreasing daily minimum positive end-expiratory pressure (PEEP) or fraction of inspired oxygen (FiO2) followed by at least 2 days of increased daily minimum PEEP or FiO2. The increase in the daily minimum PEEP must be ≥ 3cm H2O or an increase in the daily minimum FiO2 of ≥ 0.20 (20 percentage points in oxygen concentration) than the daily minimums during the baseline period.

An IVAC is defined as a VAC in which the patient has either a temperature or white blood cell count outside the expected ranges provided by NHSN and a new, eligible antimicrobial agent started and continued for four or more calendar days. Both of these triggers have to happen within the window period defined as on or after the third day of mechanical ventilation and within 2 days before or after the onset of worsening oxygenation.

Possible VAP is defined by the presence of purulent secretions or a positive lower respiratory tract culture. Probable VAP is defined by the presence of purulent secretions in addition to a positive lower respiratory tract culture meeting certain quantitative or semi-quantitative thresholds of pathogen growth (endotracheal aspirate [ETA], >105 CFU/ml; BAL, >104 CFU/ml; tissue, >104 CFU/g; protected specimen brush [PSB], >103 CFU/ml). The probable VAP definition could also be met based upon the presence of a positive pleural fluid culture, lung tissue with histopathological evidence of infection, or positive diagnostic tests for Legionella or selected respiratory tract viruses, without the concomitant requirement for purulent secretions.

This implementation project was focused on getting the needed data for BJH due to the surveillance burden at that facility. Analysis of VAE rates over the study period were evaluated using linear regression (SPSS V 21.0, IBM SPSS Inc, Armonk, NY).

Development and Implementation Processes

BJC previously developed an enterprise clinical decision support system and repository (CDS) for surveillance and real-time alerting.[4] The Pharmacy Expert Systems database (PES) has been in use since 1994 and receives registration, lab, vital sign, pharmacy, microbiology, and select nursing assessment data through a combination of some real-time and nightly batch interfaces. This data allows for the generation of batched alerts, some real-time alerts, and both prospective and retrospective surveillance monitoring for a wide range of hospital initiatives.

Prior to starting the VAE project, the implementation team did a gap analysis between the current database and the requirements of the NHSN algorithm. The analysis showed that the current CDS contained all of the necessary data for calculating the VAC and IVAC candidates, with the exception of some of the ventilator settings. In late 2012, BJC initiated a new data acquisition project to capture the ventilator settings data documented within the hospital EHR.[5]

In an effort to bring the ventilator settings data into CDS, we expanded a nightly data extract from the source clinical documentation system at BJH to include of all data elements needed for accurate determination of VAC and IVAC cases according to the NHSN specifications. Using the ventilator setting data along with existing data available in CDS (including antibiotics, vital signs, and microbiology results), we were able to implement electronic surveillance for the respiratory status (VAC) and the infection and inflammation components (IVAC) of VAE.

Due to data limitations, the electronic algorithm can only identify patients as having a VAC or an IVAC. For example, while microbiology data is available within the clinical data repository, the specific quantitative results required for the purulent respiratory secretions are not reliably available in a discrete format. It would be possible to detect a positive endotracheal aspirate culture, but it would not be possible to apply the requirement of $\geq 10^5$ CFU/ml

with the existing interface. As a result, the final determination for possible or probable VAP requires clinical review of IVAC patients by IP staff.

The manual and electronic surveillance were performed in parallel for 6 intensive care units, 1 step-down long term ventilation unit, 1 oncology unit and 1 advanced heart failure unit at BJH for 8 months, between January and August 2013. The performance of the electronic algorithm for VAC and IVAC surveillance was manually verified by IP staff. In an automated fashion, the electronic algorithm gathered and summarized daily data for every ventilated patient.

During the assessment period, an Excel report with patient identifiers, mechanical ventilator (MV) data, PEEP (minimum), FiO2 (minimum), temperature (minimum and maximum), WBC (minimum and maximum), and antibiotics was provided to IP for review on an ad-hoc basis. IP staff was blinded to the electronic algorithm's decision of VAC or IVAC status. Every patient was reviewed independently by the two sources for 8 months. The electronic algorithm went through several iterations of refinements to ensure the NHSN definitions were correctly applied. After the initial 8 months of validation were completed, the electronic algorithm was incorporated into the daily work of SA for BJH.

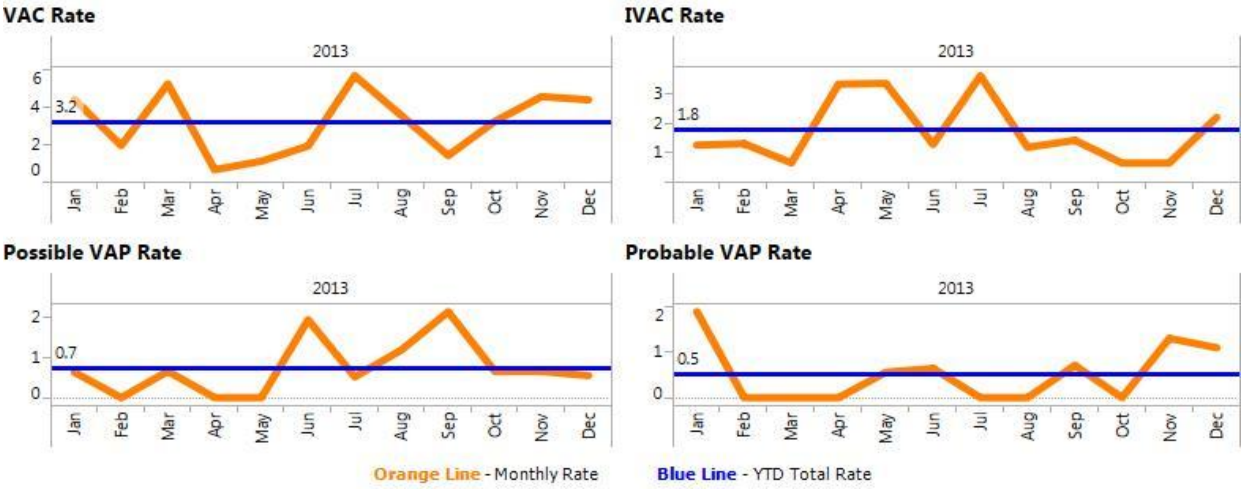
Results

In 2013, there were 3,691 ventilated patient visits, accounting for 19,105 ventilator days at BJH. Of the 3,691 patient visits, 121 were identified with 1 or more VAE events: 62 were VAC (3.2/1,000 ventilator days), 35 were IVAC (1.8/1,000 ventilator days), 14 were possible VAP (0.7/1,000 ventilator days) and 10 were probable VAP (0.5/1,000 ventilator days). VAE rates showed no increasing or decreasing trends over the study period ($p = 0.32$, Figure 1).

Figure 1

Ventilator Associated Event (VAE)

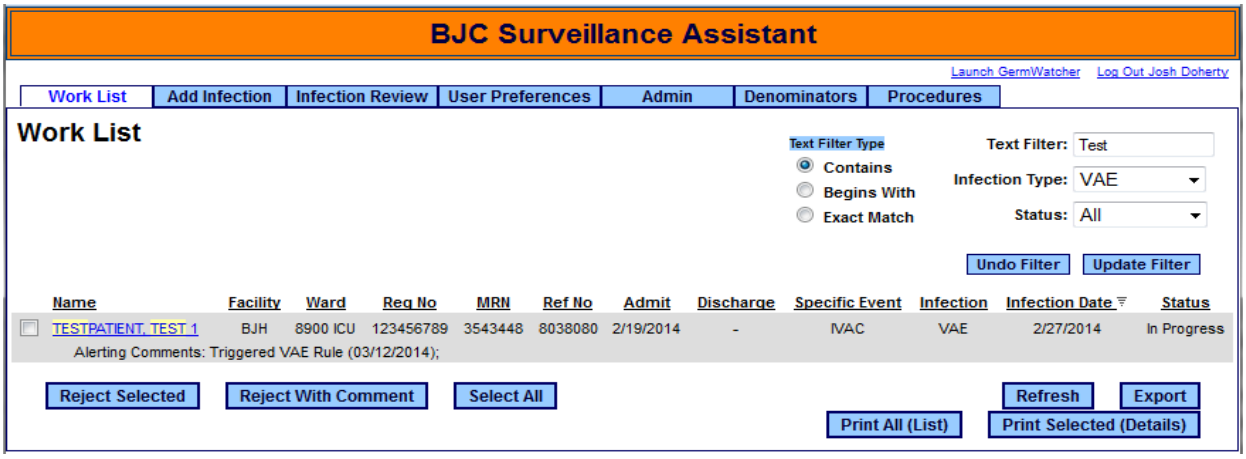
	2013				YTD
	Q1	Q2	Q3	Q4	
All VAE Events:	27	24	37	33	121
VAC	17	6	19	20	62
VAC Rate	3.7	1.3	3.8	4.2	3.2
IVAC	5	13	11	6	35
IVAC Rate	1.1	2.8	2.2	1.2	1.8
Possible VAP	2	3	6	3	14
Possible VAP Rate	0.4	0.6	1.2	0.6	0.7
Probable VAP	3	2	1	4	10
Probable VAP Rate	0.6	0.4	0.2	0.8	0.5
Vent Days	4,645	4,713	4,942	4,805	19,105
Vent Utilization	34%	36%	38%	37%	36%



The electronic algorithm required about 600 hours from an informatics intern to develop, and about 80 hours from a senior informatics analyst to complete. The senior informatics analyst was primarily responsible for a final code review and the logic to incorporate the data into the existing electronic surveillance application. About 175 hours of IP staff time was spent working with developers and validating results. The initial develop time could have been reduced by using a more experienced resource, but the overall cost would have been equivalent given the higher resource cost of a senior informatics analyst compared to an intern.

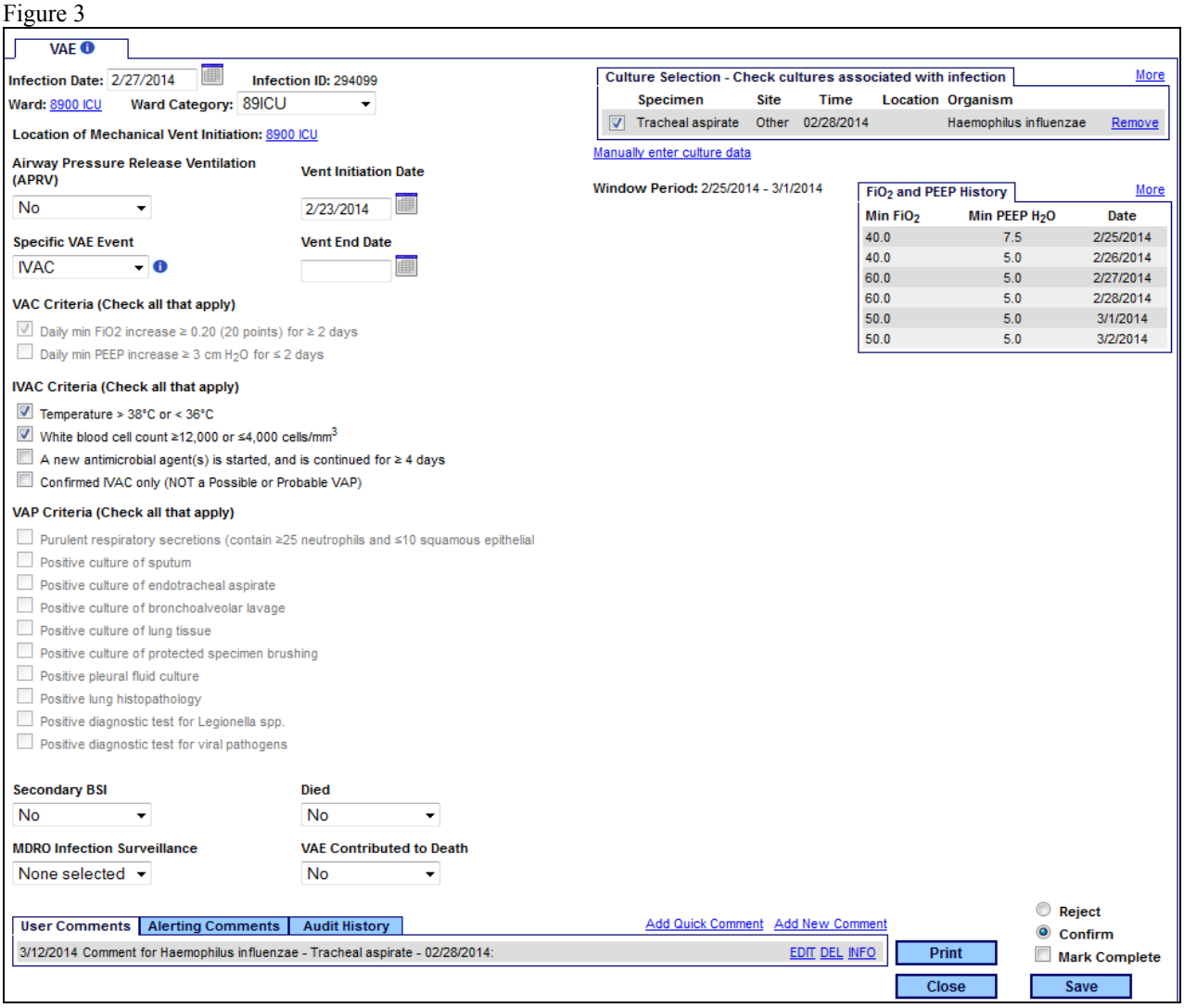
In the 8 months of duplicative review, many tweaks were made to the electronic algorithm to account for intricacies of the process. In the end, the electronic algorithm was able to correctly identify all cases of VAC and IVAC allowing IP staff to then identify possible or probable VAP. After the initial period of reviewing the VAE candidates using a daily report, work was done to incorporate the results of the query into SA that provides the IP staff with a line list that they can review on daily basis (Figure 2).

Figure 2



Initially all new events are loaded into the SA application as either VAC or IVAC. VAC cases are not displayed on the line list since they are not candidates for further classification. IVAC cases must be opened to a detail screen where a workflow module will help the user assess the IVAC case against the possible and probable VAP definitions (Figure 3). The detail page provides the IP staff with as much data as is available in the clinical repository, including microbiology data, room assignments, antibiotics, etc. It also determines the window period, which can change depending on the dates of mechanical ventilation and symptom development.

Figure 3



If the patient meets possible or probable VAP criteria, IP makes the appropriate selections and fills in details marking the case as possible or probable VAP. New VAC and IVAC cases are posted once a day on the SA web interface and IP staff can pull all previously identified cases as well as the new cases. At the time of identification, case information is given to the unit staff where the patient was located. On a monthly basis, rates of each type of VAE are fed back to all areas in a standardized format. All areas are monitored for trends in VAE.

The automated line list report reduced the IP labor efforts for surveillance. The former VAP definition surveillance took approximately 10 hours per ICU per month, equating to about 60 IP hours per month to complete ICU surveillance at BJH. The remaining 3 non-ICU units with ventilated patients generally have low census and surveillance time committed to those unit was not considered in this calculation. It is estimated that completely manual surveillance of the new VAE definition would require the same amount of time (60 hours/month). By manually using the Excel report of components of the VAE definition, that time was reduced to about 2 hours per ICU per month (12 hours/months). The complete automation of the VAE definitions in SA decreased IP time to about 30 minutes per unit per month (3 hours/month), an overall time savings of 57 hours/month.

Discussion

While the electronic VAE surveillance module was successfully deployed at BJH in November 2013, the overall effort represented a significant investment of resources from both the clinical and technical sides of the organization.

The greatest implementation cost was the development and testing of the electronic algorithm. This was in large part due to the complexity of the specifications and the lack of prior experience with the definition. At the time the electronic algorithm was being developed, there was no easy way to validate results outside of a manual review. The development process was very iterative, and questions had to be addressed either by reviewing the specifications or working with our NHSN technical contacts. While the initial investment spent on developing the electronic VAE algorithm and building an infrastructure to accommodate data collection from different hospital systems was significant, BJC is planning to leverage the investment and increase cost savings by implementing the algorithm enterprise wide and rolling it out to the other 9 adult hospitals.

It is important to note that NHSN now offers a web-based interface that takes relevant patient level data (mechanical ventilation dates, FiO2, PEEP values, etc.) and applies the VAC or IVAC criteria. This VAE Calculator provides a quick way to validate the algorithm logic and test out different rule scenarios against a standard. Having this available a year ago would have dramatically reduced the development time, and it has already provided an easier way to perform regression testing when bugs or issues have been uncovered. Furthermore, NHSN has announced that a new web service will be developed that receives de-identified patient data via an XML message. The web service returns an XML result set with the VAC/IVAC classification. While this does replace some of the functionality we have developed, the advantage to hospitals and vendors with the capacity to interact with the web service is clear in terms of development time and future maintenance as definitions change. At the time this paper was written, a prototype of the NHSN's web service was available, but no clear timeline was provided for a production version. Hospitals and vendors still have to do the work to build the required XML message and there has to be some infrastructure for storing the data that is returned. In the case of BJC HealthCare, the results of the web service would be used to populate the same infection tables that populate the detail screens from Figures 2 and 3.

Another significant challenge with the implementation involved getting access to the necessary ventilator data. The pilot hospital involved in the algorithm development already had a daily interface of ventilator data in place prior to the start of the project. This reduced the effort required to capture the additional fields required for the algorithm. In addition, the hospital had a clear champion who worked with the application team to modify the extract with the additional data fields. The interfaces for the other system hospitals require an entirely new interface to be developed. The project team had to meet with subject matter experts at the hospitals to make sure the necessary fields were included in the extract request, and Meaningful Use and other regulatory initiatives have limited access to key technical resources.

The remaining project resource investment was associated with creating the new detail web page in the SA application. The developers worked closely with the IP staff to design the screens with the goal of making it easy to apply the correct definitions. The workflow was considered, and dynamic content was added to reduce the chance of data entry errors. For example, an IVAC cannot be marked as complete unless the specific checkbox is selected that indicates no evidence of a possible or probable VAP.

There have been some concerns raised by those involved in developing and testing these new definitions. One is that healthcare workers may be able to learn to "game the system" by making small, short changes in ventilator settings daily to skew the true picture of a patient's progress. As the team has access to all ventilator setting changes for the patient population, this can be analyzed. One next step for the group is to trend patient-specific ventilator settings, looking for daily outliers.

This development project was strengthened by the fact that the validation included many patients from various medical backgrounds, including several types of intensive care and other critical care units and the oncology population. Due to this fact, and the extensive 8 month period of manual validation of the electronic algorithm, the group hypothesizes that the electronic algorithm is robustly built to correctly identify VAE in a variety of patient populations. However, as this study was completed in a large academic institution, it may not be generalizable to all populations. The application will be applied to the community hospitals that are members of BJC Healthcare, with additional manual validation of the electronic algorithm to ensure generalizability.

To date, the numbers of VAC, IVAC, possible VAP, and probable VAP have been small, and within anticipated ranges. No interventions have been specifically developed in response to this data, however there are ongoing efforts in these units to decrease adverse events associated with ventilator use. As we continue to perform surveillance and gather data, we will be able to better analyze and interpret them for use in intervention development.

Conclusion

BJC HealthCare was able to successfully implement an electronic VAE surveillance algorithm based on the CDC VAE definition and incorporate it into an existing electronic surveillance system at BJH. This saves an estimated 57 hours of IP staff work per month at just one hospital by providing decision support for identification of VAC and IVAC cases and reducing the burden of data collection. The VAE module also helps provide the additional data required to classify an IVAC case as possible or probable VAP cases. The data is stored in a local repository where it can be queried or eventually uploaded to NHSN.

The costs involved in the development and implementation must be carefully weighed against the savings obtained. Even with a robust data infrastructure, there was significant development cost in getting access to the necessary data, creating the electronic algorithm, and modifying the existing web application. In the case of a large multi-hospital system the cost can be easier to justify due to the volume of patients receiving mechanical ventilation and the ability to leverage the initial investment by implementing the electronic algorithm enterprise wide as BJC is planning on doing.

For a smaller hospital without a strong IT and data infrastructure, the costs might well outweigh the benefit. This is especially true when taking into account the new tools that NHSN is providing for hospitals and vendors to use in the classification of the VAC and IVAC cases.

References

1. Centers for Disease Control and Prevention National Healthcare Safety Network. Definitions for Ventilator-associated Events. http://www.cdc.gov/nhsn/acute-care-hospital/vae/index.html (accessed February 24, 2014).
2. Kirtland SH, Corley DE, Winterbauer RH, et al. The diagnosis of ventilator-associated pneumonia: a comparison of histologic, microbiologic, and clinical criteria. Chest 1997;112(2):445-447
3. Tejerina E, Esteban A, Fernández-Segoviano P, et al. Accuracy of clinical definitions of ventilator-associated pneumonia: comparison with autopsy findings. J Crit Care 2010;25(1):62-68.
4. Huang Y, Noirot LA, Heard KM, Reichley RM, Dunagan WC, Bailey TC. Migrating toward a next-generation clinical decision support application: the BJC HealthCare experience. AMIA Annu Symp Proc. 2007;344–348.
5. Resetar E, McMullen KM, McCormick S, Woeltje KF. Development of Electronic Surveillance for Ventilator-Associated Events (VAE) in Adults. AMIA Annu Symp. 2013.

Automatically Classifying Question Types for Consumer Health Questions

Kirk Roberts, PhD, Halil Kilicoglu, PhD, Marcelo Fiszman, MD, PhD,
Dina Demner-Fushman, MD, PhD
U.S. National Library of Medicine, Bethesda, MD

Abstract

We present a method for automatically classifying consumer health questions. Our thirteen question types are designed to aid in the automatic retrieval of medical answers from consumer health resources. To our knowledge, this is the first machine learning-based method specifically for classifying consumer health questions. We demonstrate how previous approaches to medical question classification are insufficient to achieve high accuracy on this task. Additionally, we describe, manually annotate, and automatically classify three important question elements that improve question classification over previous techniques. Our results and analysis illustrate the difficulty of the task and the future directions that are necessary to achieve high-performing consumer health question classification.

Introduction

Questions posed in natural language are an intuitive method for retrieving medical knowledge. This is especially true for consumers, who may both lack medical background knowledge and be untrained in clinical problem solving techniques. The types of questions consumers ask differ as well. Medical question answering systems targeted to professionals often utilize the PICO structure,[1,2] which is particularly useful for comparative treatment questions. Conversely, consumers often ask for general information about a disease they have been diagnosed with, potential diseases associated with their symptoms, the range of potential treatment options, the typical prognosis, and how they might have acquired the disease.[3] Additionally, the most appropriate type of answer for a consumer is qualitatively different than that for a medical professional. The PICO-based methods typically draw answers from the latest medical research, such as that available in Medline®. Since consumers often lack the necessary medical background knowledge, providing them with the latest research may not aid their understanding. Instead, answers should be taken from consumer-oriented resources such as MedlinePlus® and other medical encyclopedias. Because of the differing nature in the type of questions and the most appropriate answers, consumer health question answering systems necessarily diverge from approaches that target medical professionals. In this work, we tackle a critical aspect of consumer question answering: automatically classifying the type of question asked by the consumer. The question type can then be used to identify the most appropriate resource to retrieve answers.

The National Library of Medicine® (NLM®) provides a medical information service, largely targeted toward consumers, as part of its library operations. Confirming the findings of Zhang,[3] the requests NLM receives from consumers contain questions that are different in nature from those posed by medical professionals. The consumer health questions concentrate less on the technical details, and more on general information, such as prognosis, treatment, and symptom information for diseases. Furthermore, consumer health requests often contain multiple questions, with significant amounts of co-reference and ellipses. For example:

- *I have recently been diagnosed with antisynthetase syndrome. Could you please provide me with information on antisynthetase syndrome? I am also interested in learning about prognosis, treatment, and clinical trials.*
- *Although I have not been diagnosed with trimethylaminuria, I have been having a foul odor for about 2 years. Can you tell me more about this condition? How can I be tested? Is there a cure for trimethylaminuria?*

This work is part of a larger project to automatically provide feedback to consumers with appropriate resources for their medical questions. Currently, NLM answers these questions manually, which means that a consumer question might not be answered for several days. By performing this process automatically, consumers can have their questions answered immediately. The process of extracting and representing individual questions from the larger request has already been discussed,[4,5] as well as the handling of co-reference and ellipses,[6] but the difficult problem of accurately classifying the individual consumer health questions has, to our knowledge, not yet been studied.

In this work, we discuss our initial attempt to automatically classify consumer health questions. While many of the questions are not strictly "factoid" questions in the style of the TREC question answering competitions,[7] in

this paper we focus on identifying the structural question elements typically found in factoid questions. Our results demonstrate that recognizing these elements with high accuracy can indeed aid in automatic question classification, but further work is necessary to make significant improvements in classifying consumer health questions. In summary, the primary contributions of this work are:

(i) presenting methods for automatically classifying consumer health questions,
(ii) demonstrating the importance of understanding the key semantic elements of each question, and
(iii) discussing what further work is needed to achieve high accuracy classification of consumer health questions.

Methods

We begin by describing our question types and the data they are annotated on. Next, we describe previous approaches to question classification to provide useful baseline methods. Finally, we describe additional annotations and features proposed in this work for consumer health question classification.

A. Question Types

We classify questions using the following question types. In many respects, the question types are generalizable to non-consumer medical questions. A more complete description of each question type, along with annotation rules and many more examples, can be found in Roberts et al.[8] To illustrate how the question types correspond to encyclopedic sections, for each question type we list the sections of MedlinePlus where answers are most likely to be found (though our system uses several other sources as well).

(1) ANATOMY: Identifies questions asking about a particular part of the body, such as the location affected by a disease. (Answers to ANATOMY questions are typically found in the "Anatomy/Physiology" section.)
   – *Does IP affect any areas inside the body, such as internal organs?*

(2) CAUSE: Identifies questions asking about the cause of a disease. This includes both direct and indirect causes, such as factors that might increase the susceptibility to a disease. ("Common Causes")
   – *Can pregnancy trigger such an issue?*

(3) COMPLICATION: Identifies questions asking about the problems a particular disease causes. This primarily focuses on the risks faced by patients with the disease and does not include the signs/symptoms of a disease. ("Related Issues")
   – *Are carriers of cystic fibrosis at a higher risk for other health conditions?*

(4) DIAGNOSIS: Identifies questions asking for help making a diagnosis. Our question answering system is not designed to provide a direct diagnosis. However, the DIAGNOSIS type does include questions asking about diagnostic tests, or methods for determining the difference between possible diagnoses (differential diagnosis). ("Diagnosis", "Testing")
   – *Can Bloom syndrome be detected before symptoms appear?*

(5) INFORMATION: Identifies questions asking for general information about a disease. This includes type information about diseases, e.g., whether two disease names represent the same disease, or if one disease is a type of another disease. ("Definition", "Description")
   – *Can you please provide me with general information about hyper IgD syndrome?*

(6) MANAGEMENT: Identifies questions asking about the management, treatment, cure, or prevention of a disease. ("Treatment", "Prevention")
   – *Are there new therapies for treatment of pili torti?*

(7) MANIFESTATION: Identifies questions asking about signs or symptoms of a disease. ("Symptoms")
   – *Are there any physical characteristics associated with the disorder?*

(8) OTHEREFFECT: Identifies questions asking about the effects of a disease, excluding signs/symptoms (MANIFESTATION) or risk factors (COMPLICATION). When the question requires medical knowledge to understand a given effect is actually a MANIFESTATION or COMPLICATION, it is instead classified as an OTHEREFFECT. ("Symptoms", "Related Issues")
   – *Can tuberous sclerosis affect eye closure?*

Request:
I have been recently diagnosed with antisynthetase syndrome. Could you please provide me with information on antisynthetase syndrome? I am also interested in learning about prognosis, treatment, and clinical trials.
Decomposed Questions:
Q1) Could you please provide me with information on antisynthetase syndrome?
Q2) I am also interested in learning about prognosis.
Q3) I am also interested in learning about treatment.
Q4) I am also interested in learning about clinical trials.

Table 1: Question Decomposition Example.

(9) PERSONORG: Identifies any question asking for a person or organization involved with a disease. This can include medical specialists, hospitals, research teams, or support groups for a particular disease. Answers to PERSONORG questions are typically not found in MedlinePlus articles, but may be found in links on MedlinePlus pages (especially for support groups).
   – *I would like information about which doctors could treat me.*

(10) PROGNOSIS: Identifies questions asking about life expectancy, quality of life, or the probability of success of a given treatment. ("Expectations")
   – *I would like to know what the life expectancy is for people with this syndrome.*

(11) SUSCEPTIBILITY: Identifies questions asking how a disease is spread or distributed in a population. This includes inheritance patterns for genetic diseases and transmission patterns for infectious diseases. ("Mode of Inheritance", "Prevalence")
   – *Are people of certain religious backgrounds more likely to develop this syndrome?*

(12) OTHER: Identifies disease questions that do not belong to the above types. This includes non-medical questions about a disease, such as requests for financial assistance or the history of a disease. Answers to OTHER questions are typically found in the "Definition" and "Description" sections, though by definition these answers may be found anywhere in the encyclopedic entry.
   – *How did Zellweger Syndrome get its name?*

(13) NOTDISEASE: Identifies questions that are not handled by our question answering system.
   – *Is there any information about what genes are on chromosome 20q12?*

B. Data

As a source for consumer health questions about diseases, we used 1,467 publicly available requests on the Genetic and Rare Diseases Information Center (GARD) website. It may be argued that, since these requests are about rare diseases, they do not constitute a representative sample of disease questions. While this may be true, GARD requests are still appropriate for this task because the aspects of diseases (treatment, prognosis, etc.) that the requests are concerned with are shared between all diseases. Additionally, since there are far fewer online resources for rare diseases, these questions may be more reflective of the types of questions consumers actually ask than other disease question sources.

The GARD requests contain multiple question sentences, and many of the question sentences contain requests for different types of information, thus the questions were annotated for syntactic decomposition as described in Roberts et al.[4] Table 1 illustrates an example of question decomposition. This decomposition process results in 2,937 questions from the 1,467 GARD requests. Additionally, these requests are annotated with a FOCUS, typically one per request, the disease the consumer is interested in learning more about. In Table 1, the FOCUS is *antisynthetase syndrome*. Since our question answering approach primarily utilizes medical encyclopedias, the FOCUS is often the encyclopedic entry, while the question type is the section of the entry in which the answer is found. For cause-and-effect questions, for example, if the FOCUS is the cause, the answer is likely to be found in the effect section, and vice versa.

The process of annotating the thirteen question types on the GARD data is described more thoroughly in Roberts et al.[8] In this paper, we describe the first automatic classification methods on this data. Additionally, we have manually annotated additional elements, described below, not discussed in any of our previous work.

C. Previous Question Classification Approaches

To investigate how classifying consumer health questions differs from previous forms of question classification, we present four previous machine learning-based methods and their corresponding features. These methods provide solid baseline approaches as well as a reliable set of features from which to choose for consumer health question classification. In some cases, as described below, we slightly alter features from their original versions due to either resource availability or appropriateness. In other cases, the original feature description might not have been completely clear, and thus we chose the best interpretation based on our data.

Li and Roth[9] presented the first machine learning method for classifying questions by their answer type. Answer types specify the semantic class of the answer, as opposed to our question types which specify the search strategy (typically, the section of an encyclopedic entry). In many cases, the answer type and question type are identical.

- f_{LR1}: Words. This is the classic "bag-of-words"feature. We use the caseless form of each token
- f_{LR2}: Part-of-speech tags
- f_{LR3}: Phrase chunks
- f_{LR4}: Head chunks. The first noun phrase chunk and the first verb phrase chunk in the question
- f_{LR5}: Named entities. Li and Roth used named entity types such as PERSON and LOCATION, which make less sense for our data. Instead, we use UMLS semantic types
- f_{LR6}: Semantically related words. Here Li and Roth utilized lexicons that would capture similar words such as *actor* and *politician* for the HUMAN:INDIVIDUAL type. Instead, we approximate this feature using the cue phrases from Kilicoglu et al.[6]

Yu and Cao[10] presented a method for classifying into a similar set of classes to our own, but tailored for questions posed by physicians instead of consumers.

- f_{YC1}: Words, the same as f_{LR1}
- f_{YC2}: Stemmed (lemmatized) words
- f_{YC3}: Bigrams
- f_{YC4}: Part-of-speech tags, the same as f_{LR2}
- f_{YC5}: UMLS concepts. This feature is useful to recognize UMLS synonyms
- f_{YC6}: UMLS semantic types, the same as f_{LR5}

Liu et al.[11] presented a method for distinguishing consumer health questions from professional questions. Their features are designed more to capture the lexical tendencies of consumers and professionals against all types of questions, but we utilize them here because their work is one of the first studies in automatically classifying consumer questions.

- f_{LAY1}: Words, the same as f_{LR1}
- f_{LAY2} - f_{LAY4}: Minimum, maximum, and mean word length in the question
- f_{LAY5}: The question length in words
- f_{LAY6} - f_{LAY8}: Minimum, maximum, and mean inverse document frequency from Medline 2010. Instead, we use a 2013 version of Pubmed Central

Patrick and Li[12] presented a method for classifying clinical questions about information within an EHR. The questions in their data were posed by medical staff in an ICU. They classify clinical questions into a taxonomy and a set of template questions to facilitate answer retrieval within an EHR system. Patrick and Li used SNOMED-CT, while we use all of UMLS.

- f_{PL1}: Unigrams, same as f_{LR1}
- f_{PL2}: Lemmatized unigrams, same as f_{YC2}
- f_{PL3}: Bigrams, same as f_{YC3}
- f_{PL4}: Lemmatized bigrams
- f_{PL5}: Interrogative word. Typically, a WH-word (e.g., `what`, `how`) or verb (`is`, `could`)
- f_{PL6}: Interrogative word + next token

- f_{PL7}: UMLS semantic types, same as f_{LR5}. Patrick and Li also use a version of this feature limited to the SNOMED category "observable entity". We do not include this specific feature as it has no clear analogue nor recognizable need in our data
- f_{PL8} + f_{PL9}: verb-subject relations in both the original (f_{PL8}) and lemmatized (f_{PL9}) form. While Patrick and Li use the Enju parser, we use the Stanford dependency parser[13] In the question, "*What treatments have you recommended?*", f_{PL8} would be `nsubj(recommended,you)`, f_{PL9} would be `nsubj(recommend,you)`
- f_{PL10} + f_{PL11}: verb-object relations in both the original (f_{PL10}) and lemmatized (f_{PL11}) form, again using the Stanford dependency parser. From the question above, f_{PL10} would be `dobj(recommended,treatments)`, f_{PL11} would be `dobj(recommend,treatment)`
- f_{PL12} - f_{PL15}: Features f_{PL8} - f_{PL11}, but with the subject/object being replaced by `UMLS` if it is a UMLS concept
- f_{PL16} - f_{PL19}: Features f_{PL8} - f_{PL11}, but with the subject/object being replaced by its semantic type if it is a UMLS concept
- f_{PL20}: WordNet synonyms and antonyms of key terms found by analyzing the data. We approximate this by using a WordNet-expanded version of f_{LR6}

Patrick and Li employ three classifiers that each utilize a sub-set of these features. Their unanswerable question classifier (UQC) uses f_{PL2}, f_{PL13}, and f_{PL15}. Their answerable question taxonomy classifier (QTC) uses f_{PL2}, f_{PL13}, f_{PL15}, and f_{PL20}. Their genetic template classifier (GTC) uses f_{PL2}, f_{PL5}, f_{PL6}, f_{PL7}, and f_{PL8}.

D. Question Stems

The *question stem* is the span of text that introduces the question. It is often referred to as the WH-word or interrogative word, though it need not be a classic WH-word (e.g., "*Are the treatments effective?*") or even a single word. It provides a useful signal for indicating the question structure, while also indicating distributional differences in the question types. For example, a *where* question is far more likely to have a PERSONORG or ANATOMY type than a PROGNOSIS or CAUSE type. Examples of question stems are:

- **What** *may be the underlying cause?*
- *At* **what** *age progressive hemifacial atrophy typically present?*
- *If she needs hormonal treatment,* **which** *medication may be the safest choice?*
- **Could** *you please let me know the name of the blood test used to diagnose a vitiligo carrier?*

Not all WH-words are question stems, however, as it depends upon their syntactic function:

- **Are** *there any other recommended treatments other than what my doctor has already tried for this condition?*
- **Have** *any side effects been reported in patients who use these medications?*

We annotated all 2,937 questions with a single-token question stem. 99.2% of questions were judged to have a question stem, and in 76.6% of these the question stem was the first word. The ten most common question stems in the GARD data are *what*, *is*, *how*, *can*, *are*, *if*, *does*, *do*, *looking*, and *could*.

We use a two-step machine learning method for automatically recognizing question stems. First, a question-level SVM determines whether or not a question has a question stem using a bag-of-words model where the first word is intentionally cased and other words are lower-cased. Second, a token-level SVM is used to rank the words in the sentence, with the top-ranked word being chosen as the question stem. This ranker uses the current word (cased as before), lemma, part-of-speech, previous word, and next word as features.

Once recognized, the question stem allows for several useful features during question classification.

- f_{QS1}: Uncased Question stem
- f_{QS2}: Uncased question stem plus the next token
- f_{QS3}: Uncased question stem plus the next two tokens
- f_{QS4}: Uncased question stem plus the next three tokens
- f_{QS5}: Whether or not the question stem is in a pre-defined set of boolean stems
- f_{QS6} + f_{QS7}: Question tokens (uncased + stemmed) ignoring those tokens before the question stem

- f_{QS8} + f_{QS9}: Versions of f_{QS6} and f_{QS7}, respectively, that account for word order after the question stem
- f_{QS10} + f_{QS11}: Similar to f_{QS6} and f_{QS7}, but only over the range of the WHNP in the syntactic parse tree
- f_{QS12} + f_{QS13}: Similar to f_{QS6} and f_{QS7}, but only over the range of the clause in the syntactic parse tree

E. Answer Type Terms

The *answer type term* is the noun phrase that specifies the expected answer type. Due to its importance, a significant amount of question answering research has focused on describing and automatically identifying answer type terms.[14,15] In the traditional TREC-style factoid questions, it is typically found immediately after the question stem *what* in one of several predictable syntactic positions:

- *What* **disease** *does she have?*
- *What* **symptoms** *usually occur?*
- *What is the* **prognosis** *for this disease?*

In our data, however, answer type terms may occur in less predictable locations, or be syntactically dominated by a more general word:

- *Could you tell me some of the* **symptoms** *of cardiomyopathy hypogonadism collagenoma syndrome?*
- *How can I obtain information about* **treatment options** *for FIBGC?*

Additionally, since our interest is in question types instead of answer types, we relax the definition from an *explicit* answer type term to an *implicit* answer type term. For instance:

- *Can women have* **symptoms** *of glucose 6 phosphate dehydrogenase (G6PD) deficiency?*
- *Where can I get information on the official* **recommendations** *for pregnant women?*

Neither question above has an explicit answer type term, as their question stems convey the answer type (boolean for *Can*, location for *Where*). However, implicitly these questions are asking for symptoms and recommendations, respectively. This indirect question style is a major characteristic of our data, and reflects a less formal querying style. We thus use implicit answer type terms for classifying questions, which are more semantic in nature than the syntactically predictable explicit answer type terms. Thus our answer type term annotations resemble something closer to our question types and further from Li and Roth's answer types.

We annotated all 2,937 questions with a single-token answer type term. 62.4% of questions were judged to have an answer type term. The ten most common answer type terms in the GARD data are *information*, *treatment*, *symptoms*, *prognosis*, *more*, *treatments*, *chances*, *testing*, *risk*, and *expectancy*.

We use the same two-step process to identify answer type terms as we do question stems. Once recognized, the answer type term allows for many useful features:

- f_{ATT1}: Whether or not an answer type term is present
- f_{ATT2} + f_{ATT3}: The uncased and lemmatized answer type term token
- f_{ATT4}: The WordNet hypernyms of the answer type term token
- f_{ATT5} + f_{ATT6}: The uncased and lemmatized words in the answer type term's noun phrase
- f_{ATT7}: The WordNet hypernyms of the words in the answer type term's noun phrase
- f_{ATT8}: The full noun phrase of the answer type term

F. Answer Type Predicates

Not all questions specify their answer type with a question stem or nominal answer type term. Especially in our data, often the answer type is specified with a verb or adjective. To differentiate these from answer type terms, we refer to the verb or modifier span as the *answer type predicate*. Examples of answer type predicates are:

- *How is this condition* **treated***?*
- *If so, what is the likelihood he could* **pass** *it on to his next child?*
- *Is this a* **genetic** *disorder?*

The first two examples above illustrate verbal answer type predicates, while the third illustrates and adjective answer type predicate. The second example also shows a case where an answer type term (underlined) and an answer type

predicate both exist in the same question. In this case, the answer type predicate is more useful in determining a question type of SUSCEPTIBILITY, but often both are necessary to make a proper decision.

We annotated all 2,937 questions with a single-token answer type predicate. 54.2% of questions were judged to have an answer type predicate. The ten most common answer type predicates in the GARD data are *treated*, *have*, *is*, *diagnosed*, *causes*, *affect*, *genetic*, *cause*, *tested*, and *help*.

We use the same two-step process to identify answer type predicates as question stems and answer type terms. The feature that utilize answer type predicates (f_{ATP1} - f_{ATP7}) correspond exactly with the answer type term features. The only exception is that adjectives in WordNet, unlike verbs and nouns, cannot have hypernyms.

Collectively, we refer to the question stem, answer type term, and answer type predicate as the *key question elements*.

G. Classifier

To automatically classify question types, we utilize a multi-class SVM[16]. Due to the wide variety of features described in this paper, we use an automatic feature selection technique[17] to choose the best sub-set of features. Because many of these features convey redundant information, using every feature would not only be slower, but would result in over-training and a drop in accuracy of 3-4%. Ideally, we would have sufficient annotated data to perform these experiments on a development set and provide a final evaluation on a held-out test set. The main goal of this paper, however, is to explore the utility of various machine learning features on this problem. The feature selection technique relies on a 5-fold cross validation on the full data set using a different view (a shuffled split) of the data from that used in the Results section to ensure more generalizable results. Furthermore, since the key question elements are trained on the same data, we utilize stacking to ensure the relevant features are representative of the automatic output.

In addition to replicating the existing feature sets above, we propose three additional feature sets chosen with automatic feature selection:

- CQT_1: The best features without key question element features: f_{LR1}, f_{PL6}, f_{LR5}, f_{LAY7}, f_{LAY8}, f_{LAY4}
- CQT_2: The best features using the automatic key question elements: f_{QS4}, f_{QS7}, f_{QS10}, f_{ATT4}, f_{ATP3}, f_{ATP8}, f_{LR6}
- CQT_3: The same features as CQT_2 using the *gold* key question elements

Feature set CQT_1 allows us to determine a baseline without any key question element involvement. Feature set CQT_2 provides an expectation of how well our overall method would perform in practice. Finally, feature set CQT_3 allows us to establish a ceiling for the utility of key question elements based on the gold annotations.

Results

The results for our key question element classifiers on a 5-fold cross validation are shown in Table 2(a). Question stem recognition is the easiest task with an F_1-measure of 96.16, largely due to the limited vocabulary and the frequency of the first word acting as the question stem. The answer type term recognizer was the next best at 84.08, likely because the answer type terms are nouns that often occur in specific contexts. Finally, the answer type predicate recognizer has the most difficulty with an F_1-measure of 76.81. The fact that only around half of questions have an answer type predicate, while almost every question has at least one verb or adjective, likely leads to the lower score. While it is certainly possible to improve the automatic recognition of these question elements, below we discuss why dramatic improvements to the accuracy of these methods might not result in similar improvements to the overall question type classifier.

The results for question classification are shown in Table 2(c). The baseline bag-of-words model performs at 76.9%. The previous question classification techniques barely out-perform the baseline, with one system even under-performing this baseline. Feature set CQT_1 demonstrates slight improvements (an increase of 0.6 points) can be made over Yu and Cao's feature set by automatic feature selection. The key question element features (CQT_2), however, show a larger improvement of 2 percentage points. When gold question elements are used (CQT_3), this improves to 4 points over Yu and Cao's method and 5.5 points over the bag-of-words baseline. This ceiling, however, appears fairly low considering we used gold question elements. Originally this was thought to be a result of annotation inconsistency, but upon further analysis several interesting sources of error emerge, discussed below. Finally, Table 2(b) provides details of how well CQT_3 classifies each question type. The F_1 scores are largely reflective of the imbalances in the

	Precision	Recall	F_1
Question Stems	95.97	96.36	96.16
Answer Type Terms	84.47	83.69	84.08
Answer Type Predicates	76.90	76.71	76.81

(a) Results for automatic detection of key question elements.

Question Type	# Annotations	Precision	Recall	F_1
Anatomy	12	66.7	16.7	26.7
Cause	119	83.0	78.2	80.5
Complication	32	65.4	53.1	58.6
Diagnosis	229	83.1	75.1	78.9
Information	520	86.3	93.7	89.9
Management	673	91.4	89.7	90.6
Manifestation	103	87.3	86.4	86.8
NotDisease	16	20.0	6.2	9.5
OtherEffect	275	64.7	66.5	65.6
Other	38	63.2	31.6	42.1
PersonOrg	128	87.1	78.9	82.8
Prognosis	313	78.9	79.9	79.4
Susceptibility	420	78.0	86.0	81.8

(b) Detailed question classification results by question type using CQT_3 feature set.

	Accuracy
Bag-of-words	76.89%
Li and Roth (2002)	77.45%
Yu and Cao (2008)	78.43%
Liu et al. (2011)	76.37%
Patrick and Li (2012) – UQC	77.41%
Patrick and Li (2012) – QTC	77.76%
Patrick and Li (2012) – GTC	77.76%
CQT_1	79.01%
CQT_2	80.40%
CQT_3	82.42%

(c) Results for automatic classification of question types.

Min # Elements	# Questions	Accuracy
1	2,814	82.16
2	2,606	84.92
3	2,481	86.50
4	2,353	87.46
5	2,286	88.15
10	1,919	90.20
25	1,538	92.39

(d) Sparsity Experiment. Demonstrates how accuracy of CQT_3 model increases as questions with uncommon question elements are removed.

Table 2: Experiments

data, though OTHEREFFECT appears particularly difficult while PERSONORG, CAUSE, and MANIFESTATION appear particularly simple relative to their relative frequencies in the data.

Discussion

The fact that using gold question elements only raised the score slightly is somewhat surprising. These are the elements of a question that humans intuitively look for while annotating. The semantic gap between answer types and question types does not fully explain the 17.6% error rate when using gold elements. We performed a detailed error analysis to understand the major types of errors made when using gold elements.

The first major type of error involves word sense ambiguity, where a question element could have multiple possible interpretations. For instance:

- *My lower back doesn't seem to work, and I wonder if I will ever be able to* **run**. PROGNOSIS
- *Does CREST syndrome* **run** *in families?* SUSCEPTIBILITY
- *Can you send me a* **link** *concerning hereditary fructose intolerance?* INFORMATION
- *Is there a* **link** *between MELAS and a person who is not really strong?* OTHEREFFECT

In the first two examples, different senses of the answer type predicate *run* are used. In the first, *run* corresponds to the WordNet sense defined as "move fast by using one's feet". This corresponds to inquiring about a disease's impact on a part of one's lifestyle, which we define as PROGNOSIS. In the second case, *run* corresponds to the WordNet sense meaning "occur persistently". The question is therefore asking whether the disease is passed genetically and is thus a SUSCEPTIBILITY question. In our data, the answer type predicate *run* corresponds more with SUSCEPTIBILITY and thus the first question was mis-classified. The next two examples illustrate different senses of the answer type term *link*. The first *link* refers to a website and should be classified INFORMATION. The second *link* is referring to a causal connection between a disease and a possible effect. This question was annotated as OTHEREFFECT. In our data, the answer type term *link* corresponds more with OTHEREFFECT, and so the first *link* was mis-classified. Clearly, some form of word sense disambiguation (WSD) might prove helpful, but WSD has been shown to be a very difficult and highly domain-dependent task[18]. We therefore leave the task of WSD in consumer health questions to future work.

The word sense problem illustrates a key insight into how consumer health questions differ from professional questions. Consumers often lack the terminological familiarity to pose questions in an unambiguous manner. For example, a health professional might have written the second and fourth questions above as:

Answer Type Terms	information, treatment, symptoms, prognosis, more, treatments, chances, testing, risk, expectancy, research, chance, options, cure, test, people, studies
Answer Type Predicates	treated, have, is, diagnosed, causes, affect, genetic, cause, tested, help, treat, associated, having, rare, inherited

Table 3: Answer type terms and answer type predicates with at least 25 instances in our data.

- *Is CREST syndrome* **hereditary***?*
- *Is physical weakness a* **manifestation** *of MELAS?*

In these questions, the key question elements are entirely unambiguous, and should be easily recognized by automatic classifiers. Instead, ambiguous words like *run* and *link* are sufficiently common in the training data to result in misclassification of questions where those words might have actually been the best choice of terminology. Note that this is a different terminological problem from that discussed by McCray et al.[19] In their work, terminology issues arose from the way diseases were specified. For question classification, the disease itself is largely unimportant since we are trying to classify which aspect of the disease the user is interested in.

A second important type of error when using gold question elements has to do with data sparsity: if a question element only appears once or twice, there is not sufficient evidence for its proper question type. For instance:

- *I'm looking for a* **dermatologist** *in my area who has experience with this condition.* PERSONORG
- *What is the* **remedy** *of mixed connective tissue disorder?* MANAGEMENT

While clearly unambiguous, *dermatologist* and *remedy* each only appear once in our data. Resources like WordNet can be utilized to recognize a *dermatologist* is a *doctor*, and a *remedy* is a *cure*, and indeed f_{ATT4} does just this. Yet hypernym features also introduce a good deal of noise, and often aren't highly trusted by the classifier. We can explore the impact of sparsity on question classification by simply removing questions with uncommon elements. Table 2(d) shows this experiment. It is relatively safe to say that having at least 10 occurrences of a question element in the data removes the effects of sparsity, in which case the error rate is reduced by almost half.

Finally, the sparsity experiment provides a useful mechanism for exploring the role of the semantic gap between an answer type (the form or class of an answer) and the question type (which is more related to the topic). Table 3 lists the elements that occur at least 25 times in the data. Clearly, some are ambiguous (e.g., *have*, *options*), but many of the seemingly unambiguous words do not correspond to one specific question type. This is due to the semantic difference between answer types and question types. For example, the answer type predicate *tested* is used in 18 DIAGNOSIS, 12 SUSCEPTIBILITY, and 2 PERSONORG questions. In each case *tested* is being used in the same sense. Furthermore, the answer type term *people* is used in 11 SUSCEPTIBILITY, 9 OTHEREFFECT, 6 PERSONORG, 2 MANAGEMENT, and 2 PROGNOSIS questions. For instance:

- *How can I be* **tested** *for this condition?* DIAGNOSIS
- *Who in the family needs to be* **tested** *for the carrier gene for MLD?* SUSCEPTIBILITY
- *How many* **people** *are affected by Alexander disease?* SUSCEPTIBILITY
- *I would like to contact other* **people** *with epidermolysis bullosa acquisita.* PERSONORG

The first use of *tested* is in a question asking for a diagnostic test, and thus a DIAGNOSIS question. The second question asks *who* requires testing because they are genetically susceptible. In the third question, *people* is being used as a unit of measure for prevalence (SUSCEPTIBILITY), while the fourth question uses *people* to imply other sufferers (PERSONORG). As could be seen in these examples, in order to recognize question types with much higher accuracy than the systems presented here, methods will need to be developed to understand the role an answer type plays in determining the consumer health question's class.

Conclusion

We have presented a method to automatically classify consumer health questions into one of thirteen question types for the purpose of supporting automatic retrieval of medical answers. We have demonstrated that previous question classification methods are insufficient to achieve high accuracy on this task. Additionally, we described, annotated, and classified three important question elements that improve question classification over previous techniques. Our results showed small improvements on a difficult task. We concluded by motivating importance of overcoming word sense ambiguity, data sparsity, and the answer type/question type semantic gap for future work.

Acknowledgements This work was supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health.

References

1. Dina Demner-Fushman and Jimmy Lin. Answering Clinical Questions with Knowledge-Based and Statistical Techniques. *Computational Linguistics*, 33(1):63–103, 2007.
2. Connie Schardt, Martha B. Adams, Thomas Owens, Sheri Keitz, and Paul Fontelo. Utilization of the PICO framework to improve searching PubMed for clinical questions. *BMC Medical Informatics & Decision Making*, 7(16), 2007.
3. Yan Zhang. Contextualizing Consumer Health Information Searching: An Analysis of Questions in a Social Q&A Community. In *Proceedings of the 1st ACM International Health Informatics Symposium*, 2010.
4. Kirk Roberts, Kate Masterton, Marcelo Fiszman, Halil Kilicoglu, and Dina Demner-Fushman. Annotating Question Decomposition on Complex Medical Questions. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 2598–2602, 2014.
5. Kirk Roberts, Z, Halil Kilicoglu, Marcelo Fiszman, and Dina Demner-Fushman. Decomposing Consumer Health Questions. In *Proceedings of the 2014 BioNLP Workshop*, pages 29–37, 2014.
6. Halil Kilicoglu, Marcelo Fiszman, and Dina Demner-Fushman. Interpreting Consumer Health Questions: The Role of Anaphora and Ellipsis. In *Proceedings of the 2013 BioNLP Workshop*, pages 54–62, 2013.
7. Ellen M. Voorhees. Overview of TREC 2004. In *Proceedings of the Thirteenth Text Retrieval Conference*, 2004.
8. Kirk Roberts, Kate Masterton, Marcelo Fiszman, Halil Kilicoglu, and Dina Demner-Fushman. Annotating Question Types for Consumer Health Questions. In *Proceedings of the Fourth LREC Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing*, 2014.
9. X. Li and Dan Roth. Learning question classifiers. In *Proceedings of the 19th International Conference on Computational Linguistics*, 2002.
10. Hong Yu and YongGang Cao. Automatically Extracting Information Needs from Ad Hoc Clinical Questions. In *Proceedings of the AMIA Annual Symposium*, 2008.
11. Feifan Liu, Lamont D. Antieau, and Hong Yu. Toward automated consumer question answering: Automatically separating consumer questions from professional questions in the healthcare domain. *Journal of Biomedical Informatics*, 44(6), 2011.
12. Jon Patrick and Min Li. An ontology for clinical questions about the contents of patient notes. *Journal of Biomedical Informatics*, 45:292–306, 2012.
13. Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, 2006.
14. Vijay Krishnan, Sujatha Das, and Soumen Chakrabarti. Enhanced Answer Type Inference from Questions using Sequential Models. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 2005.
15. Zhiheng Huang, Marcus Thint, and Zengchang Qin. Question Classification using Head Words and their Hypernyms. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 927–936, 2008.
16. Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.
17. Pavel Pudil, Jana Novovičová, and Josef Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, 15:1119–1125, 1994.
18. Roberto Navigli. Word sense disambiguation: A survey. In *ACM Computing Surveys*, volume 41, pages 1–69, 2009.
19. Alexa T. McCray, Russell F. Loane, Allen C. Browne, and Anantha K. Bangalore. Terminology Issues in User Access to Web-based Medical Information. In *Proceedings of the AMIA Annual Symposium*, pages 107–111, 1999.

Using Arden Syntax to Identify Registry-Eligible Very Low Birth Weight Neonates from the Electronic Health Record

Indra Neil Sarkar, PhD, MLIS[1], Elizabeth S. Chen, PhD[1], Paul T. Rosenau, MD, MS[1,2], Matthew B. Storer[1], Beth Anderson, MBA[3], Jeffrey D. Horbar, MD[1,3]
[1]University of Vermont, Burlington, VT; [2]Vermont Children's Hospital at Fletcher Allen Health Care, Burlington, VT; [3]Vermont Oxford Network, Burlington, VT

Abstract

Condition-specific registries are essential resources for supporting epidemiological, quality improvement, and clinical trial studies. The identification of potentially eligible patients for a given registry often involves a manual process or use of ad hoc software tools. With the increased availability of electronic health data, such as within Electronic Health Record (EHR) systems, there is potential to develop healthcare standards based approaches for interacting with these data. Arden Syntax, which has traditionally been used to represent medical knowledge for clinical decision support, is one such standard that may be adapted for the purpose of registry eligibility determination. In this feasibility study, Arden Syntax was explored for its ability to represent eligibility criteria for a registry of very low birth weight neonates. The promising performance (100% recall; 97% precision) of the Arden Syntax approach at a single institution suggests that a standards-based methodology could be used to robustly identify registry-eligible patients from EHRs.

Introduction

The increased adoption of Electronic Health Record (EHR) systems for the management of patient data has largely been touted for the ability to improve health care outcomes[1-4] as well as to support research endeavors[5-7]. An essential first step in the enablement of EHR data to support research is the identification of patient cohorts that match a specified set of criteria[8,9]. Myriad approaches that leverage healthcare standards have been described for identifying such types of patient cohorts within EHR systems to serve as subjects of prospective clinical trials[8,10]. By contrast, there has been limited description of such types of vendor-agnostic approaches that may also be used to populate condition-specific registries from EHR systems.

Condition-specific registries provide a population-level view of retrospective data that may originate from clinical encounters[11,12]. Approaches that may be used for identification of eligible patients for clinical trials could potentially be used for identifying patients that fit a specified set of criteria for inclusion in a registry. Previous work has demonstrated the potential to develop automated approaches for identifying patients eligible for clinical trials based on data that are available in an EHR system[13]. The success of such approaches may also support the development of systems that can further automate the process of populating registries, taking advantage of data that are available within contemporary EHR systems.

Arden Syntax, which dates back to 1989, is a Health Level 7 (HL7) maintained standard for representing clinical knowledge such that it may be used to support clinical decision-making[14]. As an HL7 standard, it is regularly updated and supported by a formal working group that oversees the advancement of the standard in accordance with HL7 processes[15,16]. By maintaining algorithms in self-contained Medical Logic Modules (MLMs), Arden Syntax provides a means to share decision-making rules independent from technical implementation across institutions or environments[17,18]. MLMs are organized into three major categories (*maintenance*, *library*, and *knowledge*) that store information into "slots."[18] The *maintenance* and *library* categories have slots for metadata associated with the management (*maintenance* category) and categorization (*library* category) of a given MLM. The *knowledge* category contains slots that are used for representing the actual clinical knowledge. For example, the *data* and *logic* slots within the *knowledge* category define the variables that will be used within the MLM (*data* slot) and the procedural logic that is required for representing the clinical knowledge for the MLM (*logic* slot).

For an MLM written in Arden Syntax to function, it requires that the source data conform to a usable format. This has historically required custom interfaces for accessing EHR-based data[17,19-21]. By contrast, the secondary use of EHR-based data for research purposes commonly involves the extraction of data into research data repositories. Perhaps the most successful system demonstrating the value of EHR extracted data for secondary use is the Informatics for Integrating Biology and the Bedside (i2b2) platform, which provides a standard and portable interface for browsing health data that may have originated from an EHR[22]. The use of external frameworks, such as

i2b2, therefore helps position research-motivated clinical enterprises to utilize health data for secondary uses[5]. There remain other contexts, however, where the incorporation of such external systems into the health data ecosystem of a clinical enterprise is not feasible or perhaps even permissible. To address this challenge, many EHR vendors provide a "reporting database" that enables one to query data from the EHR using Structured Query Language (SQL). Similar to how previous studies have demonstrated the potential to leverage Arden Syntax for identifying clinical trial patient cohorts from EHR data[23], there may be an opportunity to use similar techniques to identify registry-eligible patient populations.

The Vermont Oxford Network (VON) is a non-profit collaboration that gathers and enables the study of neonatology data from over 900 Neonatal Intensive Care Units (NICUs) that span the globe[24,25]. The data are gathered for neonates from VON members that meet specified eligibility criteria into de-identified registries that have been used to support a range of activities, including quality improvement projects, clinical trials, and outcomes research[25]. VON members each develop a process for identifying eligible neonates according to specified VON criteria and provide data systematically using common formats or interfaces that are maintained by VON. The source systems increasingly include healthcare enterprises that utilize an EHR for primary clinical data gathering and analysis. Due to the potential range of EHR options, including both vendor and homegrown systems, there is motivation to develop a systematic and uniform process to identify eligible neonates whose data may be contributed into the VON registries.

This study explored the potential to leverage Arden Syntax for identifying eligible patients from the EHR at an academic health center EHR. Here, the process for adapting Arden Syntax to be used for identifying cohorts of eligible records from contemporary EHR systems with reporting interfaces is described. The performance of the approach is quantified based on an evaluation relative to a reference standard that included previously identified records for a VON registry. The promising results provide the motivation for developing a comprehensive data abstraction tool that can help automate the process of populating condition-specific registries.

Materials and Methods

The overall goal of this study was to explore the potential of using Arden Syntax for representing eligibility criteria to identify patients whose data might be included in a condition-specific registry. The system and approach were developed and evaluated in accordance with a human subjects protocol that was reviewed and approved by the Committee of Human Research in the Medical Sciences at the University of Vermont. The proposed approach was implemented using Java, making use of the ANother Tool for Language Recognition (ANTLR[26,27]) parser generator to enable the processing of Arden Syntax. The evaluation of the system was carried out using data from Fletcher Allen Health Care (FAHC), the clinical affiliate of the College of Medicine at the University of Vermont.

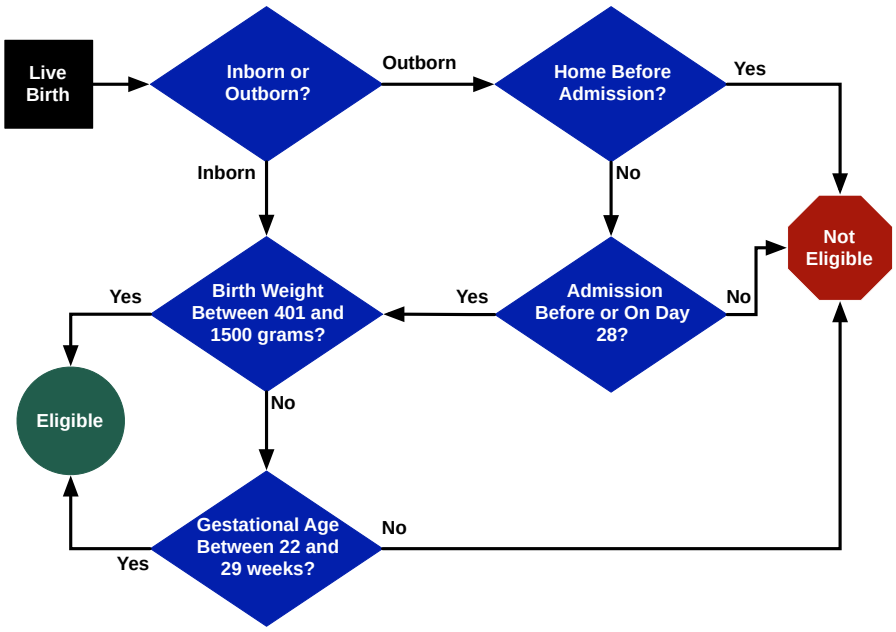


Figure 1: VON VLBW Eligibility Criteria.

VON VLBW Eligibility Criteria. This study focused on the VON Very Low Birth Weight (VLBW) registry, which gathers information on approximately 85% of all VLBW infants born in the US each year. The VON VLBW criteria for eligibility are graphically depicted in Figure 1. Briefly, the registry is focused on gathering data for live-born infants, where status of life is defined as a neonate who breathes or has any evidence for living (e.g., heart beating, umbilical cord pulsation, or definite voluntary muscle movement). Neonates are then categorized into two groups: (1) *inborn* (where the birth occurs within a particular hospital) and (2) *outborn* (where the birth occurs outside this hospital but is admitted within 28 days of birth without first having gone home). Additional criteria for inclusion into the VON VLBW registry are defined by birth weight (between 401 and 1500 grams) or gestational age (between 22 and 29 weeks).

Identifying Eligibility Data Fields in the EHR. The framework for the approach developed in this study utilized a reporting database that contains data extracted from the EHR. First, all the data elements required for determining eligibility within the MLM were manually identified from within the reporting database (i.e., date of birth, hospital name, gestational age, birth weight, admission date/time, and admission source). The chosen data fields were determined based on a manual review of the available data fields in the EHR and a validation of the chosen data fields relative to the FAHC obstetric and neonatology workflows that have been in place for more than a decade. A SQL statement was then created that retrieved all the data elements in a single query from the FAHC EHR reporting database (equivalent to generating a 'report' that included the required data elements).

Representing VON VLBW Eligibility Criteria in Arden Syntax. Within the VLBW MLM *data* slot, a patient record object was created for storing the eligibility data elements needed as well as links to other data needed for retrieving additional information about the patient (i.e., patient identifier). The object was then populated using the aforementioned SQL statement. The eligibility logic as represented by the flow diagram shown in Figure 1 was then encoded into the *logic* slot of the VLBW MLM using Arden Syntax. Conversions between local system units to those used in the eligibility criteria were also done within the *logic* slot (e.g., conversion of the birth weight from imperial ounces to metric grams). To ensure processing of all records retrieved from the EHR, a default categorization of "unknown" was added for patients for whom inborn/outborn status could not be determined (the first step in the eligibility determination as shown in Figure 1 after birth); these patients were automatically deemed not eligible.

Developing a System to Process the VLBW MLM. After eligibility criteria were encoded into the MLM, a Java-based system was developed to interact with the FAHC EHR reporting database using the jTDS[28] Java database connectivity application programming interface. Eligibility status for records subjected to the VLBW MLM was then stored into an eligibility database (either "pass"[eligible], "fail"[not eligible], or "unknown"[requires manual review]). For records deemed eligible, an additional Java routine was used to gather electronically available data from the FAHC EHR reporting database using another set of SQL queries. The resulting records were then transformed into a format that could be transmitted to the VON VLBW registry. The overall approach is shown in Figure 2.

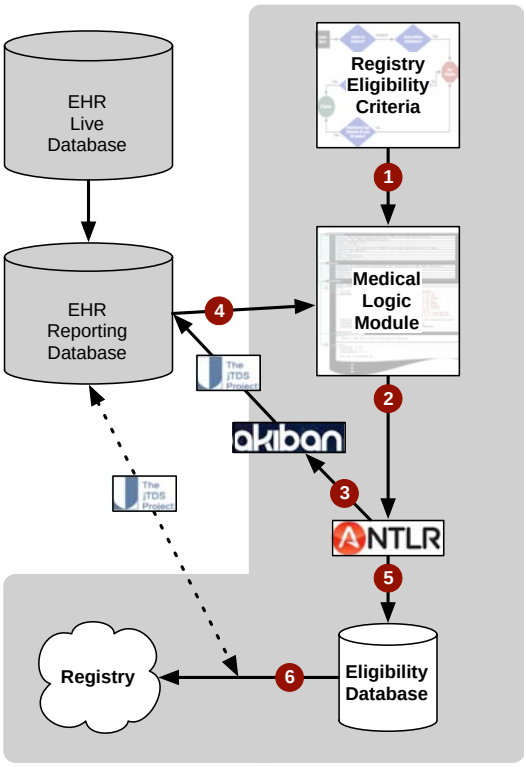


Figure 2: Overview of Approach. (1) Eligibility criteria are manually encoded into an Medical Logic Module (MLM); (2) an ANTLR parser is used to interpret the Arden Syntax; (3) SQL statements are parsed by the Akiban SQL parser and used to send queries to an SQL-92 compliant database using the jTDS Application Program Interface (API), such as an EHR reporting database that uses a relational database model (which is populated from the live EHR database that uses a hierarchical database model); (4) the returned results of the SQL query are then used by the MLM to determine eligibility; (5) eligible records are recorded in a Eligibility Database; (6) the records in the Eligibility Database are enhanced with additional data needed for the registry by additional SQL queries using the jTDS API.

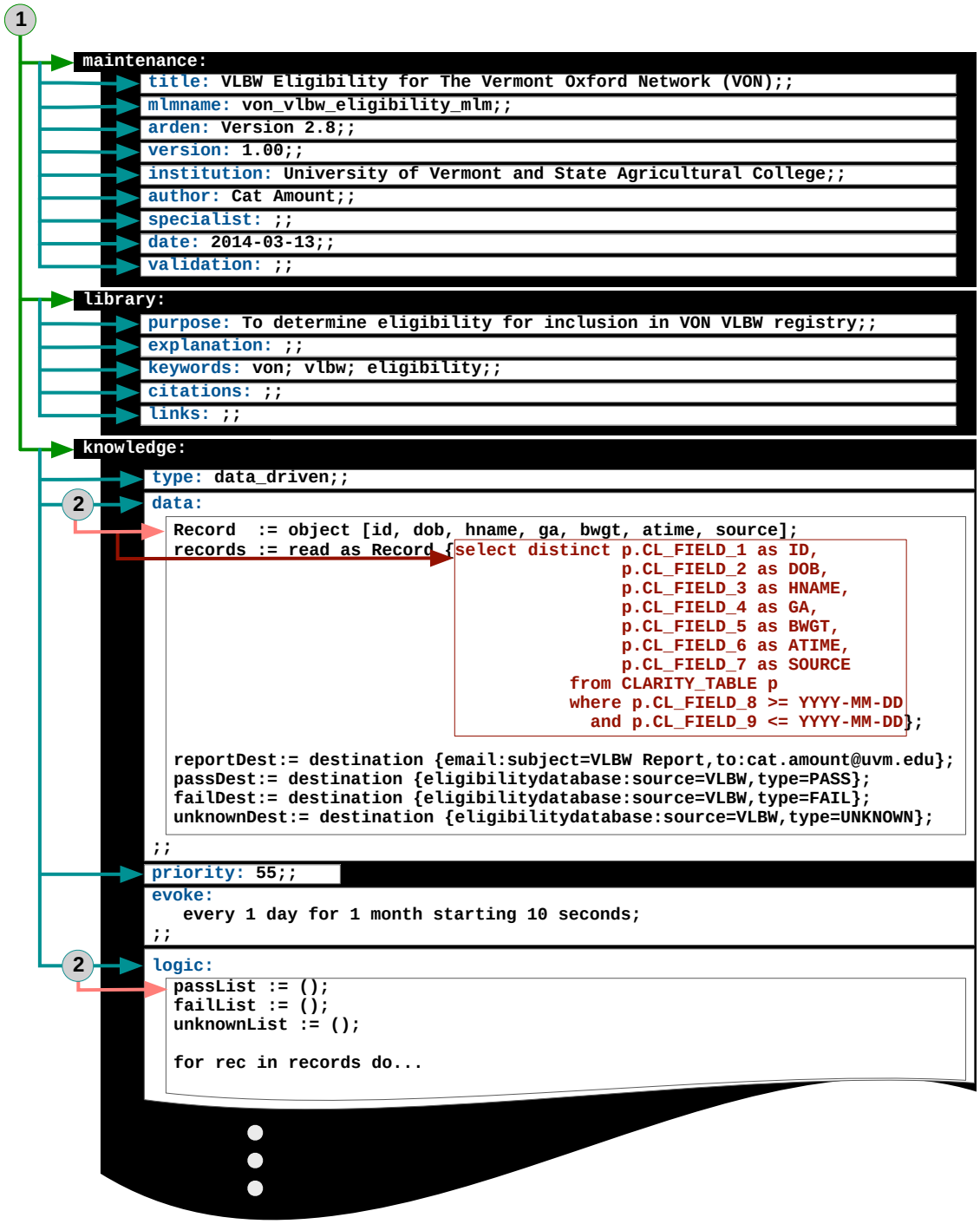


Figure 3: MLM Parsing. A system was developed that (1) used an ANTLR parser generator to identify the MLM categories (maintenance, library, and knowledge) and their contained slots; and (2) used an additional ANTLR parser generator to process the slots and leveraged the Akiban SQL Parser to process SQL statements embedded within the data slot. For this feasibility study, a SQL statement was used to retrieve data from the FAHC EHR reporting database that were used for eligibility determination (ID = Patient identifier; DOB = patient date of birth; HNAME = hospital name; GA = gestational age; BWGT = birth weight; ATIME = admission time; and SOURCE = admission source) according to the criteria depicted in Figure 1.

The actual processing of the VLBW MLM was done in a two-phased process, as depicted in Figure 3. First, an ANTLR parser generator was used to extract the MLM categories and associated slots. ANTLR uses a context-free grammar (expressed in Extended Backus-Naur Form) to recognize language, generate syntax trees, and process those trees to create machine-interpretable actions[26]. An additional ANTLR parser was then used to process the syntax embedded within specific slots (e.g., the data and logic slots). For this study, version 2.8 of Arden Syntax was used as the basis for the parsing grammar. Additionally, the Akiban SQL parser[29] was used to interpret embedded SQL-92 compliant SQL statements. The MLM parsing resulted in a parse tree for which a Java routine was developed to interpret the retrieved data from the source database and apply the encoded eligibility rules.

System Evaluation. A reference standard was generated from VON VLBW registry records associated with eligible records from the time that the Epic EHR at FAHC was fully deployed (2010) to a full complete calendar year prior to the time of this study (2012). The records contained within the VON VLBW registry for this time period were based on manual chart review and guided by a process workflow that has been in place for over a decade. The VLBW MLM was then configured to determine the eligibility of infants born between January 1, 2010 and December 31, 2012. True Positives (TP) were defined as patients who were deemed eligible by the VLBW MLM and in the reference standard; False Positives (FP) were defined as patients who were deemed eligible by the VLBW MLM but not in the reference standard; and False Negatives (FN) were defined as those patients who were in the reference standard but not deemed eligible by the VLBW MLM. The metrics of precision (TP/TP+FP) and recall (TP/TP+FN) were then used to assess the overall performance of the approach.

Results

All of the necessary fields for determining neonate eligibility for inclusion in the VON VLBW registry were identified within the relational data schema used for the FAHC EHR reporting database. The SQL statement that was used to identify the data for all neonates was then embedded within the Arden Syntax within the *data* slot of the *knowledge* category of an MLM. The VON VLBW eligibility criteria were then manually encoded into the MLM (within the *logic* slot of the *knowledge* category). Finally, the instructions for writing the status of a given processed record were defined within the *action* slot of the *knowledge* category.

A Java-based system was developed to interpret MLMs written in Arden Syntax (version 2.8) as well as embedded SQL-92 compliant SQL statements, utilizing a combination of the ANTLR parsing generator and Akiban SQL parser. The developed VLBW MLM was then used to determine the eligibility of patients born between January 1, 2010 and December 31, 2012. In total, the system processed 12,025 neonates who were either born at or transferred to FAHC for that period, of whom the VLBW MLM deemed 192 to be eligible for inclusion into the VON VLBW registry. The evaluation was done relative to the reference standard, which contained 187 neonates that were manually identified and entered in the VON VLBW registry. The total processing time used for the system to determine the eligibility for all newborn records spanning the three-year period was less than five minutes. Of the eligible patients, 187 were in agreement with the reference standard (TP), five were not found in the reference standard (FP), and no patients in the reference standard were missing from the predicted eligible patient list (FN). The overall precision and recall of the system was thus determined to be 97.4% and 100.0%, respectively.

Discussion

The primary motivation for use of Arden Syntax has historically been described in the context of clinical decision support systems[14]. Previous studies have explored other potential applications for Arden Syntax, such as for clinical trial eligibility[23] and facilitating clinical quality studies[30]; however, the majority of systems developed to date for processing Arden Syntax have been done within the context of clinical decision support[31-34]. This study thus reflects a first application of Arden Syntax for the identification of patient cohorts for inclusion in population registries. A recent survey suggests complex clinical trial eligibility representation may require a combination of multiple available formal standards[8]. In this study, the potential to leverage Arden Syntax for identifying registry-eligible patients was explored, where the eligibility rules could be expressed through procedural statements. The promising results suggest that the proposed approach can be used in place of time-consuming and manual processes to identify registry-eligible patients using eligibility criteria that are compatible with machine-encoded logic (e.g., as can be represented in Arden Syntax).

The use of an algorithm-based approach may provide a means to uniformly apply eligibility criteria for reliable data gathering as required for population-level registries that support public health initiatives. Such uniform application of eligibility algorithms may also be used to identify potential challenges in the use of applications. For example, in the evaluation performed for this study, there were five patients identified by the VLBW MLM but not reported as

eligible to the VON VLBW registry (classified in the evaluation as "False Positives"). On closer examination of the data associated with these patients, it was determined that the strict definition of the VON VLBW criteria should have included these individuals. These patients may have been excluded from the registry due to difficult situations (e.g., how are stillborns or planned terminations registered?). While each institution may establish a culture around certain criteria, it can be difficult to ensure consistency when gathering data across many sites. An MLM-driven registry eligibility approach may provide a foundation for high-quality and consistent data acquisition. An interesting future study might be to use the developed VLBW MLM across VON data providers and quantify any possible effect of inclusion/exclusion of certain patient types based on hospital culture. Such a study may help to refine the inclusion/exclusion criteria used in the algorithm to better reflect accepted and consistent practices for a given patient population.

A major potential challenge that has been noted in general deployment of MLMs is the difficulty in accommodating the heterogenic nature of EHR implementations across institutions. The ability to use MLMs is often challenged by the difficulty in completely representing all the local variables needed for the execution. As Arden Syntax specifies that local data variables are to be referenced between a set of curly braces ("{}"), which has led to an referred to as the "curly-brace problem"[35,36]. Amidst some description of systems that are able to directly interact with transactional database systems[37], the majority of previous implementations of MLM driven decision support systems leveraged local relational databases that contained information from EHR systems[36]. Contemporary solutions have also included the translation of Arden Syntax MLMs into intermediate formalisms (e.g., Drools[38]). In the approach presented here, EHR data was mapped through the use of SQL-92 compliant statements to acquire data from an EHR reporting database. This SQL statement, as well as the system-specific databases for tracking pass/fail/unknown records, was the major "curly brace" aspect of the developed approach. The increasing availability of SQL-compliant reporting databases with contemporary EHR systems (in many cases to facilitate the ability to meet Meaningful Use requirements) provides a common interface that allows for the development of an MLM that is built around the processing of a set of records that could be generated as part of an EHR report. Still, the approach described here requires database experience with the local EHR reporting database and the ability to generate a SQL statement that contains mappings to the required elements for determining eligibility. Additionally, the same challenges that one might have within a reporting context (e.g., identifying which fields are the relevant ones to retrieve) remain the same in the context of the approach developed here. It is also important to underscore that because the context for which Arden Syntax was used here is for the population of a registry that is based on information that is commonly captured within a clinical chart, it was conceivable that many of the requisite data elements would be available within a reporting interface that is updated nightly. This is in contrast to more typical scenarios where MLMs are used for more real-time clinical decision support, and would thus need to interface directly with the live EHR system. Nonetheless, the promising results demonstrate the application of Arden Syntax for retrospective patient eligibility determination and the potential utility for other contexts, such as retrospective determination of patient cohorts that may be eligible for clinical trials or who may need to be systematically alerted about a potential adverse event.

It must be acknowledged that other formalisms may be able to represent the logic for cohort identification using procedural logic. However, a significant reason for exploring the potential utility of Arden Syntax over other possible (generally more business analytic centric solutions, such as Pentaho) formalisms is because it is a maintained HL7 healthcare standard. As an established and maintained healthcare standard, Arden Syntax is likely to be more readily accepted for integration into healthcare environments. Furthermore, it is likely that one could embody the logic for cohort identification using a set of boutique SQL statements that encode the eligibility criteria, but the portability to another institution would likely be challenged by the need to modify a possibly lengthy SQL script to address variations in schema design. This could be especially difficult when considering that different reporting databases may have very different data structures and implementations of SQL dialects that may not necessarily be able to handle the full logic required for eligibility determination. The variation of SQL implementations and customization could then challenge the ability to centrally manage the process for consistent eligibility determination, since each institution will be required to maintain local customizations of the criteria to meet their data structure and SQL environment. In the context of international registries, such as the VON VLBW registry, these technical considerations are important in light of the range of technical capabilities across data contributors. Nonetheless, the minimal technical requirement that the solution presented here does require is the ability to generate a SQL statement that gathers the fields required to determine eligibility. Future studies that directly benchmark standards based (e.g., Arden Syntax) approaches against alternative approaches for patient cohort identification are needed.

The intent of the system described here was two fold: (1) identify eligible patients for inclusion in a registry; and (2) gather electronically available data from the EHR to transmit for eligible patients to the registry. In this study, the primary focus was on exploring the feasibility of developing a system that was able to identify eligible patients. The evaluation was thus focused on the ability to use the VLBW MLM to reliably identify potentially eligible patients for the VON VLBW registry. While not described here, the system is also able to gather additional data beyond the information needed for determining eligibility using SQL statements and export the full set of gathered data into a number of consumable formats (e.g., as eXtensible Markup Language or a Comma-Separated Value file). However, it should be noted that a remaining major challenge is the identification of the needed data elements from within an EHR system that should be targeted for inclusion into a registry. This requires study of the workflow at a given institution for data gathering from the respective EHR and developing appropriate data transformation techniques. One must also consider that many elements within a registry may not be available as discrete elements (e.g., they may be in natural language form) and thus advanced processing (e.g., using natural language processing [NLP]) may be required before data can be appropriately formatted into the requisite format for the target registry. Currently, if a data element required for eligibility is either inaccessible or not interpretable through simple processing (e.g., regular expressions), a record will be labeled as "unknown" and require manual review to override the status. As such, a user interface is being developed that enables manual review and override of the MLM based eligibility determination that also fits the workflow currently used to populate the VON VLBW registry. A major intent of developing the system described in this study in Java was thus to enable the potential leveraging of Java-based resources that are available for such processing (e.g., NLP could be integrated into the system using the Unstructured Information Management Architecture framework[39]).

The positive results of the performed evaluation for the VON VLBW registry using the VLBW MLM support the intention to deploy the developed system into production. It is anticipated that the system can be run in production (and set to run on a defined periodic basis through the *evoke* slot of the *knowledge* category) parallel to the current method for populating the registry. Additional evaluations will be done to compare differentials in patient classifications, and it is anticipated that if similar performance benchmarks as presented here are achieved for a defined period that the system will be used as the primary mode for identification of eligible patients for inclusion in the VON VLBW registry. While it is conceivable that the approach described here could lead to a fully automated system, it is expected that the system in practice will remain semi-automated; that is, the eligibility status of patients will be presented to users through a curation interface that allows for manual override of status before submission to the registry. Further evaluations are also being planned at other institutions that have a similar SQL-compliant reporting environment as leveraged in this study. As part of the process of deploying the process at other institutions, it is expected that a version of the system described here will be made available for other applications. In the meantime, the corresponding author (INS) may be contacted for the most recent version of the system.

Conclusion

Standards traditionally used for clinical decision support may be utilized to support the population of condition-specific registries. In this study, Arden Syntax was explored as a possible health standard for representing eligibility criteria for a neonatology registry. Based on a system developed to interact with the reporting framework associated with a reporting database that is based on data from an EHR, it was found that a Medical Logic Module written in Arden Syntax was able to perform well relative to a reference standard representing three years of registry-eligible patients. The promising results suggest that Arden Syntax may used to represent eligibility criteria and thus provide a mechanism to leverage electronically available health data (e.g., as encapsulated within an EHR) for supporting the population of patient cohort databases such as condition-specific registries.

Acknowledgements

This work was funded by a contract from the Vermont Oxford Network. Beth Anderson and Jeffrey D. Horbar are employees of the Vermont Oxford Network. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Vermont Oxford Network.

References

1. Herrin J, da Graca B, Aponte P, et al. Impact of an EHR-Based Diabetes Management Form on Quality and Outcomes of Diabetes Care in Primary Care Practices. American journal of medical quality : the official journal of the American College of Medical Quality. 2014 Jan 7.

2. Hunt JS, Siemienczuk J, Gillanders W, et al. The impact of a physician-directed health information technology system on diabetes outcomes in primary care: a pre- and post-implementation study. Informatics in primary care. 2009;17(3):165-74.
3. Shekelle PG, Morton SC, Keeler EB. Costs and benefits of health information technology. Evid Rep Technol Assess (Full Rep). 2006 Apr(132):1-71.
4. Chaudhry B, Wang J, Wu S, et al. Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. Annals of internal medicine. 2006 Jun 16;144(10):742-52.
5. Kohane IS. Using electronic health records to drive discovery in disease genomics. Nature reviews Genetics. 2011 Jul;12(6):417-28.
6. Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. Journal of the American Medical Informatics Association : JAMIA. 2013 Dec;20(e2):e206-11.
7. Wilke RA, Xu H, Denny JC, et al. The emerging role of electronic medical records in pharmacogenomics. Clinical pharmacology and therapeutics. 2011 Apr;89(3):379-86.
8. Weng C, Tu SW, Sim I, Richesson R. Formal representation of eligibility criteria: a literature review. J Biomed Inform. 2010 Jun;43(3):451-67.
9. Abhyankar S, Demner-Fushman D, Callaghan FM, McDonald CJ. Combining structured and unstructured data to identify a cohort of ICU patients who received dialysis. J Am Med Inform Assoc. 2014 Jan 2.
10. Cuggia M, Besana P, Glasspool D. Comparing semi-automatic systems for recruitment of patients to clinical trials. International journal of medical informatics. 2011 Jun;80(6):371-88.
11. Collen MF. Computer Medical Databases: The First Six Decades (1950-2010). New York: Springer-Verlag; 2012.
12. Flagg EW, Datta SD, Saraiya M, et al. Population-based surveillance for cervical cancer precursors in three central cancer registries, United States 2009. Cancer causes & control : CCC. 2014 Feb 28.
13. Embi PJ, Jain A, Clark J, Harris CM. Development of an electronic health record-based Clinical Trial Alert system to enhance recruitment at the point of care. AMIA Annu Symp Proc. 2005:231-5.
14. Samwald M, Fehre K, de Bruin J, Adlassnig K-P. The Arden Syntax standard for clinical decision support: experiences and directions. Journal of biomedical informatics. 2012 Aug;45(4):711-8.
15. https://http://www.hl7.org/implement/standards/
16. Kawamoto K, Honey A, Rubin K. The HL7-OMG Healthcare Services Specification Project: motivation, methodology, and deliverables for enabling a semantically interoperable service-oriented architecture for healthcare. J Am Med Inform Assoc. 2009 Nov-Dec;16(6):874-81.
17. Poikonen J. Arden Syntax: the emerging standard language for representing medical knowledge in computer systems. American journal of health-system pharmacy : AJHP : official journal of the American Society of Health-System Pharmacists. 1997 Mar 01;54(3):281-4.
18. Hripcsak G. Arden Syntax for Medical Logic Modules. MD computing : computers in medical practice. 1991 Mar-Apr;8(2):76, 8.
19. Hripcsak G, Johnson SB, Clayton PD. Desperately seeking data: knowledge base-database links. Proceedings / the Annual Symposium on Computer Application [sic] in Medical Care Symposium on Computer Applications in Medical Care. 1993:639-43.
20. Ohno-Machado L, Parra E, Henry SB, Tu SW, Musen MA. AIDS2: a decision-support tool for decreasing physicians' uncertainty regarding patient eligibility for HIV treatment protocols. Proc Annu Symp Comput Appl Med Care. 1993:429-33.
21. Pryor TA. The use of medical logic modules at LDS hospital. Comput Biol Med. 1994 Sep;24(5):391-5.
22. Murphy SN, Mendis ME, Berkowitz DA, Kohane I, Chueh HC. Integration of clinical and genetic data in the i2b2 architecture. AMIA Annu Symp Proc. 2006:1040.
23. Ohno-Machado L, Wang SJ, Mar P, Boxwala AA. Decision support for clinical trial eligibility determination in breast cancer. Proceedings / AMIA Annual Symposium AMIA Symposium. 1999:340-4.
24. http://www.vtoxford.org/.
25. Horbar JD, Soll RF, Edwards WH. The Vermont Oxford Network: a community of practice. Clinics in perinatology. 2010 Mar;37(1):29-47.
26. Parr T. The Definitive ANTLR 4 Reference. Dallas: Pragmatic Programmers; 2013.
27. http://www.antlr.org/.
28. http://jtds.sourceforge.net/.
29. http://www.akiban.com/.

30. Jenders RA. Suitability of the Arden Syntax for representation of quality indicators. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2008:991.
31. Gietzelt M, Goltz U, Grunwald D, et al. ARDEN2BYTECODE: a one-pass Arden Syntax compiler for service-oriented decision support systems based on the OSGi platform. Computer methods and programs in biomedicine. 2012 Jun;106(2):114-25.
32. Karadimas HC, Chailloleau C, Hemery F, Simonnet J, Lepage E. Arden/J: an architecture for MLM execution on the Java platform. Journal of the American Medical Informatics Association : JAMIA. 2002 Jul;9(4):359-68.
33. Kuhn RA, Reider RS. A C++ framework for developing Medical Logic Modules and an Arden Syntax compiler. Computers in biology and medicine. 1994.
34. Jenders RA, Shah A. Challenges in using the Arden Syntax for computer-based nosocomial infection surveillance. Proceedings / AMIA Annual Symposium AMIA Symposium. 2001:289-93.
35. Jenders RA, Corman R, Dasgupta B. Making the standard more standard: a data and query model for knowledge representation in the Arden syntax. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2003:323-30.
36. Jenders RA, Dasgupta B. Challenges in implementing a knowledge editor for the Arden Syntax: knowledge base maintenance and standardization of database linkages. Proceedings / AMIA Annual Symposium AMIA Symposium. 2002:355-9.
37. Liang YC, Chang P. The development of variable MLM editor and TSQL translator based on Arden Syntax in Taiwan. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2003:908.
38. Jung CY, Sward KA, Haug PJ. Executing medical logic modules expressed in ArdenML using Drools. Journal of the American Medical Informatics Association : JAMIA. 2012 Jul;19(4):533-6.
39. http://uima.apache.org/.

Use of Design Science for Informing the Development of a Mobile App for Persons Living with HIV

Rebecca Schnall, PhD, MPH, RN[1], Marlene Rojas, MPH, MD[1], Jasmine Travers, AGNP-C, RN[1], William Brown III, DrPH, MA[2,3], Suzanne Bakken, PhD, RN, FAAN[1,3]

[1]Columbia University, School of Nursing, New York, NY; [2]New York State Psychiatric Institute and Columbia University, HIV Center for Clinical and Behavioral Studies, New York, NY; [3]Columbia University, Department of Biomedical Informatics, New York, NY

Abstract

Mobile health (mHealth) technology presents opportunities to enhance chronic illness management, which is especially relevant for persons living with HIV (PLWH). Since mHealth technology comprises evolving and adaptable hardware and software, it provides many challenging design problems. To address this challenge, our methods were guided by the Information System Research (ISR) framework. This paper focuses on the Design Cycle of the ISR framework in which we used user-centered distributed information design methods and participatory action research methods to inform the design of a mobile application (app) for PLWH. In the first design session, participants (N=5) identified features that are optimal for meeting the treatment and management needs of PLWH. In the second design session, participants (N=6) were presented with findings from the first design session and pictures of existing apps. Findings from the Design Cycle will be evaluated with usability inspection methods. Using a systematic approach has the potential to improve mHealth functionality and use and subsequent impact.

Introduction

Despite advances, The United States (US) Human Immunodeficiency Virus (HIV) epidemic continues to take a heavy toll, which is evidenced by the 1.2 million Americans who are currently living with the disease. New York City, the setting of our study, is the epicenter of the US HIV/AIDS epidemic, accounting for 17.9% of the estimated number of persons living with HIV (PLWH) in the US [1]. The burden of HIV/AIDS is borne disproportionately by a growing number of racial and ethnic minorities[2]. Forty eight percent (48.7%) of new HIV diagnoses are among African Americans and 31.3% were among Latinos [3].

Mobile health (mHealth)

Mobile health (mHealth) is focused on the use of mobile information and communication technologies for healthcare purposes. mHealth aims to support care delivery through meeting information, communication, and documentation needs of patients, clinicians, and other healthcare workers, as well as facilitating health resource monitoring and management [4]. The development and use of mobile technologies to rapidly and accurately assess and modify health-related behavior and biological states has great potential for improving healthcare delivery. Advancements in current mobile technology allow for more memory and data storage, full color graphical interfaces with video capability, wireless access, location awareness, and integration with other computer-mediated technologies. Specifically, the use of mobile technology affords numerous advantages, including reduced memory bias, the ability to capture time-stamped data, and the potential for personalizing and tailoring information in real-time. The ubiquitous nature of mobile technologies in daily life has created opportunities for applications (apps) that were not previously possible. Thus, mHealth presents opportunities to enhance chronic illness management in real-time [5]. In fact, the use of mobile apps and technology has the potential to reduce costs, reduce geographical and economic disparities, and personalize healthcare.

Of particular relevance to this work, the use of mHealth has made a huge impact on communication, access, and information/resource delivery especially among racial and ethnic minority groups [6-8] and offers an ideal venue for meeting targeted healthcare needs for racial and ethnic minorities, as well as youth, who are the fastest growing age group affected by HIV [9]. As the HIV epidemic continues to grow among adolescents and young adults, mHealth technology can address many of the healthcare needs of PLWH including: adherence to HIV medications, prevention with positives, retention in care and self-management, all critical needs for this population. mHealth technology can bridge a divide in healthcare delivery in underserved minority groups, since ownership of a mobile

device is more common among African Americans than among Whites (87% vs. 80%) [7]. African Americans are also among the most active users of the mobile Internet and take advantage of a much greater range of their phones' features than do Whites [10]. This grants the potential to ameliorate health disparities, since in the case of HIV, health disparities are inextricably linked to age, race and ethnicity. Due to the high incidence of HIV among minorities, adolescents and underserved young adults, and their reliance on mobile technology, it is appropriate to develop health information technology (HIT) tools tailored to the needs of this population.

Theoretical Framework

Design science is the development, implementation, evaluation, and adaptation of artifacts for problem solving [11,12] and attempts to focus human creativity into the design and construction of artifacts that have utility for mobile apps. Design activities are central to most applied disciplines, including health informatics, and are particularly relevant and necessary for the development of mHealth technology that is safe, useful, and effective for end-users [13]. mHealth technology is composed of mutable and adaptable hardware, software, and human interfaces and so it provides many unique and challenging design problems that call for new and creative ideas. The design science research paradigm is highly relevant to mHealth research and supports the creation of innovative artifacts to solve real-world problems.

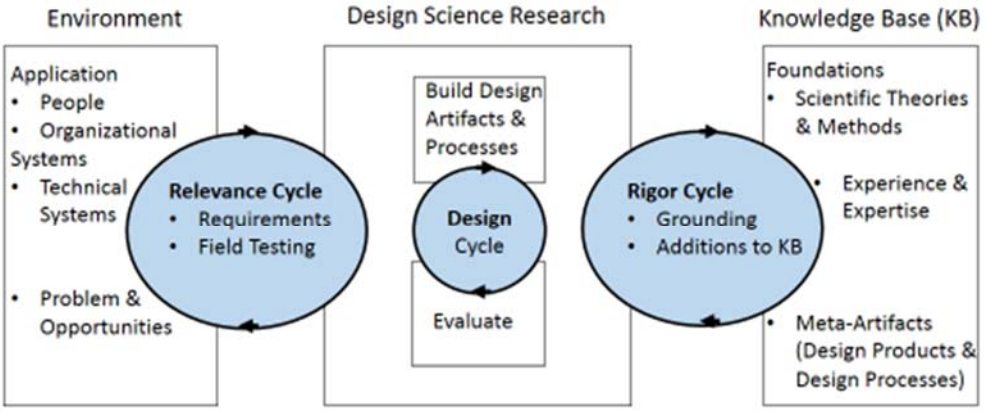


Figure 1: The Information System Research (ISR) Framework

Our study activities were guided by the Information System Research (ISR) Framework, which uses design science to inform the development of information systems [11]. The design and development cycle is repeated iteratively until a desired final product is achieved. Using the ISR Framework (shown in Figure 1), we employed three research cycles: I) the Relevance Cycle in which we seek to understand the environment of the end-user by determining requirements through a series of interactions (e.g., focus groups, interviews) with stakeholders; II) the Rigor Cycle, in which evaluation of theories and artifacts contribute to the design science and application domain knowledge base; and III) the Design Cycle in which artifacts are produced and evaluated [11]. The result of HIT-related research activities informed by the ISR framework is the purposeful creation of artifacts developed to address an important health problem.

Research Context

The activities reported here reflect a single cycle that was part of a larger research project to inform the development of two mobile apps for: 1) HIV treatment and care for PLWH and 2) HIV prevention for high-risk Men who have sex with Men (MSM). Research activities reported here focus on the development of a mobile app for PLWH. To guide this work, we used user-centered information design methods [14] and participatory action research methods [15] to ensure that the app met end-user needs. The overall goal of this two year project was to create a Design Document for the Centers for Disease Control and Prevention (CDC) to inform the development of two mobile apps for dissemination. Prior to the design sessions, we conducted study activities related to the Rigor and Relevance Cycles. The project activities discussed in this manuscript focus on Cycle III: The Design Cycle: Develop/ Build activities

for the mobile app for PLWH only.

Cycle I: The Relevance Cycle: For this cycle, we conducted six focus group sessions with PLWH (N=50) ages 18-59 and three focus group sessions with HIV care providers (N=30) to identify the desired content, features, and function of a mobile app for PLWH. Results of this work are reported elsewhere[16]. Thematic analysis of the focus group session revealed five categories of functional requirements: My Information Management, Managing My Medication, Staying Healthy, Communication (divided into provider communication and peer communication) and Resources. Participants suggested some of the following tools and functionalities for inclusion in a mobile app for meeting their health needs: reminders/alerts, lab tracking, notes, chat boxes/forums, testimonials of lived experiences, personal outreach, games/virtual rewards, coding of health tasks, and simulation on how to disclose their HIV status.

Cycle II: The Rigor Cycle: The rigor cycle guided our selection of theories and methods for designing and evaluating the Design Document [17]. We conducted a systematic review of existing studies, which used mobile technology and e-Health applications for HIV treatment and care. In addition, we did an ecological scan of the existing mobile applications that are already frequently used by many in our study population.

Cycle III: The Design Cycle: Develop/Build: The goal of this cycle is to improve the design and to increase the likelihood of technology acceptance. The design cycle is sub-divided into two phases: Develop/Build and Evaluate. Our activities reported in this paper focus on the Develop/Build phase, which centers on the creation of a highly usable artifact. To achieve the goals of this cycle, we incorporated findings from the earlier parts of our study. These include a list of content, features, and functions from the focus groups, as well as findings from the literature review. We used user-centered distributed information design methods [14] and participatory action research methods [15] for the development of our user-centered participatory design session activities. We conducted two design sessions; each successive design session was developed based on the information gathered in the previous design session.

Methods

Design Session I:

Sample

For our first design session, we recruited five PLWH who had previously participated in our focus group sessions. We had two male and three female participants whose ages ranged from 39-59 years. Two participants reported their ethnicity as being Latino and three participants reported their race as African American. One participant had never used a smartphone. For their participation, study subjects were given $25 and provided lunch as a token of appreciation for their time.

Procedures

The first design session was audio recorded and lasted approximately two hours. Study team members were present during the session and took notes. The goal of this session was to identify optimal features for improving HIV treatment and management needs of PLWH. Using the findings from the focus group sessions we developed an initial list of content, features and functions. During this session, participants discussed topics and did not look at existing apps or prototypes. For this session, we did not benchmark with other apps so as not to remove the creativity out of the intended purpose of the design session.

We asked participants to imagine each of the broad categories, identified from our focus group analysis, as a separate screen on a mobile app. We presented this information in a PowerPoint as a starting point as we sought to identify the functional requirements for a prototype of a mobile app. We presented the categories derived from our focus group sessions to the participants and asked them to discuss the content and features that would be included as functionality in a mobile app. Participants were encouraged to revise the information presented and discussion was stimulated by a set of probing questions, including: What information do you need from a mobile app related to your treatment regimen? What would make some of the current apps that you use better?

The study team met and coded the transcripts and notes from the design session to inform the content (What) and features (How) of the mobile app.

Results

Results from the first design session are categorized by topic area, derived from the Cycle I focus group sessions, and reported in Table 1. In addition to the content and features, participants also identified usability factors which would make it more likely for them to use this app. Participants mentioned it would be important not to have too many screens to access information. In addition, participants noted the user interface as well as the information presented should be "simple and straightforward." Finally, participants stressed the importance of confidentiality and privacy, reinforcing the critical need of passwords to ensure the information in the app was secure.

Table 1. Results from Design Session I

Topic Area	What	How
My Information Management	Lab results HIV and all lab work	Lab tracker
	Information on non HIV medications	Quick search feature for medications
		Health-related to-do list for HIV positives
Managing Your Medications	Medication information	Information on which medications need to be taken with or without food, and specify the type of foods
		Provide generic and brand names of medications
		Medication Interactions
		Visuals of medications with information on why to take it
	Medication reminders	Provide specific times, not just twice a day; the specific hours one dosage should be from the next
		Medication refills
	Medication tracker	Missing doses chart; how it would affect the efficacy of the medication, and the risk of not staying undetectable
Staying Healthy	Diet/Nutrition	Food log/diary
		Calorie tracker
	Exercise	Measure physical activity
	Harm reduction	Pre-Exposure Prophylaxis (PrEP)
	Measuring risk	Risk calculator
	Mental health	Stress reduction measures
	Condom information	Information about types and sizes of condoms
	STD information	Games must be sensitive to users - i.e., brain or light bulb with big eyes and bubble pop ups with information
Communication	Provider-patient communication	Virtual lectures given by providers
		Face-to-face communication with provider
		Able to share/send information with provider
		Virtual support groups with providers
		Digital audio recorder
	Peer communication	Role playing videos
		Social networking links
Resources	Support group locations	Voice activated "siri"
		GPS mapping
	Condom distribution locations	GPS mapping and information on distribution sites
	Latest HIV news	Newsfeeds

Design Session II:

Sample

For our second design session, we had six participants, all of whom participated in the earlier focus group sessions. However, only three had participated in the first design session. We had 3 male and 3 female participants whose ages ranged from 39-59 years. Two participants self-identified as Latino and four participants identified as African

American. Five of the participants were smartphone users. Of those participants who used a smartphone, social networking apps were the most frequently used applications.

Procedures

We audio recorded this session, which lasted approximately 100 minutes. The goal of the second design session was to gain information about the user interface so that prototypes could be created at the end of this design session. The investigator who ran the design session helped the end-user participants sketch a user interface of the desired mobile app. In this session, we presented the findings from the first design session as categories along with pictures of existing apps. Sticky post-it boards, size 25”x30”, were posted on the walls of the room. Each post-it board was designated for one of the topic areas listed in Table 1.

Participants were asked to identify the components to include under each topic area. The same broad categories were used as in the first design session, reported above. Once again, participants were reminded that each of the categories would be a screen on the app and were asked to describe the content, features and interface they would want to see in an app.

After participants shared their ideas, we asked probing questions to stimulate discussion about the existing apps and the need for refined content, features, and interface design. The questions were: What information do you need from a mobile app related to your treatment regimen? Which information from your healthcare visits should be viewable on your app? Participants were asked to identify app preferences including platform and design requirements, navigational features, and marketing preferences.

Figure 2: Sample Apps for Medication Reminders and Tracking

Following these questions pictures of existing apps were presented. Sample apps are in Figures 2 and 3. Participants were asked to comment on the features and interface of the existing apps.

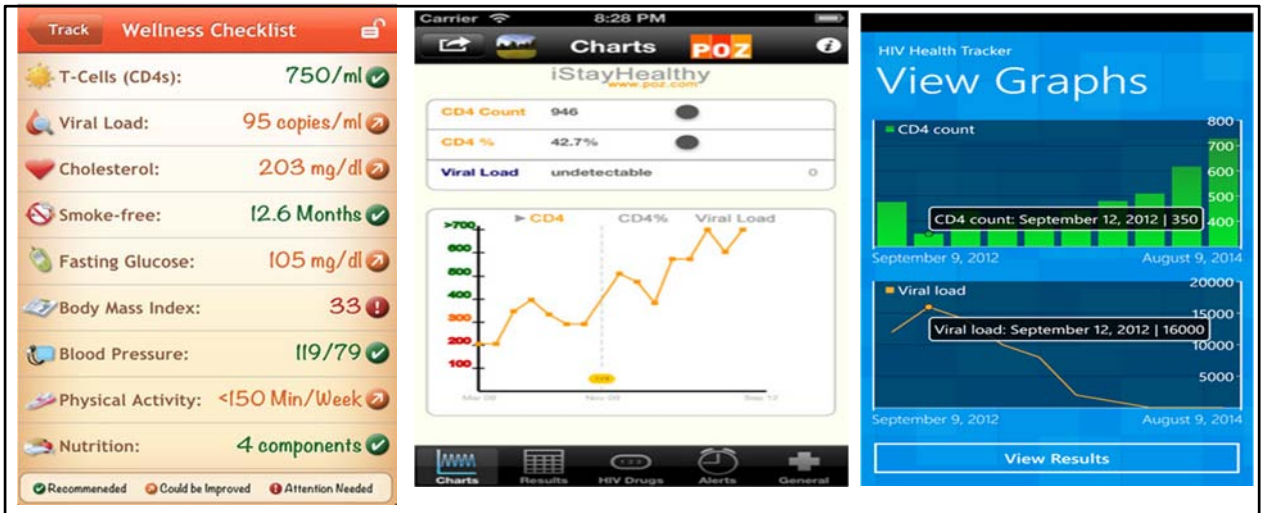


Figure 3: Sample Apps for My Information Management

Results

Participants made suggestions on the content as well as the interface of their desired app. Participants commented on the existing apps and described which features have the potential to work well for them. Moreover, participants identified features and interfaces, which were confusing and would not meet their needs. Importantly, participants had mixed views on whether they would want their app to be used as a communication tool with providers. Most participants said that they would want to be able to communicate with peers for support as well as share HIV information and other resources.

While the participants reviewed and commented on the existing apps, the design session leader sketched designs of the app on post-it boards so the participants could visualize the app. For each of the broad categories, a user interface was sketched. Sample sketches are in Figure 4.

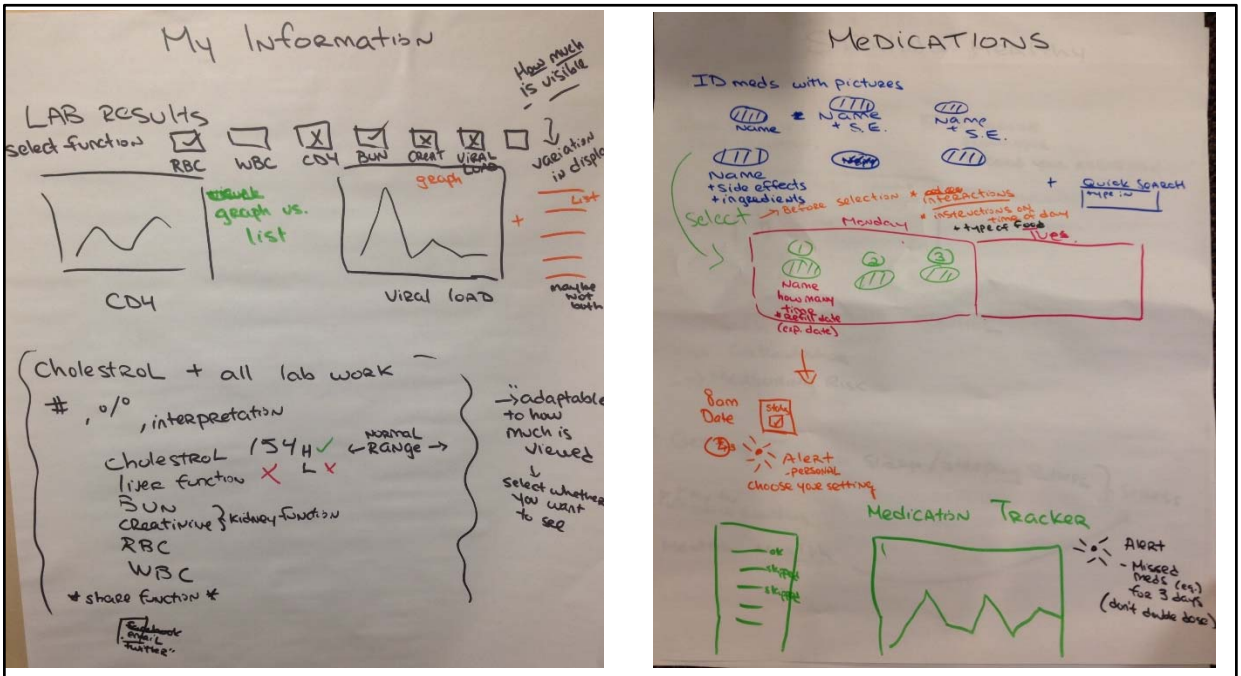


Figure 4: Sketches of Managing Your Medications and My Information Management Screens

Findings from the design sessions informed Cycle III: The Design Cycle: Build. Based on the results of the white board design, a final user interface for the prototype system was developed. We developed a low-fidelity prototype of user interface screens using a Web development tool based upon the results of the design sessions. The design sessions' findings provided information for drafting formal functional requirements, which will be refined during prototype evaluation prior to inclusion in the Design Document. The prototypes of two screens in our build phase can be found in Figure 5. These prototypes mirror the sketches presented in Figure 4. In addition, findings from our design sessions informed the creation of a map of the screen order of the mobile app. Use of a low-fidelity prototype will enable us to explore and understand our potential end-user's preferences related to content functionality (e.g. a calendar to record an appointment for HIV testing), as well as their ability to achieve tasks associated with the prototype content (e.g., scheduling a reminder in a calendar).

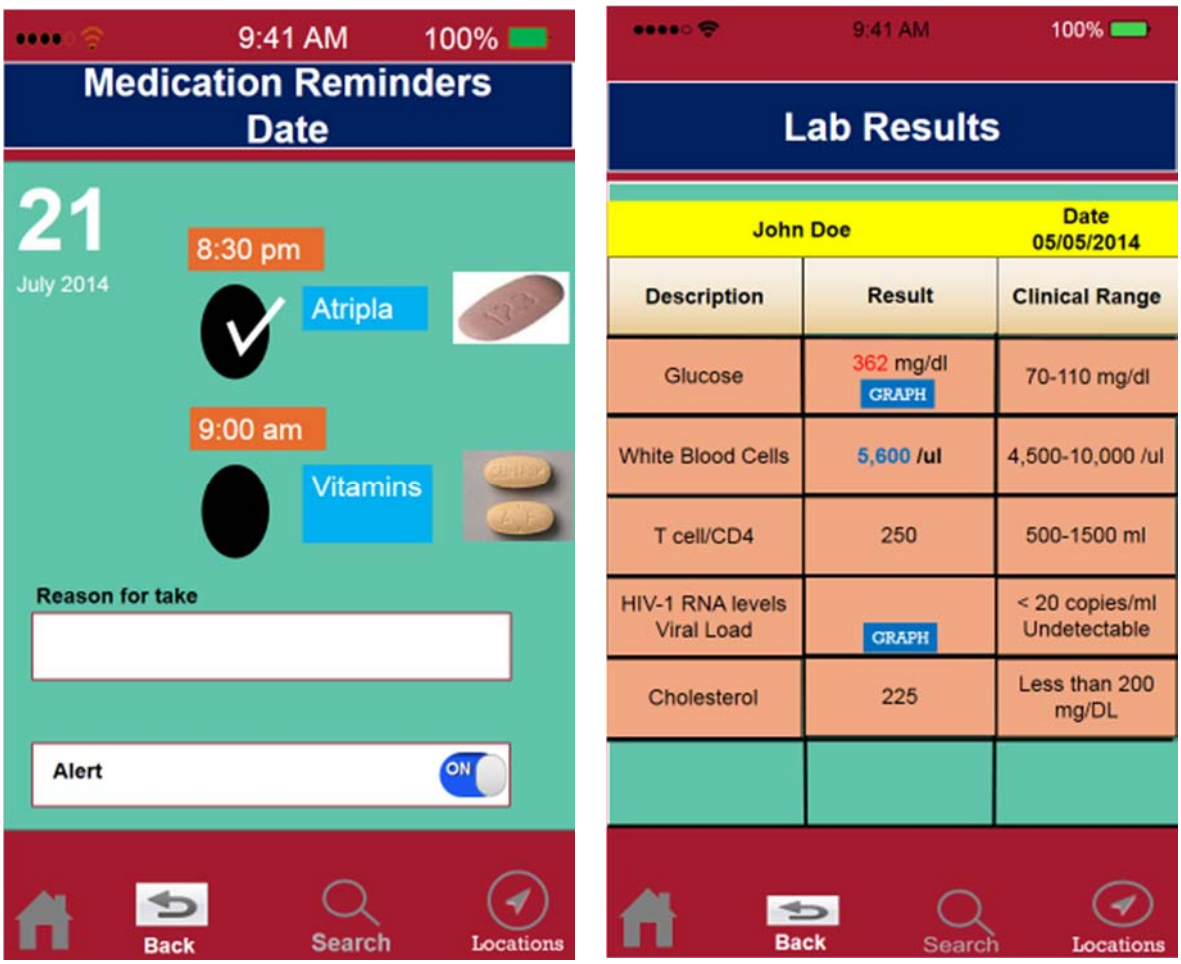


Figure 5: Mock-Ups of Medication Reminder and Lab Results Screens

The goal of Cycle III: The Design Cycle: Evaluate in order to improve the design, and to increase the likelihood of technology acceptance. To achieve this goal we will evaluate the user interface and system functions of the prototype and assess whether they are consistent with the end-users' needs. We plan to conduct two types of usability assessments: 1) a heuristic evaluation using informaticians with experience in interface design, 2) and end-user usability testing with PLWH.

Discussion

To achieve the goal of designing a mobile app for meeting the healthcare needs of PLWH, we used the ISR Framework to inform our design processes in order to build an artifact for further testing. [18] The design sessions

were an iterative process whereby study participants were considered the experts and teachers, and the investigators were the learners. The study participants contributed to the design ideas, provided feedback regarding effective interfaces, and explained which aspects of mobile interfaces, functions, and tasks they found appealing.

With increasing evidence related to the use of mHealth for improving and managing health conditions, there is recognition that scientific methods are needed to help determine how best to use mobile platforms for delivering health information. In this paper a detailed description of a novel user-centered methodology integrated with a rigorous theoretical framework is provided. Frequent interactions between developers and end-users have been noted to result in positive project outcomes. Therefore, rigorous development with end-user input is critical for the development of mHealth tools [19].

Using the ISR framework to guide our study design, we used qualitative methods, focus groups followed by end-user design sessions, which allowed us to develop an end-user perspective and not rely on the researchers' assumptions about what is important. Based on the premise that an app designed from an end-user perspective is more likely to support user needs, we applied user-centered approaches to design the content, features, and interface for a mobile app for PLWH.

Limitations

Our research focused on a single geographic area with a small sample of PLWH. Additionally, the age of those who participated in the design session activities (ages 39-59) was higher than the identified target group, adolescents and young adults. Even so, our design session activities were informed by our findings from the focus group sessions, which included a representative sample of adolescents and young adults.

This paper focuses on the process for designing a mHealth app. Further evaluation is necessary to determine whether adopting this theoretical framework for system design will improve usability and patient outcomes. The next steps for this work include a rigorous usability evaluation. Future work should also focus on the impact of a mHealth app designed through rigorous user-centered designed methods as compared to many currently available patient support apps which were not developed using a rigorous user-centered design approach.

The findings from our design session must be considered in context of the limitations of our study. Nonetheless, this work introduces a framework of system development that is perhaps more systematic and comprehensive than the development process adopted for many existing health systems. The methods that we developed and refined may be useful to others who are designing mHealth technology tools and developing HIT tools for chronically ill patients.

Conclusions

For this work, we used user-centered distributed information design methods and participatory action research methods to identify the mobile app design preferences of PLWH. Our novel approach of integrating design science and user-centered qualitative methods is unique to the design and evaluation of mHealth applications. Given our focus on design specifications from the end-user perspective, findings from this work have the potential to result in sustained use and long-term appeal for PLWH.

Acknowledgments

This publication was supported by a cooperative agreement between Columbia University School of Nursing and the Centers for Disease Control and Prevention (1U01PS00371501). Rebecca Schnall is supported by the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant Number KL2 TR000081, formerly the National Center for Research Resources, Grant Number KL2 RR024157. William Brown III is supported by NLM research training fellowship T15 LM007079 and NIMH center grant P30 MH43520. Jasmine Travers is supported by an award from the National Institute of Nursing Research of the National Institutes of Health (R01NR013687, PI: P.W.S). The findings and conclusions in this paper do not necessarily represent the views of the Centers for Disease Control and Prevention.

References

1. CDC. *US HIV/AIDS Surveillance Report, 2007.* Atlanta: Centers for Disease Control and Prevention;2009.
2. Hall HI, Song R, Rhodes P, et al. Estimation of HIV incidence in the United States. *JAMA : the journal of the American Medical Association.* Aug 6 2008;300(5):520-529.

3. New York City Department of Health and Mental Hygiene. HIV Epidemiology and Field Services. Semiannual Report. April 2011. 2011; http://www.nyc.gov/html/doh/html/dires/epi_reports.shtml. Accessed December 21, 2011.
4. mHealth Alliance. Frequently Asked Questions, 2011. 2010; http://www.mhealthalliance.org/about/frequently-asked-questions. Accessed August 3, 2011.
5. Estrin D, Sim I. Health care delivery. Open mHealth architecture: an engine for health care innovation. *Science.* Nov 5 2010;330(6005):759-760.
6. Lenhart A, Purcell K, Smith A, Zickuhr K. Social Media and Young Adults. 2010; http://www.pewinternet.org/Reports/2010/Social-Media-and-Young-Adults.aspx. Accessed March 28, 2010.
7. Lenhart A. Teens and Mobile Phones Over the Past Five Years: Pew Internet Looks Back. 2009; http://www.pewinternet.org/Reports/2009/14--Teens-and-Mobile-Phones-Data-Memo/1-Data-Memo/5-How-teens-use-text-messaging.aspx?r=1. Accessed March 28, 2010.
8. Hachman M. More Kids Using Cell Phones, Study Finds 2009; http://www.pewinternet.org/Media-Mentions/2009/More-Kids-Using-Cell-Phones-Study-Finds.aspx. Accessed March 28, 2010.
9. Lenhart A, Hitlin P, Madden M. Teens and Technology. 2005; http://www.pewinternet.org/Reports/2005/Teens-and-Technology/06-Communications-Tools-and-Teens/17-Cell-phone-text-messaging-emerges-as-a-formidable-force.aspx?r=1. Accessed March 28, 2010.
10. Brown III W. *New Media Interventions in Youth Sexual Health Promotion and HIV/STI Prevention, .* Berkley, University of California, Berkeley; 2010.
11. Hevner AR. A three cycle view of design science research. *Scandinavian journal of information systems.* 2007;19(2).
12. Simon HA. *The Sciences of the Artificial.* 3rd ed. Cambridge, Mass.: MIT Press; 1996.
13. Cross N. Designerly Ways of Knowing: Design Discipline vs. Design Science. *Design Issues.* 2001;17(3):49-55.
14. Zhang J, Patel VL, Johnson KA, Smith JW. Designing human-centered distributed information systems. *IEEE Intelligent Systems.* 2002:42-47.
15. Baum F, MacDougall C, Smith D. Participatory action research. *J Epidemiol Community Health.* Oct 2006;60(10):854-857.
16. Schnall R, Bakken S, Rojas M, et al. Design of a Mobile App as a Persuasive Technology for Persons Living with HIV. *AIDS Patient Care and STDs.* In preparation.
17. Hevner AR, Chatterjee S. *Design Research in Information Systems: Theory and Practice.* Vol 22. New York, NU: SpringerLink; 2010.
18. Hevner AR, March ST, Park J, Ram S. Design Science in Information Systems Research. *MIS Quarterly.* 2004;28(1):75-105.
19. Hyun S, Johnson SB, Stetson PD, Bakken S. Development and evaluation of nursing user interface screens using multiple methods. *Journal of Biomedical Informatics.* 12// 2009;42(6):1004-1012.

Analysis of Medication and Indication Occurrences in Clinical Notes

Sunghwan Sohn, PhD[1], Hongfang Liu, PhD[1]
[1]Department of Health Sciences Research, Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN

Abstract

A medication indication is a valid reason to use medication. Comprehensive information on medication and its intended indications has valuable potential applications for patient treatments, quality improvements, and clinical decision support. Though there are some publicly available medication resources, this medication and indication information is comprised primarily of labeled uses approved by the FDA. Additionally, linking those medications and the corresponding indications is not easy to accomplish. Furthermore, research that analyzes actual medication and indication occurrences used in real clinical practice is limited. In this study, we compiled clinician-asserted medication and indication pairs from a large cohort of Mayo Clinic electronic medical records (EMRs) and normalized them to the standard forms (ie, medication to the RxNorm ingredient and indication to SNOMED-CT). We then analyzed medication and indication occurrences and compared them with the public resource in various ways, including off-label statistics.

Introduction

A medication indication is a valid reason—most often for medical issues such as signs/symptoms or diseases/disorders—to use medication. The patients' medication history and intended indications are critical information for future medical treatment, improvements in the quality of clinical care and better clinical decision support. The linkage of medication and its intended indications and normalizing them to standard terminologies aid in clinical knowledge management [1, 2] and play an important role in enabling the secondary use of electronic medical records (EMRs) for clinical and translational research [3, 4].

A comprehensive medication and indication linkage is not straightforwardly obtained, although some freely available medication resources exist. Those medication indications are primarily labeled indications provided by manufacturers, which are later approved by the FDA. However, off-label medication use (ie, not approved by the FDA) is common in clinical practice. One in five prescribed medications in the U.S. is off-label and is most often employed as for psychiatric medication, and off-label use represents up to 31% of prescribed psychiatric medication[5].

There have been studies to aggregate indication information for prescribed medication from the data recorded by clinicians. Bashford et al. [6] observed the relation between new prescriptions for proton pump inhibitors and upper gastrointestinal morbidity in a general practitioner database and showed that new prescriptions did not necessarily reflect changes in licensed indications. Walton et al. [7] assessed a clinical decision support system based on computerized physician order entry to obtain medication indications and to record medical problems. They examined three medications to alert off-label uses and observed that the system produced less than optimal accuracy. They claimed that this result demonstrated the challenge in obtaining accurate indication information during the prescription process and suggested potential mandates for indication based prescription.

There also have been research efforts to establish a relationship between medication and indication using EMRs and public medication information resources. Burton et al. [8] processed 1.6 million EMRs from the Regenstrief Medical Record System to extract medications and diagnoses. They then linked medications and indications using RxNorm, VA National Drug File Reference Terminology (NDF-RT), and SNOMED-CT and produced 24,398 medication-indication pairs. They achieved overall sensitivity of 67.5% and specificity of 86%. Wei et al. [9] create a computable resource of medication and indication pairs, called MEDI, compiled from four publicly available medication resources, such as RxNorm, Side Effect Resource 2, MedlinePlus, and Wikipedia. They apply natural language processing and ontologies to extract indications for prescribable medication. They map medications to RxNorm ingredients and indications to the Unified Medical Language System (UMLS) and ICD9 codes. MEDI contains 63,343 medication and indication pairs. The estimated precision and recall are 56% to 94% and 20% to 51%, respectively depending on the resources. However, it should be noted that those medication-indication linkages from

the above studies are not based on the clinician-asserted data in EMRs, but, rather, are derived from using known publicly available medication information resources.

Jung et al. [10] developed the method to automatically detect off-label medication uses using a machine learner based on various features including drug and disease similarities, known medication usages, and medication-indication co-occurrence statistics in EMRs. They achieved a precision of 0.945, recall of 0.778 and F- score of 0.853 and produced 10,765 potential novel drug-indication pairs with a probability estimate cut-off of 0.95. Li et al. [11] determined the reasons for medication uses based on medical conditions in outpatient notes. They first generated the medication indication knowledge database compiled from MicroMedex, NDF-RT and Adverse Event Reporting System, and ranked indications based on the established and frequent uses. Then, they obtained an overlap between medication conditions in EMRs and the medication indication knowledge database, and selected the highest-rank medication-condition pair as the most likely reason for prescribing medication. They achieve an F-measure of 0.739 in the pilot study.

A large-scale study that analyzed diverse medication and indication pairs used in real clinical practice was limited due to technical difficulties in extracting these medications and indications from EMRs and limited access to a large amount of appropriate clinical data. In this study, we compiled medication and indication occurrences from a cohort consisting of 140K patients whose medical home is at Mayo Clinic and normalized them to the standard terminologies (ie, medication to the RxNorm ingredient (IN) and indication to SNOMED-CT). We then analyzed the medication and indication occurrences in various ways, including off-label statistics and also compared them with the public medication indication resource, MEDI.

Backgrounds

This study utilized the medical terminology and ontology to normalize medications and their indications, and the existing tools to extract them. We briefly describe them as follows:

RxNorm

RxNorm is a terminology for normalized medication names developed by the U.S. National Library of Medicine (NLM). It contains the prescription medications and many nonprescription formulations approved for human use. RxNorm uses term types to describe generic and branded names at different levels of specificity and conceptually unique medication descriptions are assigned by a concept unique identifier (RxCUI). RxNorm provides normalized medication names that link to medication variants frequently used in pharmacy management such as First Databank [12], Micromedex [13], MediSpan [14], Gold Standard [15], Multum [16], and NDF-RT [17]. Hence, RxNorm enables various systems using different medication terminologies to share and exchange data. RxNorm is becoming part of Meaningful Use to support the expanding functionality of health record technology [18].

SNOMED-CT

SNOMED-CT (Systematized Nomenclature of Medicine-Clinical Terms) is a systematically organized collection of medical terminology providing codes, terms, synonyms and definitions used in clinical documentation. SNOMED-CT provides the core general terminology for EMRs. SNOMED CT is one of the standard uses in U.S. Federal Government systems for the electronic exchange of clinical health information. It covers clinical findings, symptoms, diagnoses, procedures, body structures, organisms and other etiologies, substances, pharmaceuticals, devices and specimen (http://www.ihtsdo.org/snomed-ct). It also supports in organizing EMRs by mapping and encoding various medical concepts to the standard for clinical care and research [19].

MedXN

MedXN (Medication eXtraction and Normalization) is a UIMA (Unstructured Information Management Architecture) based pipeline to extract medication information and map it to the most specific RxNorm concept [20] (http://sourceforge.net/projects/ohnlp/files/MedXN/). MedXN focuses on medication normalization by mapping the comprehensive medication description to the best matching RxCUI using flexible matching, abbreviation expansion, inference, etc. MedXN uses externalized resources (ie, medication dictionary, attribute definitions, and regular expression attribute patterns) to allow a simple customization process for the needs of end users. For medication RxCUI assignment, MedXN produced an F-measure of 0.932. In this study, we used MedXN for normalizing medications (ie, mapping medications to RxNorm).

MedTagger

MedTagger is the open-source pipeline (http://sourceforge.net/projects/ohnlp/files/MedTagger/) that contains a suite of programs including three major components: indexing based on dictionaries, information extraction based on patterns, and machine learning-based named entity recognition [21, 22]. In this study, we used MedTagger's indexing functionality to map medication indications to SNOMED-CT.

MEDI

MEDI (MEDication Indication) is an open-source computable medication indication resource that is compiled from four publicly available medication resources, such as RxNorm, Side Effect Resource 2, MedlinePlus, and Wikipedia (http://knowledgemap.mc.vanderbilt.edu/research/content/MEDI). In this study, we compared Mayo medication-indication pairs with MEDI (MEDI_01212013_UMLS.csv) that contains 3,112 medications, 4,396 indications, and 53,106 medication-indication pairs. MEDI also contains a maker, 'possible label use,' which denotes whether a given indication is highly likely on-label or not. We utilized this information to analyze potential off-label uses in Mayo medication-indication pairs.

Materials and Methods

Data

We use all clinical notes of a cohort consisting of 140K patients whose medical home is at Mayo Clinic. The clinical notes consist of a variety of sections, such as Impression/Report/Plan, Current Medications, History of Present Illness etc., and each section contains specific content. Medication information appears in numerous ways across sections. However, most medication mentions are found in the Current Medication section, which contains medications currently being taken by patients. Therefore, it is the most important section regarding to medication information for patient care. The format of the Current Medication section is much like a grocery list. However, it contains the most diverse medication description patterns [23]. Some medication entries in the Current Medication section also occur with clinician-asserted indications, followed by "Indication:" or "Indication, Site, Instruction:" (See the example in Figure 1). In this study, we investigated the clinician-asserted medication and indication pairs extracted from the Current Medication section in clinical notes.

Aspirin 325-mg tablet enteric-coated 1 tablet by mouth one-time daily.
Indication: stroke prevention.
Bactrim-DS 160-800 mg tablet 1 tablet by mouth one-time daily.
Ativan 1-mg tablet 1 tablet by mouth once-a-month.
Indication, Site, Instruction: anxiety. take 1-mg one-hour every month.
Azopt 1 % drop suspension 1 drop ophthalmic two times a day.
Site: Both eyes.
Indication: glaucoma.

Figure 1. Medication and indication descriptions in the Current Medication section

Medication-indication pair extraction

The MedXN system was employed to extract medications and map them to RxNorm. An indication was extracted using regular expressions in the following ways: (1) extracted strings after 'Indication:' and (2) extracted strings after 'Indication, Site, Instruction:' In (1) we treated all extracted strings as an indication, but in (2) we selected only the concept classified as 'disease' and 'finding' in UMLS as an indication because there might be non-indication mentions, such as 'site' or 'instruction' that are generally not a disease or disorder.

The medication name extracted from MedXN is a generic (ie, ingredient, precise ingredients, multiple ingredients) or brand name (ie, trade name), and an indication is free-text signs/symptoms and disease/disorder expressed in various ways. In order to properly analyze overall medication-indication occurrence statistics, the synonyms of medication names and indications need to be normalized to the standard terminologies.

After we compiled all medication and indication pairs, each medication was mapped to an RxNorm ingredient (IN), and an indication was mapped to a SNOMED-CT concept. A medication name that is not an ingredient (ie, brand name) was mapped to an ingredient using RxNorm relationship (ie, "tradename_of" "has_precise_ingredient" in

RXNREL.rrf). We used MedTagger to process free-text indication strings and map them to SNOMED-CT with UMLS CUI.

Finally, we aggregated all normalized indications (SNOMED-CT with UMLS CUI) along with their frequency for a given normalized medication (RxNorm IN with RxCUI) to obtain medication and indication pairs—ie, each indication for a given medication is counted across all clinical notes to obtain the global number of occurrences. Figure 2 shows the overall process employed to obtain medication and indication pairs from clinical notes.

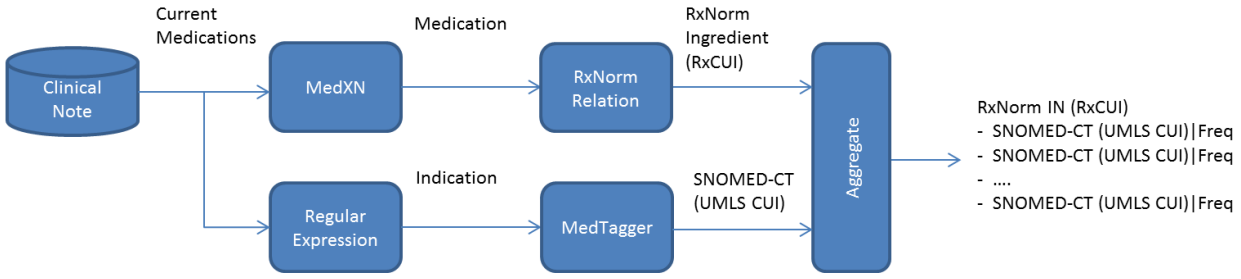


Figure 2. A workflow to extract medication and indication pairs.

Comparison of Mayo medication-indication with MEDI

We compared the normalized Mayo medication-indication pairs with MEDI (MEDI_01212013_UMLS.csv) to determine to what extent they share the same medication-indication pairs. In MEDI, medication and indication are normalized to RxNorm IN and UMLS CUI, respectively. The normalized indications of the Mayo data are SNOMED-CT concepts that are a part of UMLS and that also have UMLS CUI assigned. Hence, we first found medications that have the same RxNorm IN (ie, RxCUI) and we then checked to see if they shared the same UMLS CUI for an indication.

Medication-indication pair assessment

Various statistics, including numbers of medications/indications/medication-indication pairs, comparisons (ie, match and mismatch) with MEDI, and potential off-label medication indications, have been analyzed. For comparison with MEDI, we used two types of matches—ie, 'Exact match' and 'Flexible match.' In exact match, if both Mayo's medication (RxCUI) and the corresponding indication (UMLS CUI) appear in MEDI, we treat them as a 'match,' otherwise, we treat them as a 'mismatch.' In flexible match, a 'match' means that two of indications each in Mayo and MEDI for the same medication are within two steps of navigation in the network where the nodes are the FIND and DISO concepts in the UMLS and the edges are concept relations defined in the UMLS concept relation table (ie, MRREL), otherwise, it is a 'mismatch.'

For the normalization of indication, we used SNOMED-CT in the UMLS. SNOMED-CT is known to have a fine granularity—ie, the detailed classifications for similar medical concepts although they are semantically the same in general. Since indications in clinical notes generally are not described in such detail or precise manner, this fine granularity classification may cause mismatches when compared with the data that are built from different sources. For this reason, we also investigated the flexible match that may represent more reasonable match statistics.

Results

Mayo medication-indication statistics

Table 1 contains basic statistics delineating Mayo's medication and indication occurrences. Each column represents the unique number of medications, indications, and medication-indication pairs, respectively. The term 'Original' denotes the medication and indication descriptions as they appear in clinical notes. In 'Normalized,' medications were mapped to RxNorm ingredients, and indications were mapped to SNOMED-CT. In the normalization process, 434 medications out of 5,128 original medications did not have the ingredient relationship in RxNorm's RXNREL.rrf and 7,769 indications out of 88,522 original indications were not able to map to SNOMED-CT using our dictionary lookup algorithm. After the normalization process (without considering mismatches), the number of medication, indications, and medication-indication pairs were decreased significantly (to approximately 29%, 6%, and 21% of the original size, respectively)

Table 1. Mayo medication and indication statistics

	# medications	# indications	# medication-indication pairs
Original	5,128	88,522	140,499
Normalized	1,494	5,066	29,823

Mayo vs. MEDI medication-indication occurrences

We compared Mayo's normalized medication and indication occurrences with those of MEDI (MEDI_01212013_UMLS.csv). Since MEDI's normalization is the same as Mayo's (ie, MEDI's medications are normalized to RxNorm ingredient and indications are normalized to UMLS CUI), we were able to identify matches between Mayo's and MEDI's medication and indication pairs. Table 2 shows overall matched and mismatched medication-indication pairs between Mayo and MEDI. Mayo's medication-indication pairs are stratified by the log of the patient-level occurrences. Each row represents the cases that have greater or equal to the log of the number of patients who have a given medication-indication pairs.

Table 2. Mayo's medication-indication pairs compared with MEDI.

		Exact match		Flexible match	
$\log_2$(#pts)	# total	# matches (%)	# mismatches (%)	# matches (%)	# mismatches (%)
0 (1)	29823	4347 (15%)	25476 (85%)	17285 (58%)	12538 (42%)
1 (2)	15677	3244 (21%)	12433 (79%)	9482 (60%)	6195 (40%)
2 (4)	8095	2262 (28%)	5833 (72%)	5101 (63%)	2994 (37%)
3 (8)	4715	1601 (34%)	3114 (66%)	3084 (65%)	1631 (35%)
4 (16)	2850	1108 (39%)	1742 (61%)	1911 (67%)	939 (33%)
5 (32)	1710	712 (42%)	998 (58%)	1159 (68%)	551 (32%)
6 (64)	997	452 (45%)	545 (55%)	675 (68%)	322 (32%)
7 (128)	608	291 (48%)	317 (52%)	412 (68%)	196 (32%)
8 (256)	331	171 (52%)	160 (48%)	226 (68%)	105 (32%)

(%) denotes a percentage of the given case out of # total.

Table 3. Statistics of potential off-label indications in Mayo.

		Exact match		Flexible match	
$\log_2$(#pts)	# total	# potential off-labels in matches (%)	# potential off-labels in mismatches (%)	# potential off-labels in matches (%)	# potential off-labels in mismatches (%)
0 (1)	29823	1901 (6%)	23905 (80%)	9733 (33%)	10967 (37%)
1 (2)	15677	1349 (9%)	11706 (75%)	5106 (33%)	5468 (35%)
2 (4)	8095	897 (11%)	5498 (68%)	2643 (33%)	2659 (33%)
3 (8)	4715	613 (13%)	2917 (62%)	1516 (32%)	1434 (30%)
4 (16)	2850	387 (14%)	1619 (57%)	886 (31%)	816 (29%)
5 (32)	1710	236 (14%)	928 (54%)	510 (30%)	481 (28%)
6 (64)	997	144 (14%)	501 (50%)	267 (27%)	278 (28%)
7 (128)	608	82 (13%)	285 (47%)	150 (25%)	164 (27%)
8 (256)	331	38 (11%)	146 (44%)	70 (21%)	91 (27%)

(%) denotes a percentage of the given case out of # total.

Mayo's medication indications that appear in a low number of patients might be incorrect or simply noise, because of human mistakes or the system error in extracting or normalizing indications. Hence, we investigated statistics at a

different level based on patient volume, as shown in Table 2. The higher the number of patients who have the given medication indication—ie, the higher the number in $\log_2$(#pts)—the greater the reliability of the given medication-indication pairs. We can also assume that the portion of the number of matches between Mayo and MEDI would be increased as the given medication indications occur in the higher number of patients. As expected, the match ratio of Mayo's medication-indication pairs with MEDI increases when the given medication indications appear in a higher number of patients (see Figure 3 and 4). This phenomenon is more significant in the original match than the flexible match.

In MEDI, there exists a marker that denotes 'possible label use' whether the given indication can be considered as a labeled use or not. In Table 3, the column '# potential off-labels in matches' denotes the number of matched Mayo medication-indication pairs with MEDI, but their indications are not considered as a 'possible label use' in MEDI. The column '# potential off-labels in mismatches' denotes the number of the mismatched Mayo medication-indication pairs, in that the Mayo medication is in MEDI but its corresponding indication is not in MEDI. In light of these facts, we may consider those two cases above (ie, '# potential off-labels in matches' and '# potential off-labels in mismatches') as the potential off-label candidates. However, those indications should be further reviewed and validated thoroughly to determine true off-label uses.

Figures 5 and 6 show the portions of potential off-label candidates from Mayo's medication-indication pairs when $\log_2$(#pts) is 5 when using the exact match and flexible match, respectively. For the exact match, the ratio of potential off-label candidates is 68% (= 54% in mismatch + 14% in match, Figure 5). Whereas for the flexible match, the ratio of potential off-label candidates is 58% (= 30% in mismatch + 28% in match, Figure 6).

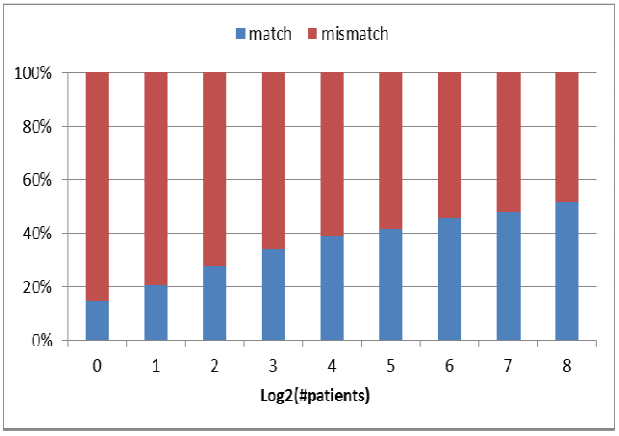


Figure 3. A ratio of match and mismatch Mayo's medication-indication with MEDI (Exact match)

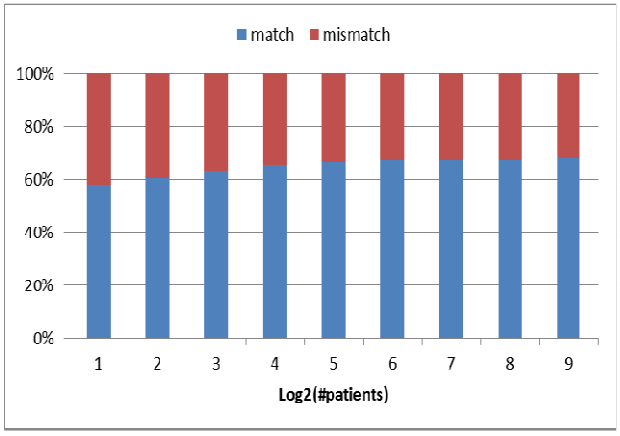


Figure 4. A ratio of match and mismatch Mayo's medication-indication with MEDI (Flexible match)

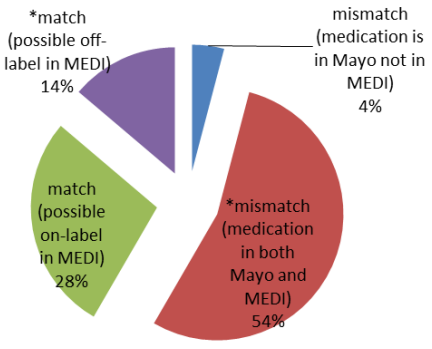


Figure 5. Off-label candidates (starts with *) in Mayo ($\log_2$(#pts)=5, Exact match)

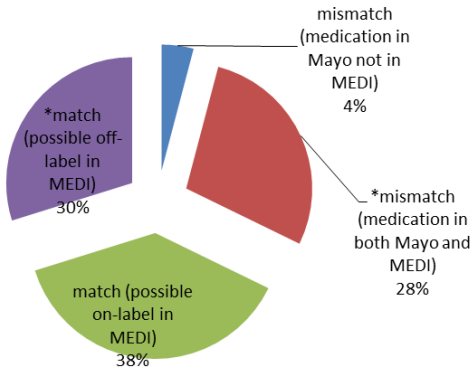


Figure 6. Off-label candidates (starts with *) in Mayo ($\log_2$(#pts)=5, Flexible match)

Table 4. Manual match of mismatched indications between Mayo and MEDI in flexible match.

$\log_2$(#pts)	Medication	Mayo indication	MEDI indication (close to Mayo)
0 ~ 1	lidocaine	vertebral compression fractures	-
	clobetasol	insect bite	-
	codeine	pain in right breast	chest pain, generalized pain
	gabapentin	PTSD	-
	hydrocortisone	nose sore	-
1 ~ 2	omeprazole	pregnancy nausea	Nausea alone
	amoxicillin	foot wound	-
	hyoscyamine	stomach cramps	unspecified disorder of stomach
	insulin lispro	dms	diabetes mellitus
2 ~ 3	nabumetone	ankle pain	generalized pain
	prednisone	facial swelling	-
	mupirocin	skin irritation	unspecified disorder of skin
	rifaximin	small bowel bacterial overgrowth	irritable bowel syndrome
	topiramate	chronic daily headache	cluster headache syndrome
	neomycin	conjunctivitis	other ill-defined disorders of eye
3 ~ 4	prednisone	urinary tract infection	-
	etodolac	back pain	acute pain, generalized pain
	amlodipine	vasospasm	-
	levalbuterol	cough	-
	digoxin	heart rate control	congestive heart failure (CHF)
4 ~ 5	glyburide	gestational diabetes	type II diabetes mellitus
	tramadol	arthritis pain	acute pain, chronic pain, generalized pain
	lidocaine	blood draw	-
	loratadine	asthmatic	-
	benzocaine	mouth sores	candidiasis of mouth
	clobetasol	scalp dermatitis	contact dermatitis and other eczema
5 ~ 6	fluticasone	sinus congestion	-
	docusate	hard stools	constipation
	bupropion	anxiety depression	major depressive disorder
	prednisone	poison ivy	-
	methylphenidate	idiopathic hypersomnia	ADHD
6 ~ 7	oxycodone	knee pain	chest pain, chronic pain, generalized pain
	fluticasone	vasomotor rhinitis	chronic rhinitis, allergic rhinitis
	tacrolimus	psoriasis	Other atopic dermatitis
	triamterene	hypertensive	hypertension
7 ~ 8	quetiapine	panic attacks	schizophrenia, bipolar
	ketorolac	headaches	generalized pain
	polymyxin b	pink eye	other ill-defined disorders of eye
	varenicline	quit smoking	unspecified drug dependence
	epinephrine	bee stings	-
8 ~	lorazepam	panic attacks	panic disorder without agoraphobia
	citalopram	anxiety depression	anxiety state unspecified
	albuterol	SOB	shortness of breath
	diphenhydramine	sleep	persistent disorder of sleep
	acetaminophen	cold symptoms	acute nasopharyngitis (common cold)

Table 4 shows some examples from the mismatch indications where medications are in both Mayo and MEDI but Mayo's indications were not found in MEDI in flexible match. We manually examined MEDI's indications for a given medication to see if there exist any similar indications to those of Mayo in a broad sense. If we found any similar indications, we place them into the column 'MEDI indication close to Mayo,' otherwise, we place '-' in this column. As can be seen in Table 4, manually identified MEDI's indications share common high-level indications with Mayo (eg, 'back pain' vs. 'general pain'; 'pink eye' vs. 'other ill-defined disorders of eye'). However, there are still some indications that were not able to be matched with MEDI in this manual process (ie, the cells with '-'). Most of these cases seem to be specific medical problems asserted by clinicians rather than general indications, which more likely appear in the low number of patient frequency. For example, Lidocaine's general use is 'numbness or loss of feeling' but Mayo's indication is 'vertebral compression fractures,' Clobetasol's general use is 'itching, redness' but Mayo's indication is 'insect bite.'

Based on the previous comparison between Mayo and MEDI, we investigated actual examples of potential off-label candidates from Mayo's medication-indication data. Tables 5 and 6 show potential off-label indications (the second column) and actual labeled indications (the third column). Table 5 contains examples of medication indications that appear in Mayo but not in MEDI, which are regarded as potential off-label uses. Table 6 contains examples of medication indications that appear in both Mayo and MEDI but which MEDI indicates not 'possible label use.' Since these indications are not marked as a possible label by MEDI, we might also assume them to be potential off-label uses.

We have examined these potential off-label indications through Wikipedia and Internet searches and have verified that they are treated as off-label uses. Through this investigation, we also found many psychiatric medications are prescribed for off-label uses.

Table 5. Examples of potential off-label indications (exist in Mayo but not in MEDI)

Medication	**Indication in Mayo (off-label)**	**On-label Indication**
Brimonidine	photophobia	glaucoma
Bupropion	smoking cessation assistance	depression
Albuterol	cough	asthma
Doxepin	allergies, hives	depression, anxiety, sleep disorders
Modafinil	wakefulness, alertness	narcolepsy
Sertraline	bipolar disorder	depression

Table 6. Examples of potential off-label indications (exist in Mayo and MEDI's possible off-label)

Medication	**Indication in Mayo and MEDI (off-label)**	**On-label Indication**
Azithromycin	traveler's diarrhea	certain bacterial infections
Gabapentin	hot flashes	certain types of seizures
Trazodone	difficulty sleeping	antidepressant
Methylphenidate	depression	ADHD, narcolepsy
Duloxetine	anxiety	depression

Discussion

A large volume of medication and indication occurrences from Mayo clinical notes have been analyzed and compared with MEDI that compiled from four public medication resources. Unlike previous studies, we used actual medication-indication pairs asserted by clinicians in real clinical practice. For exact and flexible matches (Table 2),

Mayo's medication-indication pairs match MEDI approximately 15% to 52% and 58% to 68% of the time, depending on the number of patients with the given indication, respectively.

Mayo's medication-indication pairs demonstrate that they have both off- and on-label medication uses. The ratio of potential off-label candidates when considering the indications that occurs in greater than or equal to 32 patients ($\log_2$(#pts)=5) is approximately 68% and 58% for exact and flexible matches, respectively. However, these represent rough 'potential candidates' that may be off-label indications which require further validation.

It should also be noted that the overall ratio of potential off-label indications found in our results does not represent the actual ratio of off-label uses in Mayo as a whole, because we only investigated the medication descriptions that include indications, but these are only a small portion of the total medication descriptions. Also, we do not know when the clinicians actually add indications to in the Current Medication section of our clinical notes.

Instead of medication information in clinical narratives, which is often described with intended indications, we used medication information in a specific section—ie, Current Medication—that contains a majority of patient-intake medications in a list format (as shown in Figure 1). This provides us with a relatively straightforward method of extracting medication-indication pairs. Using current medications may provide another benefit. Current medications may infer active diagnoses of patients and therefore support the compilation of the patient's comprehensive problem list, which is often difficult to maintain and manage in clinical practice. If we were able to obtain missing indications in the Current Medication section through examining other portions of clinical narratives, we would be able to compile comprehensive patient medication-indication lists, which would aid in better clinical care and decision support.

Off-label medication use is common in clinical practice, but most of these off-label uses lack rigorous evidence or studies to support them. However, off-label medication use is legal and often useful. Table 5 and 6 show good examples of potential off-label use as they show that clinicians use them differently from the original pharmaceutical use. Analyzing actual clinical practice of off-label use through a well-managed database would likely prove beneficial to patient treatments, patient safety, and quality improvements. Our study may serve as a foundation for further investigation and the eventual development of such a database.

There are some limitations to this study. Our analysis was based on medication-indication pairs automatically extracted by the system. We believe that most are correct, because medication indication descriptions in our clinical notes are relatively straightforward for extraction purposes. However, there exists the possibility that some medications and their indications could be incorrectly linked. We used normalized Mayo medication-indication pairs to compare with those of MEDI but this normalization has not been thoroughly evaluated in this study. Although we employed flexible matches between Mayo and MEDI, some mismatches may occur because of normalization differences between the two data sets. Also, false matches may occur because of over-generalization (ie, within two steps of navigation in the UMLS network). One-step distance or using certain types of relationship would be alternative ways to investigate. The appropriate medical-concept normalization would be necessary to better match indications between different medication resources.

Acknowledgements

This work was made possible by joint funding from National Institute of Health (R01GM102282A1 and R01LM011369) and National Science Foundation (ABI:0845523).

References

1. Ghitza UE, Sparenborg S, Tai B: **Improving drug abuse treatment delivery through adoption of harmonized electronic health record systems**. *Substance abuse and rehabilitation*, **2011**(2):125.
2. Roth CP, Lim Y-W, Pevnick JM, Asch SM, McGlynn EA: **The challenge of measuring quality of care from the electronic health record**. *American Journal of Medical Quality* 2009, **24**(5):385-394.
3. Tracy RP: **'Deep phenotyping': characterizing populations in the era of genomics and systems biology**. *Current opinion in lipidology* 2008, **19**(2):151-157.
4. Wilke R, Xu H, Denny J, Roden D, Krauss R, McCarty C, Davis R, Skaar T, Lamba J, Savova G: **The emerging role of electronic medical records in pharmacogenomics**. *Clinical Pharmacology & Therapeutics* 2011, **89**(3):379-386.

5. Radley DC, Finkelstein SN, Stafford RS: **Off-label prescribing among office-based physicians**. *Archives of Internal Medicine* 2006, **166**(9):1021-1026.
6. Bashford JN, Norwood J, Chapman SR: **Why are patients prescribed proton pump inhibitors? Retrospective analysis of link between morbidity and prescribing in the General Practice Research Database**. *Bmj* 1998, **317**(7156):452-456.
7. Walton S, Galanter W, Rosencranz H, Meltzer D, Stafford R, Tiryaki F, Sarne D: **A trial of inpatient indication based prescribing during computerized order entry with medications commonly used off-label**. *Applied clinical informatics* 2011, **2**(1):94.
8. Burton MM, Simonaitis L, Schadow G: **Medication and indication linkage: a practical therapy for the problem list?** In: *AMIA Annual Symposium Proceedings: 2008*. 86-90.
9. Wei W-Q, Cronin RM, Xu H, Lasko TA, Bastarache L, Denny JC: **Development and evaluation of an ensemble resource linking medications to their indications**. *Journal of the American Medical Informatics Association* 2013, **20**(5):954-961.
10. Jung K, LePendu P, Shah N: **Automated Detection of Systematic Off-label Drug Use In Free Text of Electronic Medical Records**. *AMIA Summits on Translational Science Proceedings*, **2013**:94.
11. Li Y, Salmasian H, Harpaz R, Chase H, Carol Friedman: **Determining the reasons for medication prescriptions in the EHR using knowledge and natural language processing**. In: *AMIA Annual Symposium Proceedings: 2011*. 768-776.
12. **First DataBank. http://www.firstdatabank.com/**.
13. **Micromedex. http://www.micromedex.com.**
14. **Medi-Span. http:/www.medispan.com/**.
15. **Gold Standard Drug Database. http:/www.goldstandard.com/product/gold-standard-drug-database/**.
16. **Cerner Multum. http://www.multum.com/**.
17. Brown SH, Elkin PL, Rosenbloom S, Husser C, Bauer B, Lincoln M, Carter J, Erlbaum M, Tuttle M: **VA National Drug File Reference Terminology: a cross-institutional content coverage study**. *Medinfo* 2004, **11**(Pt 1):477-481.
18. Bodenreider O, Nguyen D, P C, Chuang P, Madden M, Winnenburg R, McClure R, Emrick S, D'Souza I: **The NLM value set authority center**. *Studies in Health Technology and Informatics* 2013, 192:1224.
19. Ruch P, Gobeill J, Lovis C, Geissbühler A: **Automatic medical encoding with SNOMED categories**. *BMC Medical Informatics and Decision Making* 2008, **8**(Suppl 1):S6.
20. Sohn S, Clark C, Halgrim S, Murphy S, Chute C, Liu H: **MedXN: an Open Source Medication Extraction and Normalization Tool for Clinical Text**. *Journal of the American Medical Informatics Association* 2014, **Epub ahead of print**.
21. Torii M, Wagholikar K, Liu H: **Using machine learning for concept extraction on clinical documents from multiple data sources**. *Journal of the American Medical Informatics Association* 2011, **18**(5):580-587.
22. Liu H, Bielinski S, Sohn S, Murphy S, Wagholikar K, Jonnalagadda S, KE R, Wu S, Kullo I, Chute C: **An information extraction framework for cohort identification using electronic health records**. In: *AMIA Summits Transl Sci Proc: 2013 Mar 2013; San Francisco, CA*. 149-153.
23. Sohn S, Clark C, Halgrim S, Murphy S, Jonnalagadda S, Wagholikar K, Chute C, Liu H: **Analysis of Cross-Institutional Medication Description Patterns in Clinical Narratives**. *Journal of Biomedical Informatics Insights* 2013, **accepted**.

Reducing Wrong Patient Selection Errors: Exploring the Design Space of User Interface Techniques

Awalin Sopan[1,2], Catherine Plaisant, PhD[1], Seth Powsner, MD[3], Ben Shneiderman, PhD [1,2]
[1]Human-Computer Interaction Lab and [2]Department of Computer Science
Univ. of Maryland, College Park MD
[3]Psychiatry, Emergency Medicine, & Center for Medical Informatics, Yale University

Abstract
Wrong patient selection errors are a major issue for patient safety; from ordering medication to performing surgery, the stakes are high. Widespread adoption of Electronic Health Record (EHR) and Computerized Provider Order Entry (CPOE) systems makes patient selection using a computer screen a frequent task for clinicians. Careful design of the user interface can help mitigate the problem by helping providers recall their patients' identities, accurately select their names, and spot errors before orders are submitted. We propose a catalog of twenty seven distinct user interface techniques, organized according to a task analysis. An associated video demonstrates eighteen of those techniques. EHR designers who consider a wider range of human-computer interaction techniques could reduce selection errors, but verification of efficacy is still needed.

Project webpage with video demonstration: http://www.cs.umd.edu/hcil/WPE/

Introduction

Patient selection errors can be defined as actions (orders or documentation) which are performed for one patient that were intended for another patient[1]. Such wasteful and potentially life-threatening errors are a well-documented problem. Koppel et al. categorized 22 types of scenarios where CPOE increased the probability of prescription errors, including wrong patient selection[2]. The wrong patient can be selected when referring to patient profiles, lab results, or medication administration records[3.] According to a study by Hyman et al., placing orders in the incorrect patients chart comprised 24% of the reported errors[4]. Case-reports from the Veterans Health Administration showed that 39% of their "laboratory medicine adverse events" were caused by wrong patient order entry and 8% of these were due to reporting back the results to the wrong patient medical record[5]. Lambert et al. [6] projected that 14247 cases of wrong drug errors happen every day in USA, and many of them were caused by a wrong patient selection error. Two studies[7,8] estimated that about 50 per 100,000 electronic notes are entered in the wrong patient record. We believe that providing cognitive support though interface design for patient selection can reduce the frequency of harmful outcomes, thereby improving performance and safety.

Method

We reviewed sources of user interface selection errors from academic literature, interviews with clinicians, and inspection of existing EHR interfaces. Then, guided by a task analysis (ranging from recall of patient identity to error recovery or reporting of errors), we propose 27 user interface techniques. Eighteen techniques are illustrated in a prototype and available on video. Finally the techniques were tagged with an estimated level of implementation difficulty and estimated payoff, based on past findings of Human-Computer Interaction research[9.] Verification of efficacy is clinical environments is still needed.

Prior Work

Sengstack[10] provides a 46-item checklist for CPOE system designers to follow, categorized into clinical decision support, order form configuration, human factor configuration, and work flow configuration. Like Sengstack we found many descriptions of the safety problems associated with patient selection but very little prior work describing empirical evaluation of user interface design guidelines. The scope of the suggested techniques was limited and many proposed solutions had shortcomings. For example Adelman et al.[7] showed that ID-reentry (i.e. keying the

patient information twice) could reduce wrong patient selection errors, but this technique takes a substantial amount of additional time and therefore is likely to cause significant user frustration.
Lane et al.[11] list 'Wrong Selection' as a type of error in their taxonomy of errors in hospital environments, with wrong patient selection being only one of the errors mentioned. Most of those errors were attributed to poorly designed systems for patient selection[7]. More general work by Reason[12] categorized human errors as violations, mistakes and slips. Norman[13] asserts that the dividing line is the intention: it is a slip when the intention is correct but mechanical factors lead to error, while it is a mistake when the intention itself is wrong. Wrong patient selection can be a result of either a slip or a mistake. It is a slip if the clinician accidentally selects the patient in an adjacent row or hits the wrong number key when entering a patient ID number[14]. Slips are more frequent when the text is hard to read or small buttons are hard to select. Mistakes are more frequent when two patients are listed with the same first and last name[15] or inconsistent Medical Record Numbers (from different data sources).

Various human factors such as visual perception or short-term memory can lead to confused intentions. For example when names are sorted alphabetically similar names can coalesce visually and lead to intentional selection of the wrong one. Hospitals often have shared computers where clinicians need to log in to work and then to log off. Failure to log off can leave then next clinician's memory being cued by the wrong list. To minimize confusion between patients, it is possible to display other identifier such as date of birth, room number, gender, admission date, or attending physician name[3]. Clinicians often identify their patient by room number / location[16]. The Joint Commission now mandates that at least two patient identifiers be used[17]. Unfortunately, clinicians do not routinely verify patient identity after selection[18, 19]. They may be interrupted in the midst of selecting a patient or toggle between the records of two patients[15].

The general literature on errors can also inform better design of EHR systems[13.] Slips can be reduced by providing feedback (e.g. highlighting the item before the selection occurs). Mistakes can be reduced providing memory aids (e.g. providing pictures to remind the clinician of the patient) especially when interruptions / distractions between the time of the intent and the time to perform the task. Visual attention guidelines can also be useful. For example animation, if used appropriately, can focus users attention[20, 21]. Schlienger et al.[22] presented empirical evidence that use of animation and sound improves perception and comprehension of a change. One recent example of the use of animation is TwinList, that was designed to assist the medication reconciliation process[23].

Task Analysis

While a detailed clinician task analysis depends on a particular system and its user interface, we focus on general steps for CPOE, namely patient selection, verification, order entry and confirmation (Figure 1). In an ideal hospital, order entry would occur at the patient's bedside assuring identity confirmation, or if not at the bedside using proper documentation. Unfortunately, in reality, interruptions are frequent, documentation incomplete, and human nature tends to shortcuts. Inevitably, recall will be used[16], so it seems prudent to engineer EHR interactions to maximize odds of correct patient selection from recall. We therefore include the following tasks:

1. *Recall patient:* Clinicians must first recall the patient's identity (e.g. John Smith, or the heart failure patient in Room 201, or the patient whose ID# is on this sheet of paper). They may not recall the correct patient name.
2. *Select patient:* This is done either by typing a name or record ID number or by selecting in a list of patients.
3. *Verify selection:* After selecting a patient, clinicians may or may not verify whether they selected the correct patient. Unfortunately the patient name may be invisible (e.g. scrolled away from view). To verify may require checking the age, chief complaint, or room number, all of which may not be displayed at that time. They may not recognize the error, and continue working with the wrong patient's record.
4. *Place order:* One or more orders are entered for the patient. This can be a long and complex task which can be interrupted multiple times, sometimes to work on other patient records.
5. *Confirm order:* A confirmation dialogue may be displayed before or after order submission. This is likely to be the last chance to stop a wrong patient order.
6. *Report error:* If clinicians realize later on (minutes or days later) that they made an error after submitting an order, they should be able to cancel the wrong order and then place the same order for the correct patient. They should also be able to report the error and notify stakeholders.

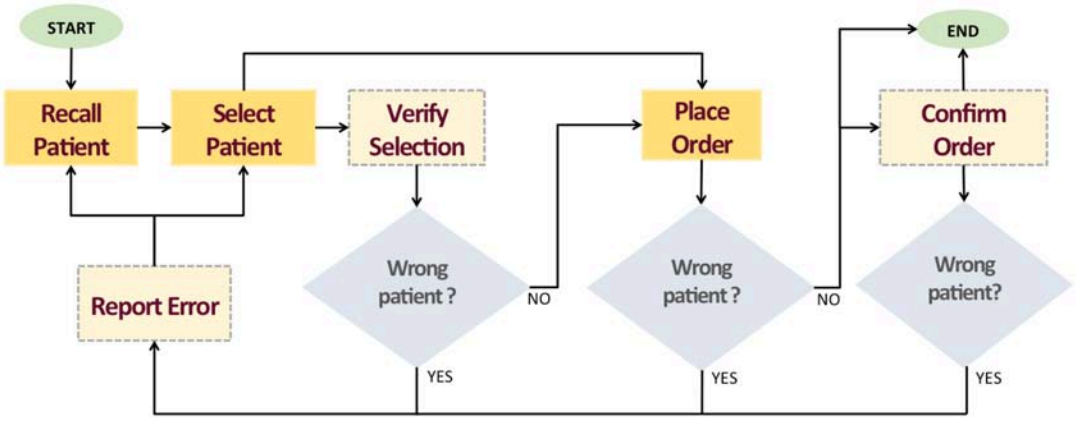


Figure 1: Task analysis of the selecting a patient and placing an order, shown as a flow-diagram. Lighter colored dotted rectangles represent tasks that may not be taken.

Proposed User Interface Techniques to Reduce Wrong Patient Selection Errors

Based on the error types reported in the prior work and the task analysis, we provide a design space of 27 potentially helpful user interface techniques (see Table 2). We grouped the techniques by the task for which they would be most useful. Some of the techniques have been successfully implemented, either in some EHR systems or in other non medical applications, but most have not. Eighteen techniques with comparatively higher pay-offs and lower implementation cost were implemented in a prototype, refined based on clinician feedback, and recorded on video (see http://www.cs.umd.edu/hcil/WPE). The prototype has a control panel that allows designers to combine techniques or check them individually. The prototype was developed in JavaScript and HTML. The animated features use the JavaScript library D3 and jQuery. For user interface and layout, we used the Bootstrap library from Twitter and jQueryUI. Source code is available.

During Task 1: Recall patient

* *Use patients' photos and other information.* Patients' photos can be used along with the patients' information both in the patient list and in the header of the order screen. Photos likely aid the recall and verification step[28]. More contextual information, like the location, admission date, attending clinician's name, chief complaint, etc. may also help clinicians recall the patient correctly. Color coding patient age and using visual clues indicating how long each patient has been in the hospital can also help clinicians identify the correct patient. Providing more information helps resolve confounding cases where two patients have similar names. Usually room number is shown in the patient selection list. But if the room numbers do not provide any contextual information, they will not be helpful.
* *Show the floor plan.* Patients are often recalled by the location where they were seen. Clinicians dealing with a new patient and unable to remember the patient name after returning to their desk should be allowed to select from a floor plan to narrow the list of patients to only see the patients at that location. In small facilities the patient information might appear directly on the map, while in large facilities the map can be used to filter patients by ward or room (Figure 2).

Other techniques can be used to facilitate recall as well as selection by narrowing the list of patients, so when clinicians only recall that a patient had a French name they can more easily scan the list to recall the exact name. In general users find their targets faster in smaller lists, and are more likely to miss a target in a large list[24].

* *Provide a personalized list of patients.* Showing only the patients assigned to the currently logged-in clinician is preferable. Different clinicians may use the same computer to log in so showing the clinician's name in the selection screen is critical. Further visual differentiation is likely to be needed to help clinicians recognize that someone else is logged in – especially when they do not expect it, or have been interrupted. Overlaying the clinicians' name in large characters over the screen after a brief period of inactivity might be useful. A distinct background color or border style could be recommended to each user during initial setup, which could then be applied to all their screens.
* *Allow sorting.* Sorting patient lists by attributes such as date of birth, date of admission, name of provider, etc. can help match the list to the clinician's mental organization of their care, for example clinicians often think of young

and elderly patients.

* *Allow filtering.* Repetitively scanning long lists takes time, and clinicians also need to be able to recall patients' identity based on other attributes than name. They should be able to use filters to narrow down the list of patients. For example, if clinicians remember that a patient was in cardiology, they can filter the list to see only the patients from that department. Our prototype provides three examples of filters (Figure 3): by patient name or ID, room number, and attending clinician name (which may be relevant for a nursing station). Filters need to be rapid, incremental and reversible. For example typing a string of letters should progressively filter down the list to only show names that include the string. Menus and controls should be optimized for rapid selection[9].

* *Allow categorical grouping.* Meaningful groups also reduce search time[25]. The patient list can be grouped into smaller lists based on categories, such as gender, floor, room number, service, department, attending physician name, etc. We show three examples of grouping in our prototype: gender, clinician and ward.

Task	User Interface Technique	Estimated Effort	Estimated Safety impact	In demo?
Recall patient	Use patients' photos and other information	**HIGH**	**HIGH**	x
	Show the floor plan	**medium**	**medium**	x
	Provide a personalized list of patients	low	**medium**	
	Allow sorting	low	low	x
	Allow filtering	low	low	x
	Allow categorical grouping	low	low	x
Select patient	Provide clues that similar names exist	**HIGH**	**HIGH**	x
	Use RFID technology	**HIGH**	**HIGH**	
	Always show patient's full name	low	**HIGH**	x
	Consider ID reentry	low	**medium**	
	Include buffer space between rows	low	**medium**	x
	Increase row height	low	**medium**	x
	Highlight row under cursor	low	**HIGH**	x
	Consider using a 2D grid instead of a list	**medium**	low	
Verify selection	**Highlight on departure**	low	**HIGH**	x
	Highlight on arrival	low	**medium**	x
	Use animated transition	low	**medium**	x
	Use a visual summary of the patient history	**medium**	TBD	
Place order	Keep the patient header visible at all times	low	**medium**	x
	Maintain consistency between screens	low	low	x
	Use clinical decision support	**HIGH**	**HIGH**	
	Consider side-by-side display of detailed patient information and order information	**medium**	TBD	
Confirm order	**Include the identity of the patient in Submit button**	low	**HIGH**	x
	Consider placing the submit button near the patient information	low	**medium**	x
Report error	Report errors with proper feedback	**medium**	TBD	
	Speed up placing the same order for the correct patient	**HIGH**	Low	

Table 2: Techniques to reduce wrong patient selection, with our estimates of the level of effort and safety impact (TBD = to be determined). Techniques with high estimated safety impact and low estimated effort are highlighted with bold text. The In-demo checkmarks (x) indicate which techniques are demonstrated in the prototype or video.

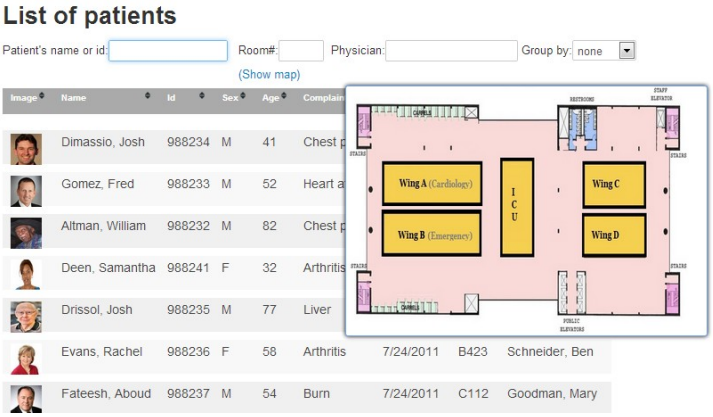


Figure 2: As clinicians click on a section of the map, the patient list is filtered to show only patients in that section. In small facilities individual rooms might be selected.

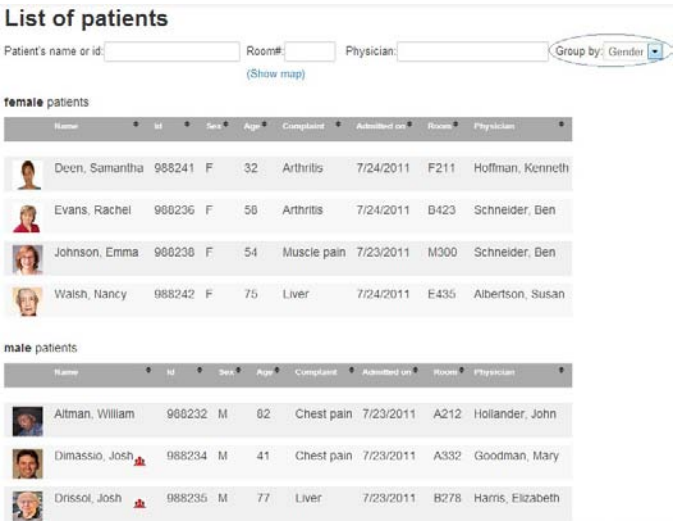


Figure 3: A list of patients with pictures and examples of controls for searching (by name, ID, room or physician) and grouping (here by gender). Each row includes a patient's photo, room number, date of birth, gender and the name of the attending clinician.

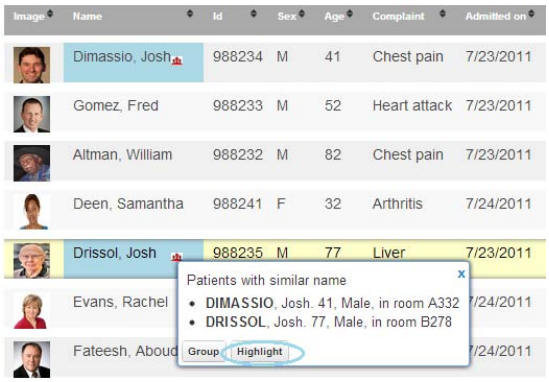


Figure 4: Notification about similar names: here, a red little icon is placed next to names that are similar to other names. The cursor is hovering over Josh Drissol, showing that the name was found similar to Josh Dimassio.

During Task 2: Select patient
After clinicians have recalled the identity of the patient whose record they want, they need to easily select that record. Several user interface techniques can help reduce errors.

* *Provide clues that similar names exist.* The system can alert clinicians of the existence of similar names. Imagine that a clinician recalls the name Adam Davis, then see David Adams at the top of the list. Those names will not sort together. Clinicians may never realize that there are two patients with similar names. Both orthographic (similar spelling) and phonological (similar sounding) similarity are detectable. Our prototype and video illustrates an example technique to notify clinicians about this similarity. It shows a small icon next to the name if there are other patients with similar names. Hovering the cursor over the icon reveals the list of similar names and additional information about those patients. The prototype also allows users to highlight the similar names in the list, or to group the rows of the patients with a similar name close together to facilitate comparison (Figure 4).
* *Use RFID technology*. Some hospitals give wristbands with their ID and demographic information. An RFID (radio frequency identification) scanner can be used to read the information. Although the technology is still costly and has its technical limitation, it has the potential to help reduce patient misidentification. RFID tags might also be used to read the location of the patient (e.g. a room) or the clinician ID.
* *Always show patient's full name.* Some tabular lists crop patients' names to fit it in the limited horizontal space. This will lead to confusion among similar names. The row displaying the patient names should be automatically resized to be large enough to show the entire names. Even omitting titles or suffixes can lead to a wrong selection. If a name in the list is abnormally long (to the point of not leaving room for the other columns) it should be wrapped inside the cell using a double height row.
* *Consider ID reentr*y. Re-entering the patient's ID a second time on the order screen has been shown to helpful[7]. Nevertheless we believe that the additional time it takes is likely to produce a lot of user frustration.

Other design techniques can be used to facilitate the mechanical aspects of target selection in lists:
* *Include buffer space between rows.* If there is "dead" space between the rows and the mouse click occur, nothing will be selected. This reduces the chance of selecting the wrong row by mistake. This could be accomplished by added white space between rows (e.g. in figure 2-4) or by only accepting clicks in the center of the row. A proper balance should be found: when the spacing becomes too large, more scrolling will become needed, and if the click-able area is too small selection becomes more difficult
***** *Increase row height.* If the row height is big enough, there is less chance of clicking the adjacent rows as we tend to click in the middle of a target. In general target selection performance follow Fitts' law, i.e. speed and accuracy decreases proportionally with the smallest dimension of the target area[26] (i.e. in this case the height of the selectable area). Again, if each row takes more vertical space, the screen can contain fewer number of rows.
***** *Increase font size.* Readability can be an issue while working under pressure and older or tired users with even mild vision impairments may have trouble. Users should be allowed to increase the font size easily so that the text is always clearly readable. In our prototype, users can use a keyboard shortcut to customize the font-size.
* *Highlight row under cursor*. This is a standard technique for list selection which is sometime omitted. Highlighting the row under the cursor draws attention to the impending selection. It also makes more apparent an inadvertent slip to an adjacent row. In our prototype, the row under the cursor is highlighted by changing the background color to yellow and changing the font to a darker gray (see Figure 4).
* *Consider using a 2D grid instead of a list.* Instead of listing patients with one patient per row, the interface can use a 2-dimensional grid with all the patient information (name, room, age etc.) presented in a single taller and narrower cell. Square targets are easier to select than long skinner ones, which would reduce the slip errors. The drawback is that it may be less obvious how the patients are sorted or to compare particular attributes (e.g. age).

During Task 3: Verify selection
Clinicians should be given opportunities to detect selection errors before order entry.

***** *Highlight on departure*. Highlighting the selected name or row while departing from the patient list can help users verify that the selection was correct[28]. In our prototype, when users selects a patient row by a mouse click, all the other rows fade out in 500 millisecond, leaving only the selected row visible on the screen (Figure 5).
* *Highlight on arrival.* The ordering screen can also display the patient information header first so the clinician can verify that the correct record has been opened. The temporary highlight only takes a fraction of a second so it does not slow down the interaction or require additional action. In our prototype, the patient name appears on the top-left corner of the screen (where readers usually start scanning or reading a screen) then the rest of the patient

identification information appears next to the name. After one second, the rest of the interface becomes visible allowing clinicians to place the order. Our prototype allows adjustment of all animations delays and informal testing suggests that keeping all delays below a second is preferable. This is in contrast with techniques such as ID Reentry might require 10-15 additional seconds for each patient selection.

***** *Use animated transition.* Clinicians' attention can be drawn to the patient name during the entire transition between screens. For example an animated transition can smoothly glide the selected name from its position in the patient list to its final position in the header of the order screen. Again the animation should be fast enough not to hamper the workflow, and slow enough to serve its purpose (i.e. attract attention to the selected name, and teach the final location of the name in the header of the order screen to new users). An adaptive approach might be helpful. For the system could detect the probability of a slip error (e.g. when the mouse click was very near the edge of a row) and slow the transition a bit more to increase the chance of error detection.

**Use a visual summary of the patient history.* A thumbnail of the timeline of the patient history could be displayed next to the patient's photo. A visual summary is faster to "read" than a textual description of the patient's history. Instead of the full history the last two or three significant events in the time line might help clinicians distinguish between patients.

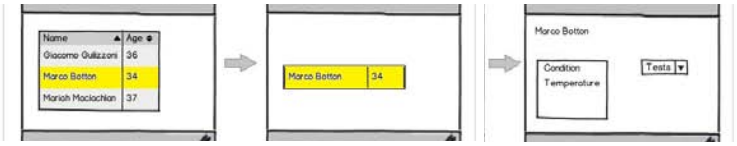


Figure 5: Highlight on departure (mock-up)

During Task 4: Place order

* *Keep the patient header visible at all times.* The header should not be allowed to scroll out of view

* *Maintain consistency between screens*. When switching from the patient list to the order screen, font-type, capitalization and color used to display the patient's basic information should not change abruptly. Such changes are visually distracting and make it harder to perceive differences between the selection screen and the order screen.

* *Use clinical decision support.* For example Galanter et al. demonstrated a reduction of wrong-patient medication errors after implementing a clinical decision support system to prompt clinicians for indications when certain medications were ordered without an appropriately coded indication on the problem list.[27.]

* *Consider side-by-side display of detailed patient information and order information*. For example, a patient's medication history is useful for clinicians ordering a new medication, and that medication history will be fairly unique to each patient so may help recognize selection errors.

During step 5: Confirmation after order entry

The goal here is to remind users of the patient identity before they finish their work:

* *Include the identity of the patient in the Submit button.* This simple technique increases the chances that clinicians will pay attention to the name or photo as they place the order. If there is no space to include the patient information on the submit button a tooltip can be used, but this is less desirable and it should be carefully designed to pop up consistently as the mouse approaches the submit button. A confirmation dialog box can serve the same purpose, but would require an additional reorientation and action from the user. It seems preferable to leave the dialog box option for situations where Decision Support can draw attention to a specific potential error (such as ordering a pap-smears for a male patient)

***** *Consider placing the submit button near the patient information.* Alternatively the submit button can be positioned near the header and the patient information so clinicians are more likely to glance at the patient's name and photo before the order submission. This may require changing button placement guidelines for the entire application.

During Task 6: Report error

* *Allow placing the same order for the correct patient without starting entirely from scratch.* Correcting the error is another aspect of the problem. Canceling or invalidating the order and re-entering a new order from scratch takes time; so we recommend that if clinicians recognize that an error was made, they be able to keep at least some of the order information and assign it to the correct patient. This technique will not only speed up the creation of the new order and will make it easier to track the occurrence of patient selection errors.

* *Report errors with proper feedback.* When an error has been made, allow reporting of the error. If a clinician

simply cancels an order then the reason of the error is not clear. It can be that the wrong patient was selected, or a wrong medication or simply that the situation has changed. The cancellation dialog box can be augmented to include a list of possible reasons for the cancellation. Collecting reasons for cancellation will provide a basis for requesting improvements based on what type of errors are most common for that system. For example, if most of the errors were due to selecting an adjacent row then the interface can be improved by inserting gap between the rows or resizing the row height. If a majority of errors are caused by similar names, then adding similarity algorithms and warnings about similar names will be most helpful. Most of the techniques proposed here have implementation costs, so error reports will be most useful to guide the selection of improvements to be made.

Limitations and Evaluation Challenges

This study has several limitations. We catalog many possible user interface techniques, and discuss why they are likely to be useful but the new techniques we propose have not been tested in clinical situations. Some of the techniques are very easy to implement and obvious payoffs (e.g. putting patient's names on the Submit button, or not truncating names) so we are very confident they can be implemented immediately without extensive evaluation. Other techniques require more complex implementation and we cannot predict the exact payoff (e.g. calculating similarity of names) so clinical evaluations may be needed to measure the benefits, but they can be safely deployed after adequate interface usability testing. The benefits of other techniques such as animated transitions or list filtering is fairly well documented in Human-Computer Interaction research (see prior work section) and used extensively in modern interfaces such as mobile devices. Those techniques may not require formal evaluations in clinical settings before they can be deployed widely, but will require careful usability evaluations (e.g. animation needs some user testing to adjust transition speed). On the other hand, long-term monitoring and evaluations in clinical settings will be beneficial to measure if clinicians' attention is still attracted to the patient identity after months of animation or highlighting use.

The main challenge faced by Human-Computer Interaction researchers is that it is difficult to simulate a hospital environment. Spontaneous errors are rare and unlikely to occur in laboratory settings. One approach is to "plant" errors and measure how the interface helps users to notice and recover from those errors[18, 28]. Another possible technique might be to simulate extremely disrupted situations (e.g. very high levels of interruptions or dual parallel tasks) so errors are more likely to be made and experiments can measure which techniques help reduce the effect of interruption. None of the laboratory methods can adequately evaluate the long-term effect of the proposed techniques.

Long-term in-situ evaluations will require both accurate detection of the errors and safety monitoring (as recommended by Sittig and al.[29]) and participation of vendors to build novel techniques into their user interfaces. Both Wilcox[8] and Adelman[7] used self-corrections to estimate the number of errors made (i.e. orders placed that were retracted within 10 min, and then reordered by the same provider on a different patient within 10 min of retraction.) Wrong patient errors are often only detected when reported by pharmacists or patients, sometimes after the patients have suffered the consequences of the errors. Elder et al.[30] showed that the perceived benefit of improving the patient care system can encourage clinicians to report errors. A voluntary error reporting and tracking system established in Duke University's Department of Community and Family Medicine succeeded in increasing the rate of error reporting by encouraging the clinicians to improve their system, involving them in the care improvement process, and keeping their identity confidential[31]. Instead of blaming the clinicians they directed their focus towards improving the overall health care system.

A/B testing[32] is a new method now commonly used to improve e-commerce websites. It is a type of between-subject test where participants are presented with one of two versions of an interface while doing their work. With adequate measurement of the error rates conducting long term A/B testing might then become possible within operational EHR systems, and lead to evaluations in real situations. With adequate IRB review each novel technique presented here could be tested by presenting the original version A to one group of clinicians and version B (with the suggested techniques) to another group of clinicians.

Conclusion

After reviewing the prior work on wrong patient selection and applying our expertise in Human-Computer Interaction design, we cataloged 27 user interfaces techniques that have the potential to reduce the occurrence of

costly and potentially life-threatening errors. Multiple techniques are available at every task of the process of recall, selection, verification, order entry, confirmation and reporting. Eighteen of those techniques were implemented and demonstrated. Feedback from clinicians and user interface designers enabled us to refine our designs, but scientific evaluation of those techniques remains a challenge. While some techniques may be evaluated in laboratory settings, more rapid progress will be made when hospitals and clinics put in place adequate incentives for reporting errors or near misses, and researchers have agreements with vendors to assist in the testing of the proposed techniques. Nevertheless many techniques have low implementation costs, low risk and high probable impact and we believe do not require long-term evaluation before they can be implemented. One such example we hope to see implemented in all EHR systems is the inclusion of the name of the patient on Submit buttons. Drawing attention to and collecting accurate data about medical care errors, is exactly what advocates of the Learning Health System promote to accelerate continuous improvement. In addition, these same techniques could have payoffs for many consumer and business applications, in which selections of people (passengers, college applicants, etc.), products (books, films, etc.), or services (flights, shipping, etc.) are made.

Acknowledgments

This work was partially supported by Grant No. 10510592 for Patient-Centered Cognitive Support under the Strategic Health IT Advanced Research Projects Program (SHARP) from the Office of the National Coordinator for Health Information Technology. We also want to thank Zach Hettinger, Meirav Taieb-Maimon, and all our Sharp project colleagues who helped along the way by provided feedback on the prototype and paper.

References

1. Schumacher RM, Patterson ES, North R, Zhang J, Lowry SZ, Quinn MT, Ramaiah M. 2011. Technical Evaluation, Testing and Validation of the Usability of Electronic Health Records. Baltimore, MD: *National Institute of Standards and Technology*. (Report No. NISTIR 7804)
2. Koppel, R., J. P. Metlay, A. Cohen, B. Abaluck, A. R. Localio, S. E. Kimmel, and B. L. Strom. Role of computerized physician order entry systems in facilitating medication errors. *JAMA: The Journal of the American Medical Association*, 293(10):1197–1203, 2005.
3. Grissinger, M., Oops, sorry, wrong patient!: Applying the joint commission's "two-identifier" rule goes beyond the patient's room. *Pharmacy and Therapeutics: a peer-reviewed journal for formulary management*, 33:625–651, 11 2008.
4. Hyman, D. , M. Laire, D. Redmond, and D. W. Kaplan. The use of patient pictures and verification screens to reduce computerized provider order entry errors. *Pediatrics*, 130(1):e211–e219, 2012.
5. Dunn, E. J. and P. J. Moga. Patient misidentification in laboratory medicine: A qualitative analysis of 227 root cause analysis reports in the veterans health administration. *Archives of Pathology and Laboratory Medicine*, 2010.
6. Lambert, B. L. , L. W. Dickey, W. M. Fisher, R. D. Gibbons, S.-J. Lin, P. A. Luce, C. T. McLennan, J. W. Senders, and C. T. Yu. Listen carefully: The risk of error in spoken medication orders. *Social Science and Medicine*, 70(10):1599 – 1608, 2010.
7. Adelman, J. S. , G. E. Kalkut, C. B. Schechter, J. M. Weiss, M. A. Berger, S. H. Reissman, H. W. Cohen, S. J. Lorenzen, D. A. Burack, and W. N. Southern.*Journal of the American Medical Informatics Association*, 2012.
8. Wilcox AB, Chen YH, Hripcsak G. Minimizing electronic health record patient-note mismatches. *J Am Med Inform Assoc* 2011;18:511-14.
9. Shneiderman, B., Plaisant, C., *Designing the user Interface* (2010) Addison Wesley
10. Sengstack, P., CPOE configuration to reduce medication errors. *Journal of Healthcare Information Management*, 24(4), 2010.
11. Lane, R., N. A. Stanton, and D. Harrison. Applying hierarchical task analysis to medication administration errors. *Applied Ergonomics*, 37(5):669 – 679, 2006.
12. Reason, J., *Human error*. Cambridge university press, 1990.
13. Norman, D. A. Design rules based on analyses of human error. *Commun. ACM*, 26(4):254–258, Apr. 1983.
14. Thimbleby H. and P. Cairns. Reducing number entry errors: solving a widespread, serious problem. *Journal*

of The Royal Society Interface, 7(51):1429–1439, 2010.
15. McCoy, A. B. , A. Wright, M. G. Kahn, J. S. Shapiro, E. V. Bernstam, and D. F. Sittig. Matching identifiers in electronic health records: implications for duplicate records and patient safety. *BMJ Quality & Safety*, 22(3):219–224, 2013.
16. Phipps, E., M. Turkel, E. R. Mackenzie, and C. Urrea. He thought the lady in the door was the lady in the window: A qualitative study of patient identification practices. *Joint Commission Journal on Quality and Patient Safety*, 38(3):127–134, 2012
17. The Joint Commission. National Patient Safety Goals. Available at: http://www.jointcommission.org/PatientSafety/NationalPatientSafetyGoals/ (Accessed 20 December 2012).
18. Hettinger, A. Z. and R. J. T. Fairbanks. Recognition of patient selection errors in a simulated computerized provider order entry system. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 56(1):1743–1747, 2012.
19. Henneman PL, Fisher DL, Henneman EA, et al. Providers do not verify patient identity during computer order entry. *Acad Emerg Med* 2008;15:641-8.
20. Novick, J. Rhodes, and W. Wert. The communicative functions of animation in user interfaces. In *Proceedings of the 29th ACM international conference on Design of communication*, SIGDOC '11, pages 1–8, New York, NY, USA, 2011. ACM.
21. Tversky, B., Visualizing Thought, *Topics in Cognitive Science*, vol. 3, pp. 499-535, 2011.
22. Schlienger, S. Conversy, S. Chatty, M. Anquetil, and C. Mertz. Improving users' comprehension of changes with animation and sound: An empirical assessment. *Proc. Human-Computer Interaction INTERACT 2007*, volume 4662 of *Lecture Notes in Computer Science*, pages 207–220. Springer Berlin Heidelberg, 2007.
23. Plaisant, C., Chao, T., Wu, J., Hettinger, A., Herskovic, J., Johnson, T., Bernstam, E., Markowitz, E., Powsner, S., Shneiderman, B., Twinlist: Novel User Interface Designs for Medication Reconciliation, *Proceedings of AMIA Annual Symposium* (2013) 1150-1159
24. Halverson, T. and A. J. Hornof. Local density guides visual search: Sparse groups are first and faster. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 48(16):1860–1864, 2004.
25. Marian, A. , F. Dexter, P. Tucker, and M. Todd. Comparison of alphabetical versus categorical display format for medication order entry in a simulated touch screen anesthesia information management system: an experiment in clinician-computer interaction in anesthesia. *BMC Medical Informatics and Decision Making*, 12(1):46, 2012.
26. Wobbrock, J. O., E. Cutrell, S. Harada, and I. S. MacKenzie. An error model for pointing based on fitts' law. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 1613–1622, 2008
27. Galanter W1, Falck S, Burns M, Laragh M, Lambert BL. Indication-based prescribing prevents wrong-patient medication errors in computerized provider order entry (CPOE). *J Am Med Inform Assoc*. 2013 May 1;20(3):477-81
28. Taieb-Maimon, M., C. Plaisant, Z. Hettinger, and B. Shneiderman. Increasing recognition of wrong patient errors while using a computerized provider order entry system an experiment. *Under review- contact authors*, 2013.
29. Sittig DF, Singh H. Electronic health records and national patient-safety goals. *N Engl J Med*. 2012;367(19):1854-1860
30. Elder, N. C. , D. Graham, E. Brandt, and J. Hickner. Barriers and motivators for making error reports from family medicine offices: A report from the American Academy of Family Physicians National Research Network (AAFP NRN). *The Journal of the American Board of Family Medicine*, 20(2):115–123, MarchApril 2007.
31. Kaprielian, V., T. Østbye, S. Warburton, D. Sangvai, L. Michener, et al. A system to describe and reduce medical errors in primary care. *Advances in Patient Safety: New Directions and Alternative Approaches*, 1, 2008.
32. Kohavi, R., R. Longbotham, D. Sommerfield, and R. Henne. Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery*, 18:140–181, 2009

Concordance of Electronic Health Record (EHR) Data Describing Delirium at a VA Hospital

Joshua Spuhl[1,2], Kristina Doing-Harris, PhD[1,2], Scott Nelson, PharmD[1,2], Nicolette Estrada, RN, PhD[1], Guilherme Del Fiol, MD, PhD[1,2], Charlene Weir, RN, PhD[1,2]
George E. Wahlen VA Medical Center, Salt Lake City UT
University of Utah, Department of Biomedical Informatics, Salt Lake City UT

Abstract

BACKGROUND: Delirium is a common syndrome in elderly hospitalized patients that is correlated with poor outcomes and higher costs yet health care teams often overlook its diagnosis and treatment. Poor data quality in EHR systems can be contributing to this as a common tool teams use to communicate and record data about their patients.
METHODS: Data were gathered from 30 patients chosen randomly that spanned various data domains in the EHR. These were analyzed for concordance as an indicator of data quality.
RESULTS: Concordance was high between the physician and nursing narrative documentation. The other domains of data were drastically less concordant.
DISCUSSION: The low concordance between structured and narrative data domains suggests that clinicians are forgoing the features available in modern EHR systems and opting to work in narrative. For informatics, this can be troubling as narrative data are difficult to compute.

Background

Delirium is a transitory syndrome that is characterized by a sudden and temporary change in mental status exhibiting a deficiency in cognition or attention. Individuals with this syndrome can have symptoms such as distracted attention or difficulty holding conversations. Highly prevalent in hospitalized patients, especially in those over age 65 [1], delirium has been shown to correlate with increased hospital length of stay, increased mortality at discharge and at 12 months, reduced independence, and greater risk of institutionalization [1-4]. Delirium is a significant burden for nursing staff, as sitters or increased monitoring is often required to care for the patient [5]. All of these effects lead to higher costs of care and poorer outcomes [6], making prevention, mitigation and treatment a high priority. Risk factors of delirium include medications, acute disease, location (e.g. intensive care units), age over 65, and comorbidities [6].

The transitory nature of delirium makes it difficult to diagnose and treat, as the affected individual may not display symptoms during any particular evaluation by a health care provider. Over 50% of delirium may go undetected [7] and 50% of patients are discharged home with delirium [2]. Difficulties in detection make a definitive diagnosis of delirium difficult to ascertain. Therefore, the diagnosis of delirium is largely a subjective process that varies depending on the experience of the health care provider working with the individual [7]. Many objective tools exist that can assist with diagnosis, such as the Confusion Assessment Method (CAM), and can consistently evaluate and score a subject's level of delirium [8]; however, these tools are not typically utilized unless initial signs and symptoms of delirium are already detected [9]. Specialized teams of mental health experts can also be engaged to improve the diagnosis of the presence of delirium and to explore mechanisms for treatment. However, these consults must be ordered which again relies on initial detection of symptoms. These factors, coupled with delirium's inconsistent presentation, make routine identification and diagnosis poor and unreliable. Awareness of the risk factors and symptoms must be present at all times by all members of the health care team to detect any subtle signs of its presence. Medications and infection are the two most common cause of delirium. Early detection would significantly decrease patient harm for both situations. Furthermore, once symptoms are identified they must be communicated to the entire team so appropriate prevention and mitigation can be enacted. Without this constant awareness and communication, the immediate risk of an adverse event due to delirium is increased significantly [10].

Communication within the health care team regarding delirium remains poor [10]. Many different forms of communication exist through which team members can convey suspicion or progression of delirium. While some of the communication occurs face-to-face, facilities that utilize electronic health records (EHRs) can capture some communication electronically and make it available to the entire team. The EHR becomes a comprehensive record

of the patient's status and progress. Health care team members utilize the EHR to stay current with their patients. Ultimately, the EHR is a tool for communication for the individual clinician, team and institutional level stewardship

Previous studies show that documentation quality within an EHR is generally poor. Data are often recorded inconsistently or not read by other team members [11, 12]. The result is that clinicians must expend considerable energy and cognitive effort in looking for inconsistent information, validating the information and ensuring accuracy [13]. Currently no tools exist that display the degree to which recorded information is inconsistent, a valuable tool to support decision-making for clinicians. In the case of delirium, its transitory nature coupled with it being a change in mental status (making determination of baseline data a key piece of information), as opposed to physical ailment, lead to vagueness and uncertainty regarding the quality of mental health data. As a result, it is commonly undiagnosed and untreated. Given the difficulty of identifying delirium and the deficiency of quality in EHR documentation, it is not surprising that delirium is grossly undertreated.

Data concordance plays a major role in documentation quality [14]. We conducted this study to determine the concordance of data on delirium as an indicator of document quality within the Veterans Affairs (VA) EHR system. We compiled data from International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM) codes, problem lists, physician and nursing narrative notes, physician orders, and CAM assessments.. By examining the concordance of data on delirium we can inform health care team members about discrepancies. This information may motivate them to improve the quality of the data in the EHR.

Methods

We used a cohort study design to determine the degree of concordance within EHR data. A cohort of 30 patients were randomly chosen out of 528 patients admitted to the two medical units at the VA Medical Center in Salt Lake City, Utah. The cohort consisted of patients who were over the age of 65 (a risk factor for delirium), and an inpatient stays that occurred during the 2013 calendar year. We chose a single, whole, contiguous year to eliminate the possibility of temporal confounding. Finally, patients included were from the Acute Medicine and Telemetry units as these are the main medical units at this facility. We specifically excluded ICU, psychiatry and surgery for this pilot work in order to identify data collection practices on these vulnerable units.

For the 30 randomly selected patients in the cohort, we performed retrospective chart reviews for their most recent inpatient stay. These reviews gathered data from EHR domains including the following: ICD-9-CM diagnosis codes (Table 1 lists the codes used), problem list entries, physician orders, physician and nursing narrative notes, CAM assessments, and mental status assessments. To determine the presence of delirium in diagnosis codes we relied on the set of terms derived from Hope et al. [15], who performed a similar enumeration (but did not randomly select patients). As in this study, orders for restraints and sitters and CAM assessments were included. In addition, we included medication review that was not done in their study. Finally, physician and nursing narrative notes were included since they are predominantly used to collect the status and progress of the patient throughout their stay as well as the thought processes of the health care team.

Once the data were gathered, the narrative note reviews were analyzed for inter-rater reliability to ensure consistent evaluation. We then performed an analysis of heterogeneity between the different data domains. The data were categorized on each patient for each domain with a binary evaluation: evidence of delirium exists or not.

Table 1. ICD-9-CM Codes Considered Evidence of Delirium

Code	Description
290.3	Senile dementia with delirium
291	Alcohol-induced mental disorders
291.0	Alcohol withdrawal delirium
292	Drug-induced mental disorders
292.81	Drug-induced delirium
293	Transient mental disorders due to conditions classified elsewhere
293.1	Subacute delirium
780.09	Other alteration of consciousness

Results

Overall, a total of 10 patients had some reference to delirium in their medical record. Of those 10, 9 had agreement between the narrative notes of nurses and physicians. Within those 9, 3 had ICD9 codes representing delirium and 1 had a problem list entry (it was not one of the patients identified in the ICD9 codes). Table 2 presents the distribution of positive findings.

Table 2. Occurrence of delirium

Source	Cases Identified	Notes Reviewed	Distinct Terms
MD Notes	9	178	15
RN Notes	9	298	14
ICD-9-CM Codes	3		
Problem List Entries	1		
Orders	0		

To test concordance an overall Kappa was computed across all sources using presence/absence as the categories and the sources as the "judges". The Kappa was moderately low, despite a baseline prevalence of about 35%. (Kappa= 0.33). Specific contrasts were done to assess the degree of association between nurses and physicians (82% agreement or Kappa = 0.83) and between ICD9 codes and ANY narrative note (27% agreement and Kappa = 0.36) and between ICD9 and the problem list (Kappa= 0.47). Statistical inference was limited due to small sample size.

Patients in the cohort took an average of 15 (range = 0 to 33) unique medications as outpatients the year prior to their hospitalization. While inpatient, the average number of unique medications remained the same at 15 medications per patient during their stay. The number of unique medications did not correlate with the presence or absence of delirium symptoms in our cohort OR the presence of an ICD-9-CM diagnosis code for delirium. Two patients in the cohort (20% of patients with delirium) received pharmacological treatment with an antipsychotic (haloperidol) to help manage symptoms of delirium/agitation, one with an ICD-9_CM code for delirium and the other without. Both patients had documentation of delirium in the narrative text.

Further assessment of the narrative text (495 progress notes) revealed significant variation in how the different roles used narrative text to document delirium. Nurses almost uniformly used somewhat structured terms, such as oriented times 3 or 4. In contrast, the terms that physicians were used were more collogue and informal. Table 3 lists most of the terms used by the different roles in their text.

Table 3. Terms and references used by different roles in narrative text

MD Terms	RN Terms
changes in mental status	*AAO to self*
confused…delirium	*alert and confused*
delirium in the setting	*confused on lortab*
delirium not worse…yesterday	*disoriented*
fluctuating orientation	*combative*
increasing cognitive impairment	*mittens/agitation*
mild cognitive impairment	*cognitively impaired*
muttering…does not respond to qt.	*showing confusion*
not oriented	*sitter*
seem to be experiencing delirium	*talking to his 1:1 [sitter]*
AAOx1	*cognitive impairment*
aggressive	*aox1*
weird answers	*CAM-ICU Positive*
dementia/tremors/Parkinson's	*sundowner*
hx delirium/delirium precautions	

Discussion

The results show there is a high degree of concordance regarding the presence/absence of delirium within the clinical narrative documentation but little outside of that domain. Differences were found in the nature of the

narrative. In contrast, ICD9 codes only represented a little less than a third of those documented in text and did not particularly correlate with physician notes, orders (there were no orders in this set) or medications. All of the documents for the reviewed patients had minor variability in regards to frequency and agreement of mental status evaluations.

The consistent evaluation and recording of mental health status, especially for nurses., in the narrative notes may have been facilitated by the regular use of note templates within VA hospitals. These templates include a section for mental health evaluation. Thus nearly all physician and nurse narrative documents reviewed noted the mental health status of the patient. The only exceptions to this rule were notes documenting a non-routine event such as helping a patient to the bathroom. The ubiquity of mental status evaluations led to the detection of acute episodes of delirium in these narrative notes. The substantial literature attesting to the neglect of delirium detection and treatment made this a welcome finding. The result shows that the VA making a concerted effort to keep mental health status at the forefront of their care model. It also demonstrates that the most concordant and, hence, reliable information is still locked away in narrative notes. This information can only be programmatically utilized through complicated natural language processing (NLP) technologies. Ultimately, we can interpret this concordance as an indicator of high quality narrative documentation at the VA by the nurses.

Discrete and coded data, were almost completely deficient in representing any changes to the patient's mental status. Interestingly these are data most likely to be used for computerized decision support (CDS). Orders for the pharmacological management of delirium did not correspond to the number of patients with documentation of delirium. Medications typically associated with delirium adverse events did not seem to correlate with the presence or absence of delirium either however. However, the study was under powered to detect such a difference. Additionally, only a small percentage of case records were found to have delirium indicated in either the billing ICD-9-CM codes or the active problem list. Other studies, such as Inouye et al., had comparable results. They found ICD-9-CM codes had a sensitivity of 3% and specificity of about 99% [16]. These findings mean that where delirium was coded the code was accurate, but that up to 93% of delirium cases were missed. Typically ICD-9-CM coding is done after the patient has been discharged as a way to enumerate the diagnoses that were treated during the inpatient episode. At non-VA institutions these codes are required to justify charges. However, clinical personnel do not assign ICD-9-CM codes. They are determined by trained coders, who scan the documentation looking for diagnoses. Since Delirium does not often appear on problem lists, the evidence we detected in the narrative notes was related to the recording of mental health status checks. Coders are by law prohibited from generating diagnoses not listed explicitly in the record. Therefore, they may not read them in the level of detail required to find the necessary evidence to contact a physician to inquire whether a patient had delirium. In this case, using concordance as an indicator, the quality of ICD-9-CM codes are poor, leading one to doubt the validity of knowledge, research or otherwise, gained from them.

The lack of delirium being recorded in the problem list is intriguing but not uncommon [17]. Problem lists are a common component in modern EHR systems. They are meant to facilitate a shared mental model of the issues surrounding a patient. The lack of recording suggests that they are seldom used at the VA for delirium. This is unfortunate as there is evidence that shows consistent problem list use improves care [18]. As many CDS systems utilize coded data in problem lists, a lack of accurate information could lend to more harm than good being done for the patient. Government initiatives such as Meaningful Use seek to mandate the use of problem lists. Overall, without a change in the integration of problem lists are integrated into routine patient care they will most likely continue to provide minimal usefulness to the health care team and informatics in general.

There are a few limitations in this study worth noting. The first is that the data analyzed were limited to a single VA hospital and the results may not be generalizable to other hospitals. The methodology with which this VA hospital cares for and document its patients could be unique. We were limited by a small sample size due to the time and expense required to manually review narrative notes. The time and effort limitation reinforces the need for reliable discrete data. Finally, this study focused on delirium, it cannot be assumed that the same pattern occurs for any other diagnoses. Future studies could broaden the scope of this study to determine if similar patterns exist in other health care facilities as well as with other diseases and diagnoses.

Conclusion

This study demonstrates that the highest quality data are those that are the hardest to obtain. Moreover, it reaffirms the clinician's preference for narrative in recording patient information. For initiatives involving CDS, research, or other data dependent ventures the implications of this research are that accessing the information within narrative notes will be vitally important.

The data pertaining to delirium in the VA EHR system have very little concordance between the narrative documents and the discrete coded data, while the narrative domains alone demonstrate a high degree of concordance, but a substantial degree of 'fuzziness". The EHR as a tool to facilitate communication and collaboration between health care team members has not been embraced and continues to function as merely a document repository. This challenges the utility of feature-rich systems compared to the cost to implement and maintain them as well as restates the clinicians' preference for narrative-based information exchange.

References

1. Inouye SK. Delirium in older persons. N Engl J Med. 2006 Mar 16;354(11):1157–65.

2. Ely EW, Shintani A, Truman B, Speroff T, Gordon SM, Harrell FE, et al. Delirium as a predictor of mortality in mechanically ventilated patients in the intensive care unit. JAMA. 2004 Apr 14;291(14):1753–62.

3. Siddiqi N, House AO, Holmes JD. Occurrence and outcome of delirium in medical in-patients: a systematic literature review. Age Ageing. 2006 Jul;35(4):350–64.

4. McCusker J, Martin C, Abrahamowicz M, P F, Belzile E. Delirium predicts 12-month mortality. Arch Intern Med. 2002;162:457–63.

5. Carr F. The Role of Sitters in Delirium : an Update. Can Geriatr J. 2013;16(1).

6. Fong T, Tulebaev S, Inouye S. Delirium in elderly adults: diagnosis, prevention and treatment. Nat Rev Neurol. 2009;5(4):210–20.

7. Cavallazzi R, Saad M, Marik PE. Delirium in the ICU: an overview. Ann Intensive Care. Annals of Intensive Care; 2012 Jan;2(1):49.

8. Morandi A, McCurley J. Tools to Detect Delirium Superimposed on Dementia: A Systematic Review. J Am Geriatr Soc. 2012;60(11):2005–13.

9. Sanders AB. Missed delirium in older emergency department patients: A quality-of-care problem. Ann Emerg Med. 2002 Mar;39(3):338–41.

10. Detweiler MB, Kenneth A, Bader G, Sullivan K, Murphy PF, Halling M, et al. Can Improved Intra- and Inter-team Communication Reduce Missed Delirium? Psychiatr Q. 2013 Dec 6.

11. Mamykina L, Vawdrey DK, Stetson PD, Zheng K, Hripcsak G. Clinical documentation: composition or synthesis? J Am Med Inform Assoc. 2012;19(6):1025–31.

12. Hripcsak G, Vawdrey DK, Fred MR, Bostwick SB. Use of electronic clinical documentation: time spent and team interactions. J Am Med Inform Assoc. 2011;18(2):112–7.

13. Keenan G, Yakel E, Dunn Lopez K, Tschannen D, Ford YB. Challenges to nurses' efforts of retrieving, documenting, and communicating patient care information. J Am Med Inform Assoc. 2013;20(2):245–51.

14. Stetson P, Morrison F. Preliminary development of the physician documentation quality instrument. J Am Med Inform Assoc. 2008;15(4):534–41.

15. Hope C, Estrada N, Weir C, Teng J, Damal K, Sauer B. Documentation of Delirium in the VA Electronic Health Record. BMC Res Notes. Forthcoming 2014.

16. Inouye S, Leo-Summers L, Zhang Y, Bogardus S, Douglas L, Agostini J. A Chart-Based Method for Identification of Delirium : Validation Compared with Interviewer Ratings Using the Confusion Assessment Method. J Am Geriatr Soc. 2005;53:312–8.

17. Onofrei M, Hunt J, Siemienczuk J, Touchette D, Middleton B. A first step towards translating evidence into practice : heart failure in a community practice-based research network. Inform Prim Care. 2004;12(3):139–45.

18. Banerjee ES, Gambler A, Fogleman C. Adding Obesity to the Problem List Increases the Rate of Providers Addressing Obesity. Fam Med. 2013;45(9):629–33.

Pediatric Readmission Classification Using Stacked Regularized Logistic Regression Models

Gregor Stiglic, PhD[1], Fei Wang, PhD[2], Adam Davey, PhD[3], Zoran Obradovic, PhD[3]
[1]University of Maribor, Maribor, Slovenia; [2]IBM T.J. Watson Research Center, Yorktown Heights, NY; [3]Temple University, Philadelphia, PA

Abstract

Background: *Regulations and privacy concerns often hinder exchange of healthcare data between hospitals or other healthcare providers. Sharing predictive models built on original data and averaging their results offers an alternative to more efficient prediction of outcomes on new cases. Although one can choose from many techniques to combine outputs from different predictive models, it is difficult to find studies that try to interpret the results obtained from ensemble-learning methods.*

Methods: *We propose a novel approach to classification based on models from different hospitals that allows a high level of performance along with comprehensibility of obtained results. Our approach is based on regularized sparse regression models in two hierarchical levels and exploits the interpretability of obtained regression coefficients to rank the contribution of hospitals in terms of outcome prediction.*

Results: *The proposed approach was used to predict the 30-days all-cause readmissions for pediatric patients in 54 Californian hospitals. Using repeated holdout evaluation, including more than 60,000 hospital discharge records, we compared the proposed approach to alternative approaches. The performance of two-level classification model was measured using the Area Under the ROC Curve (AUC) with an additional evaluation that uncovered the importance and contribution of each single data source (i.e. hospital) to the final result. The results for the best distributed model (AUC=0.787, 95% CI: 0.780-0.794) demonstrate no significant difference in terms of AUC performance when compared to a single elastic net model built on all available data (AUC=0.789, 95% CI:0.781-0.796).*

Conclusions: *This paper presents a novel approach to improved classification with shared predictive models for environments where centralized collection of data is not possible. The significant improvements in classification performance and interpretability of results demonstrate the effectiveness of our approach.*

Introduction

The widespread use and availability of Electronic Health Record (EHR) data are responsible for numerous studies and should result in measurable improvements of the healthcare quality level in the coming years. However, most of the available data from healthcare repositories nowadays are still very limited in terms of heterogeneity and most studies use local data repositories [1, 2]. The Health Information Technology for Economic and Clinical Health HITECH act of 2009 was one of the most recent incentives of the US government to establish healthcare data exchange systems. However, in many cases legal and privacy concerns is still the main reason against data exchange between hospitals or healthcare providers [3]. On the other hand, researchers are often faced with a complex task of data integration even in cases where an agreement for data integration is reached [4]. On the other hand, one can observe a growing number of studies that use millions of records to build predictive models that will be used on the future repositories of EHRs [5-7].

Multiple privacy-preserving distributed classification models with applications in healthcare were proposed recently. Mathew and Obradovic [8] proposed a distributed knowledge-mining framework based on a decision tree classifier. Their approach allows heterogeneous data schemas to build a decision tree using locally abstracted data – i.e. no raw data needs to leave the hospital. Another distributed approach was proposed in [9] where distributed distance metric learning was used to assess patient similarity. A recent approach by Wang et al. [10] used a Bayesian approach to online learning based on logistic regression. High-level privacy preserving was ensured by encrypted posterior distribution of coefficients during the exchange between the server and the client. Additionally, the proposed model supports asynchronous communication between hospitals and allows dynamic model updating – i.e. there is no need to rebuild the model for each new patient. However, the complexity of the proposed approach rises with the number of included hospitals and [10] only presents results with up to 8 hospitals contributing to the global logistic

regression model. Rider and Chawla [11] use probabilistic graphical models to facilitate transfer learning between distinct healthcare data sets by parameter sharing while simultaneously constructing a disease network for interpretation by domain experts. Their approach is primarily used to rank the patient disease risk for multiple diseases simultaneously. A recent study by Wiens et al. [12] presents a more empirical evaluation of a transfer-learning approach using data from multiple hospitals to enhance local hospital predictions. A large sample of 132,853 admissions from three hospitals was used to test different scenarios on sharing the data or using models built on data from a specific hospital to predict on data from other participating hospitals. Although the study does not address privacy directly, it offers interesting results demonstrating high performance gains when data from all hospitals can be used in the final prediction.

Our study utilizes a large dataset of hospital discharge data to propose a novel approach in distributed predictive modeling that allows asynchronous exchange of models in a peer-to-peer or centralized environment. The proposed predictive model consists of two levels and is based on deep learning architectures [13, 14] that originate from a stacked generalization approach [15]. It was evaluated using data from 54 hospitals in California to demonstrate the large-scale deployment possibilities of the proposed approach. Compared to similar frameworks, we allow an additional high-level interpretability of results to obtain additional hospital level information. In contrast to most related work our approach allows combinations of different predictive models.

Background

Beginning October 1, 2012 under section 3025 of the Affordable Care Act, hospital reimbursements became tied to performance relative to preventable 30-day Medicare hospital readmission rates compared with hospitals having similar predicted risk profiles. Initially, readmission rates are tracked for three specific adult diagnoses: acute myocardial infarction (MI), congestive heart failure (HF), and pneumonia (PN). This change in the structure of Medicare reimbursements places increasing importance on the ability of health care providers to identify predictors of 30-day hospital readmissions as well as to identify characteristics of individuals and providers associated with above-average levels of readmission risk. Under-performing hospitals will see reduction of up to 1% in Medicare base reimbursements for services related to all diagnostic-related groups (DRGs). In 2010, these targets would have placed half of all hospitals in the under-performing group.

There are now plans to expand this approach to consider pediatric populations and growing interest in considering the value of pediatric readmissions rates as hospital quality indicators [16-18]. Research on pediatric readmission rates suggests that a small number of cases account for a disproportionate number of hospital readmissions [19]. Similarly, wide hospital-level variation is also seen. Considering sickle-cell anemia readmissions in a sample of more than 12,000 hospitalizations of some 4,762 children from 33 hospitals, Sobota et al. [20] found that even after adjusting for individual-level characteristics such as age, treatment, and complications, there was 4.2-fold variation in readmission rates between hospitals. Considering only admissions for appendicitis, Rice-Townsend et al. [21] found 3.8-fold variation between hospitals in readmissions rates after adjusting for disease severity and insurance status. In a larger study including 568,845 all-cause admissions at 72 children's hospitals, Berry et al. [22] found 28.6% greater adjusted readmission rates in hospitals with high versus low readmission rates. Overall, these findings point to the importance of considering differences between hospitals.

Despite considerable interest in this topic, the accuracy of predictive models for 30-day hospital readmission is not particularly strong. Horwitz et al. [23], for example, found in-sample prediction by area under the curve (AUC) of 0.61, 0.63, and 0.61 for MI, HF, and PN, respectively using Medicare claims data. More recently, focusing only on MI readmissions, Krumholz et al. [24] used 2006 Medicare claims data to compare models relying on claims data versus the combination of claims data and medical record data. These authors found high agreement (r=.98), but their overall model had an AUC of just 0.63.

Methods

Combining outputs of different predictive models to improve the performance of classification has been widely addressed in the research literature for the past two decades. The simplest approach to combining predictions from different models is majority voting, also called bagging [25] when combined with bootstrap sampling from the training dataset. In cases where classifiers are built from disjoint sets of data, Ting and Witten [26] name the

approach "dagging." In the same paper they propose an approach called dag-stacking, where an additional model is built to combine the outputs of low-level models instead of majority voting.

Stacked generalization approach inspired many novel, so-called deep learning frameworks [13, 14]. This paper introduces a stacked generalization based classification approach, inspired by dag-stacking, with a few important modifications from the original implementation by Wolpert [15] and Ting and Witten [26]. As described above, stacked generalization was originally proposed to leverage different types of classifiers that are used on the lower level of the stacking framework using a high-level classifier. Our approach aims to combine different classifiers of the same type built on disjoint sets of data (i.e. each classifier is built on data from a specific hospital). Ting and Witten [27] already noticed that it is possible to improve the results of the originally proposed stacking framework by combining confidence levels in contrast to predicted class labels. In our case, we use predicted risk of readmission obtained from regularized logistic regression models to prepare a high-level dataset (Figure 1). Additionally, we use a high level classifier (i.e. sparse logistic regression) that allows us to interpret the results of the high level classifier.

Suppose we have K different clinical sites or hospitals, on each site k there are n_k patients characterized by a matrix $X_k = \left[x_1^k, x_2^k, \cdots, x^{k_{n_k}} \right]^T$. We construct a local model $f_k : R^d \rightarrow R$ at each site *k* such that $f_k(x_i^k)$ returns the probability of hospital readmission for the *i*th patient in the *k*th clinical site, *d* is the dimensionality of the patient feature vector. For example, in the case of logistic regression, it learns a linear decision function

$$f_k(x_i^k) = w_k^T x_i^k + b_k$$

by minimizing the following logistic loss

$$\ell_{org}^k(w_k, b_k) = \frac{1}{n} \sum_{i=1}^{n_k} \log\left[1 + \exp\left(-y_i^k \left(w_k^T x_i^k + b_k\right)\right)\right]$$

where $y_i^k \in \{0,1\}$ is the label of x_i^k, such that $y_i^k = 1$ if the *i*th patient in the *k*th clinical site is re-admitted to hospital within 30 days, and $b_k \in R$ is the offset. The optimal solution of (w_k, b_k) by minimizing $\ell_{org}^k(w_k, b_k)$ can be obtained by iterative optimization methods such as gradient descent, Newtown method, or coordinate descent. For a detailed comparison of those methods one can refer to [28].

Generally there are many factors involved in every patient during the predictive modeling procedure, which makes *d* fairly large. In most of the cases only a small portion of factors would play important roles. It is highly desirable if the prediction model can also identify the set of important factors. Therefore Sparse Logistic Regression model is proposed, which aims to get the optimal (w_k, b_k) by minimizing the following objective

$$\ell_{sp}^k(w_k, b_k) = \frac{1}{n} \sum_{i=1}^{n_k} \log\left[1 + \exp\left(-y_i^k \left(w_k^T x_i^k + b_k\right)\right)\right] + \lambda \|w_k\|_1$$

where $\lambda > 0$ is the parameter trading off model sparsity and accuracy, and $\|\bullet\|_1$ is the vector ℓ_1 norm. The objective can be minimized by accelerated gradient descent as described in [29]. However, simple ℓ_1 regression has some limitations in small sample high dimensional case, as well as in the case when there are a group of highly correlated variables.

To overcome these limitations, Zou *et al.* [30] proposed elastic net regularization, which solves the optimal (w_k, b_k) by minimizing the following objective

$$\ell_{enet}^k(w_k, b_k) = \frac{1}{n} \sum_{i=1}^{n_k} \log\left[1 + \exp\left(-y_i^k \left(w_k^T x_i^k + b_k\right)\right)\right] + \lambda_1 \|w_k\|_1 + \lambda_2 \|w_k\|_2$$

where $\lambda_2 > 0$ is a regularization parameter and $\|\bullet\|_2$ is the ℓ_2 norm of a vector.

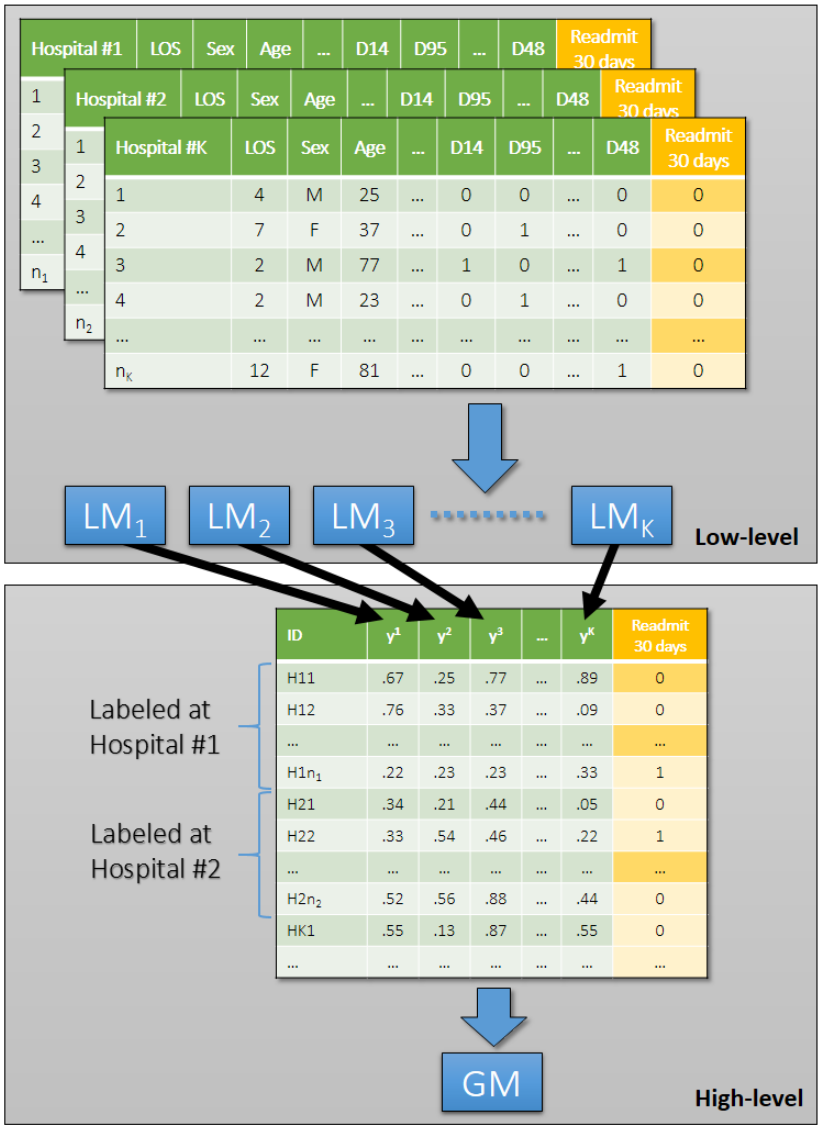


Figure 1. Two-level classification framework for distributed hospital based predictive modeling.

With this formulation, we can get the objective of original logistic regression with $\lambda_1 = \lambda_2 = 0$, sparse logistic regression with $\lambda_2 = 0$, and ridge logistic regression with $\lambda_1 = 0$. $\ell^k_{enet}(w_k, b_k)$ can be minimized by existing software packages such as *Glmnet* [31]. If (w_k^*, b_k^*) is the optimal solution, then the probability that the i^{th} patient in the k^{th} clinical site is re-admitted to hospital within 30 days can be computed as

$$p(y_k^i = 1 \mid x_k^i) = \frac{1}{1 + \exp\left[-(w_k^T x_k^i + b_k)\right]}$$

After we got the optimal prediction model $f_k^*(1 \le k \le K)$ for each site *k*, we collect all those models and form a set

$$F^* = \left\{ f_1^*, f_2^*, \cdots, f_K^* \right\}$$

Those models will be used as the low-level models. Then for each patient vector $x \in R^d$, we can form a *K*-dimensional vector

$$F^*(x) = \left[f_1^*(x), f_2^*(x), \cdots, f_K^*(x) \right]^T$$

We can train another prediction model $g : R^K \to R$ on those *K*-dimensional vectors, which will be used as the decision function on the higher level. Finally, we stack *F* and *g* together to make a classification decision.

Results

This section introduces the experimental settings of all experiments, followed by experimental results and interpretation of the high-level model to demonstrate the effectiveness of the proposed approach.

Experimental Settings

Hospital discharge data from California, State Inpatient Databases (SID), Healthcare Cost and Utilization Project (HCUP), Agency for Healthcare Research and Quality [32] was used in all experiments. The SID is a component of the HCUP, a partnership between federal and state governments and industry, tracking all hospital admissions at the individual level. We used all data from January 2009 through December 2011 in the pre-processing phase. Pediatric patients up to 10 years of age were used in this study due to specific regulations that allow age reporting in months instead of years only for patients younger than 132 months. Patients were excluded from the analysis if they died prior to discharge, were discharged on the same day as admission, were transferred to another institution, or were missing data on the unique patient identifier, age or sex. After pre-processing, we obtained the final dataset containing 66,994 discharge records with 11,184 positive (readmitted within 30 days) and 55,810 negative records.

The bottom 80% of all ICD-9-CM diagnosis codes, ranked by observed frequency, were removed from the dataset, leaving the top 122 diagnosis codes as binary diagnosis features. An additional 21 features (e.g., sex, age, month of admission, length of stay, total charges in USD, etc.) were also included (Table 1). Log transformed values for three numerical features (age, length of stay and total charges in USD) were also included. By recoding nominal features to binary values, we obtained 185 features that were used to build the models presented below.

Each experimental run included randomized training and test dataset where 2/3 of samples were used for training and 1/3 for testing. The holdout testing could cause extremely low number of positive cases in smaller hospitals. Therefore, we removed all records belonging to hospitals with less than 150 records. After removal of records from smaller hospitals, we obtained a final dataset with 61,111 records (10,675 positive and 50,436 negative). Due to a large number non-pediatric hospitals, only 54 out of the initial 205 hospitals were retained for the experiments. The test dataset was not used at all during the process of stacking or model development.

Classification Performance

To evaluate the performance of the built classifiers, we used Area under ROC Curve (AUC) metric. Holdout testing was repeated 1000 times to obtain more robust comparison of the AUC scores. In each run, we calculated the results for the following classification approaches:

- Best performing local model (BLM),
- Simple averaging of the local model outputs (AVG),
- Two-level deep learning architecture (DLA),
- Advanced two-level deep learning architecture with two different local models per hospital - i.e. elastic net and generalized boosted regression models [33] (DLA2) and

- Global sparse logistic regression model (elastic net) built on data from all hospitals. Figure 2 presents the distribution of AUC results for the three compared approaches (SLRA).

It can be observed that plain deep learning approach (AUC=0.781, 95% CI: 0.773-0.789) with a single elastic net classifier on hospital level does not significantly outperform a simple averaging approach (AUC=0.762, 95% CI: 0.746-0.775). However, it does perform much better on average. The proposed approach is also better in comparison to the best performing model from a single hospital (AUC=0.768, 95% CI: 0.758-0.777). We also observed the performance of the weakest local classifiers with an average AUC of 0.416 (95% CI: 0.355-0.481). These results point out that there are hospitals with models that cannot be used for practical application and would gain significantly if they can evaluate the risk for their patients with the proposed approach. When we added an additional, conceptually different model (i.e. generalized boosted models), for each hospital, in DLA2 (AUC=0.787, 95% CI: 0.780-0.794) we were able to significantly outperform AVG and BLM.

We were also interested in how much performance is lost when we compare our approach to a global model that would use all available data (simulating a scenario with no data exchange restrictions between hospitals). It turns out that neither DLA nor DLA2 significantly differ in terms of AUC performance when compared to single elastic net model built on all available data (AUC=0.789, 95% CI:0.781-0.796). On the other hand, both DLA2 and SLRA significantly outperformed averaged and best single models from hospitals.

Table 1. List of 143 features used for building and testing the proposed predictive models.

Feature name	Description	Feature name	Description
DSHOSPID	Unique hospital identifier	PAY1	Primary payer
TOTCHG	Total charge in USD	MEDINCSTQ	Quartile classification of the patient's estimated median household income
AGEMONTH	Age in months (12 - 131)	ASCHED	Scheduled hospitalization
LOS	Length of stay in days	PL_UR_CAT4	Four category urban-rural designation for the patient's county of residence
TOTCHG_LOG	Log transformed total charge in USD	Race	Race and ethnicity (White, Black, Hispanic, Asian or Pacific Islander, Native American, Other)
NPR	Number of procedures on hospital discharge record	MDC	Major Diagnostic Category
NCHRONIC	Number of chronic conditions	ORPROC	Operating room procedures
LOS_LOG	Log transformed length of stay	NECODE	Number of ICD9 E codes
ASOURCE	Source of admission	TRAN_IN	Type of admission
HCUP_ED	Presence of emergency department codes	HospitalUnit	Six hospital unit categories
FEMALE	Identification of gender	D1 – D122	Binary variables for presence of the most frequent diagnoses

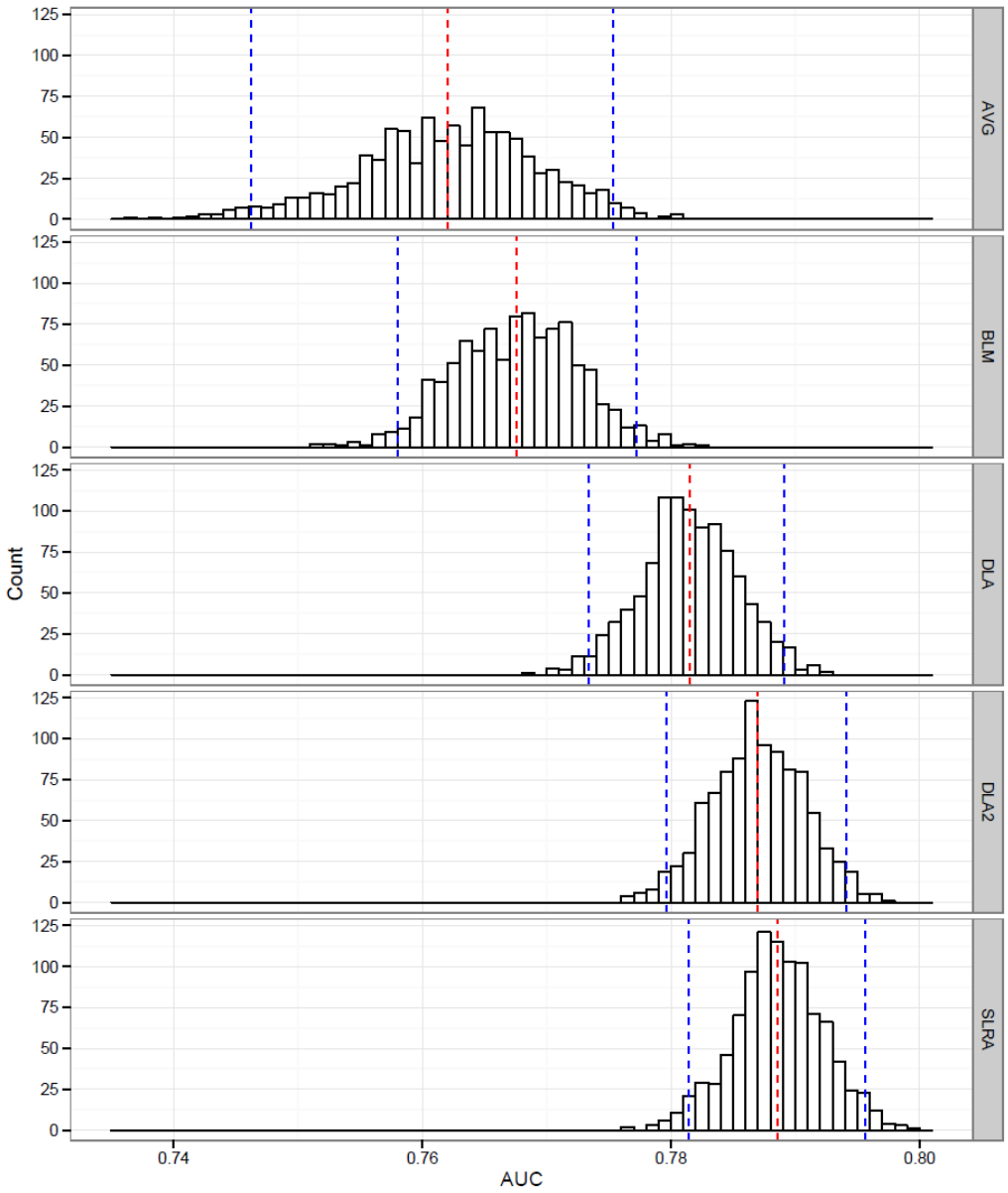


Figure 2. Distribution of AUC results on 1000 hold-out runs for averaged local models (AVG), best local model (BLM), deep learning approach (DLA), deep learning approach with two classifiers (DLA2) and single sparse logistic regression on all samples (SLRA) with mean AUC (red dotted line) and 95% CI (blue dotted line).

Interpretation of the High-level Classifier Results

In contrast to proposed approaches in distributed learning for medical applications [8-12], our approach additionally allows healthcare experts to interpret the results of the high-level classifier. The only constraint is the comprehensibility of the high-level classifier. In case of sparse logistic regression, we can obtain important information on inclusion of the local hospital models in the global model. Nevertheless, this is not the only information we can obtain – one can observe the relative influence of the specific local models in the global solution. For further interpretation of the global regression model, we calculated relative influence of all 54 local models based on their inclusion in the global model in 1000 holdout runs. Relative Hospital Influence (RHI) was calculated as a percentage of holdout runs when the specific hospital was included in the high-level sparse logistic

regression model (i.e. hospital's coefficient was non-zero). Table 2 demonstrates high level of heterogeneity among hospitals.

Table 1. Descriptive overview for 54 hospitals included in the study.

	Min	Max	Median	Mean	SD
Number of records	151	7,884	346.5	1,130.82	1,747.32
Average length of stay	1.90	12.68	3.42	3.95	1.79
Average number of chronic diseases	0.32	3.88	1.50	1.60	0.85
Average total charge (in USD)*	6,615	123,700	32,410	38,230	28,333
Average age of children (in months)	34.76	115.50	55.56	56.97	16.12

*12 hospitals with missing total charge data were excluded

Figure 3 presents correlation of RHI with the most interesting patient characteristics averaged for each of 54 hospitals that were selected based on their increasing or decreasing temporal trend. It can be observed that hospitals with higher cost per patient (TOTCHG) on average contribute more influential models to the final solution. There could be multiple reasons for correlation of influence and cost per patient (larger sample size in such hospitals, specialization of hospitals treating complex conditions, etc.). Therefore, some further analysis would be needed to explain this correlation. Likewise, average number of procedures on patient's discharge records correlates with RHI. This correlation is not difficult to explain as more procedures on the discharge records also means more details for the classification algorithm that can be used. Another positive correlation – i.e. with the average percentage of scheduled patients can be observed in the left lower chart of Figure 3. One should conduct further research into characteristics of hospitals where a large proportion of admission are scheduled to explain this correlation. The only negative correlation presented is the one relating RHI with percentage of children with pneumonia. It turns out that the high-level classifier rarely used outputs from hospitals with lower percentage of pneumonia. It is known from previous studies [34] that prediction of 30-day readmission represents a difficult problem with AUC performance of 0.63. Therefore, it might be possible that models from hospitals with high percentage of pneumonia perform relatively weak and are therefore rarely selected for inclusion in the final model. Readmission rates for hospitals also correlate with the RHI, pointing out that models built on data with stronger class imbalance will have a lower probability of inclusion. The last chart in lower right corner of Figure 3 presents another positive correlation between percentage of patients with gastrostomy (one of the most prevalent diagnoses in the observed population) and RHI. Berry et al. [22] found a similar relation in a study on recurrent readmission within children hospitals, where they confirmed a correlation between readmission frequency and the percentage of technology assistance. The most prevalent technologies among the patients with four or more readmissions were digestive related (30.7%), including gastrostomy tube.

Conclusions

In this study, we present a novel approach to distributed predictive modeling with application to 30-day all-cause readmission in children hospitals. Our approach is based on stacked generalization, dag-stacking and recently proposed deep-learning architectures. Using the proposed approach it is possible to significantly outperform a simple averaging as well as the best performing models from single hospitals. The results demonstrate that there is no significant difference in terms of AUC performance between the global model where data from all hospitals can be used and our approach to distributed predictive modeling. Additionally, our proposed models can be interpreted on high-level, offering an additional insight into the characteristics of specific hospitals. As such, the additional information can be used on policy-making levels to observe hospital quality on a more global level.

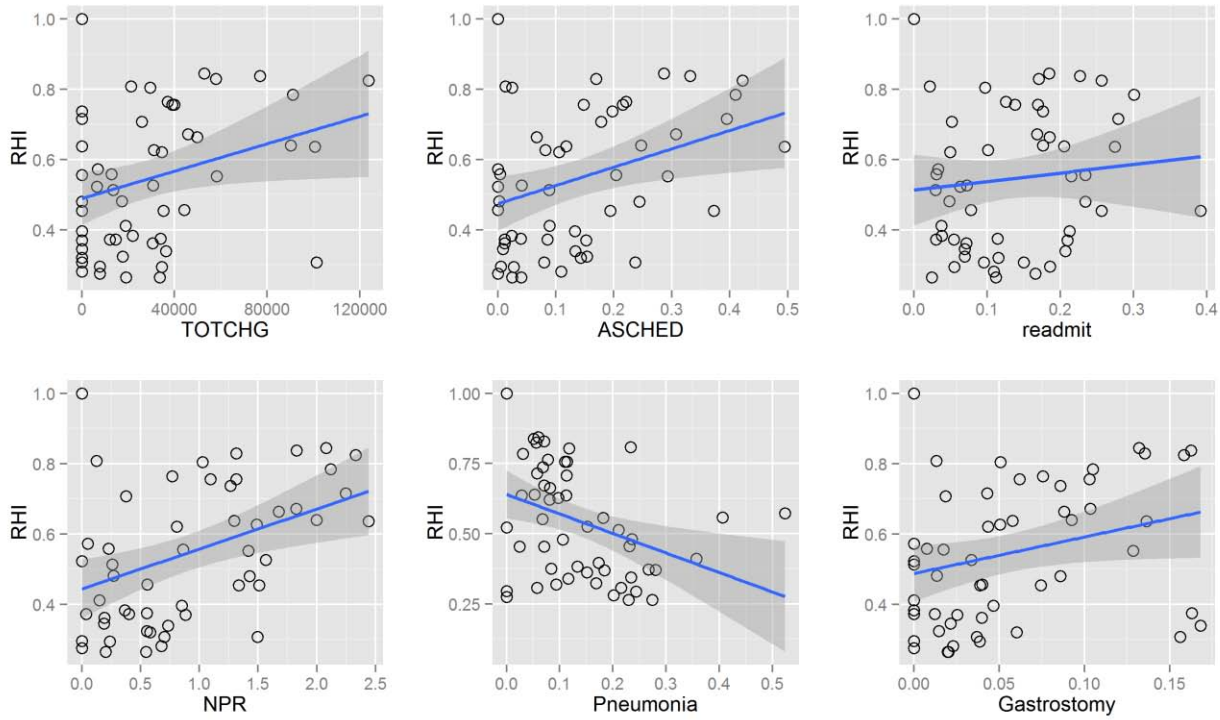


Figure 3. Trends of Relative Hospital Influence (RHI) in relation to average total charge per hospital (TOTCHG), percentage of records with diagnosed pneumonia (Pneumonia), average number of procedure codes on the record (NPR), rate of 30-day readmissions (readmit), percentage of scheduled admissions (ASCHED) and percentage of records with gastrostomy (Gastrostomy).

Acknowledgements

This study was partially supported by the Swiss National Science Foundation through a SCOPES 2013 Joint Research Projects grant SNSF IZ73Z0_152415. We also acknowledge partial financial support from grant #FA9550-12-1-0406 from the U.S. Air Force Office of Scientific Research (AFOSR) and the Defense Advanced Research Project Agency (DARPA). Healthcare Cost and Utilization Project (HCUP), Agency for Healthcare Research and Quality provided data used in this study.

References

1. Cole TS, Frankovich J, Iyer S, LePendu P, Bauer-Mehren A, Shah NH. Profiling risk factors for chronic uveitis in juvenile idiopathic arthritis: a new model for EHR-based research. Pediatric Rheumatology. 2013; 11(1), 45.
2. Sun J, Hu J, Luo D, et al. Combining knowledge and data driven insights for identifying risk factors using electronic health records. AMIA Annu Symp Proc 2012; 2012:901-910.
3. Menachemi N, Collum TH. Benefits and drawbacks of electronic health record systems. Risk Manag Healthc Policy. 2011; 4, 47-55.
4. Coloma PM, Schuemie MJ, Trifirò G, et al. Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. Pharmacoepidemiology and drug safety. 2011; 20(1), 1-11.
5. Davis DA, Chawla NV, Christakis NA, Barabási, AL. Time to CARE: a collaborative engine for practical disease prediction. Data Mining and Knowledge Discovery. 2010; 20(3), 388-415.
6. Stiglic G, Pernek I, Kokol P, Obradovic Z. Disease prediction based on prior knowledge. In Proceedings of the ACM SIGKDD Workshop on Health Informatics, in Conjunction with 18th SIGKDD Conference on Knowledge Discovery and Data Mining. 2012.
7. Wang L, Porter B, Maynard C, et al. Predicting risk of hospitalization or death among patients receiving primary care in the Veterans Health Administration. Medical care. 2013; 51(4), 368-373.
8. Mathew G, Obradovic Z. A privacy-preserving framework for distributed clinical decision support. In IEEE 1st International Conference on Computational Advances in Bio and Medical Sciences (ICCABS). 2011; 129-134.

9. Wang F, Sun J, Ebadollahi S. Composite distance metric integration by leveraging multiple experts' inputs and its application in patient similarity assessment. Statistical Analysis and Data Mining. 2012; 5(1), 54-69.
10. Wang S, Jiang X, Wu Y, Cui L, Cheng S, Ohno-Machado L. EXpectation Propagation LOgistic REgRession (EXPLORER): Distributed privacy-preserving online model learning. Journal of biomedical informatics. 2013; 46(3), 480-496.
11. Rider AK, Chawla NV. An Ensemble Topic Model for Sharing Healthcare Data and Predicting Disease Risk. In Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics (ACM BCB). 2013; 333.
12. Wiens J, Guttag J, Horvitz E. A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. Journal of the American Medical Informatics Association. 2014; doi:10.1136/amiajnl-2013-002162.
13. Ferrucci D, Brown E, Chu-Carroll J, et al. Building Watson: An overview of the DeepQA project. AI magazine. 2010; 31(3), 59-79.
14. Bengio Y. Learning deep architectures for AI. Foundations and Trends in Machine Learning. 2009; vol. 2, pp. 1–127.
15. Wolpert DH. Stacked Generalization. Neural Networks. 1992; vol. 5, pp. 241-259.
16. Alverson BK, O'Callaghan J. Hospital Readmission: Quality Indicator or Statistical Inevitability? Pediatrics. 2013; 132(3), 569-570.
17. Bardach NS, Vittinghoff E, Asteria-Peñaloza R, et al. Measuring hospital quality using pediatric readmission and revisit rates. Pediatrics. 2013; 132(3), 429-436.
18. Srivastava R, Keren R. Pediatric readmissions as a hospital quality measure. JAMA. 2013; 309(4), 396-398.
19. Berry JG, Hall DE, Kuo DZ, et al. Hospital utilization and characteristics of patients experiencing recurrent readmissions within children's hospitals. JAMA. 2011; 305(7), 682-690.
20. Sobota A, Graham DA, Neufeld EJ, Heeney MM. Thirty-day readmission rates following hospitalization for pediatric sickle cell crisis at freestanding children's hospitals: Risk factors and hospital variation. Pediatric blood & cancer. 2012; 58(1), 61-65.
21. Rice-Townsend S, Hall M, Barnes JN, Lipsitz S, Rangel SJ. Variation in risk-adjusted hospital readmission after treatment of appendicitis at 38 children's hospitals: an opportunity for collaborative quality improvement. Annals of surgery. 2013; 257(4), 758-765.
22. Berry JG, Toomey SL, Zaslavsky AM. Pediatric readmission prevalence and variability across hospitals. JAMA. 2013; 309(4), 372-380.
23. Horwitz L, Partovian C, Lin Z, et al. Hospital-wide (all-condition) 30-day risk-standardized readmission measure. Draft measure methodology report. Yale New Haven Health Services Corporation. Center for Outcomes Research and Evaluation (YNHHSC/CORE). 2011.
24. Krumholz HM, Lin Z, Drye EE, et al. An administrative claims measure suitable for profiling hospital performance based on 30-day all-cause readmission rates among patients with acute myocardial infarction. Circulation: Cardiovascular Quality and Outcomes. 2011; 4(2), 243-252.
25. Breiman L. Bagging Predictors, Machine Learning. 1996; vol. 24, pp. 123-140.
26. Ting KM, Witten IH. Stacking Bagged and Dagged Models. In Proc. 14th International Conference on Machine Learning. 1997; 367-375.
27. Ting KM, Witten IH. Issues in Stacked Generalization. Journal of Artificial Intelligence Research. 1999; 10, 271-289.
28. Minka TP. A Comparison of numerical optimizers for logistic regression. 2003; Available at http://research.microsoft.com/en-us/um/people/minka/papers/logreg/minka-logreg.pdf.
29. Liu J, Chen J, Ye J. Large Scale Sparse Logistic Regression. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. 2009; 547-556.
30. Zou H, Hastie T. Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society, Series B. 2005; 301–320.
31. Friedman J, Hastie T, Tibshirani R. Glmnet: Lasso and elastic-net regularized generalized linear models. R package, version 1.9-5. 2013.
32. HCUP State Inpatient Databases (SID), Healthcare Cost and Utilization Project (HCUP). 2009-2011. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/sidoverview.jsp.
33. Ridgeway G. Gbm: generalized boosted regression models. R package, version 2.1. 2013.
34. Lindenauer PK, Normand SLT, Drye EE, et al. Development, validation, and results of a measure of 30-day readmission following hospitalization for pneumonia. Journal of Hospital Medicine. 2011; 6(3), 142-150.

Text Classification towards Detecting Misdiagnosis of an Epilepsy Syndrome in a Pediatric Population

Ryan Sullivan, MS[1], Robert Yao[1], Randa Jarrar, MD[2]
Jeffrey Buchhalter, MD[3], PhD, Graciela Gonzalez, PhD[1]
[1]Arizona State University, Phoenix, AZ; [2]Phoenix Childrens Hospital, Phoenix, AZ;
[3]Alberta Childrens Hospital, Alberta, Canada

Abstract

When attempting to identify a specific epilepsy syndrome, physicians are often unable to make or agree upon a diagnosis. This is further complicated by the fact that the current classification and diagnosis of epilepsy requires specialized training and the use of resources not typically available to the average clinician, such as training to recognize specific seizure types and electroencephalography (EEG)[1–4]. Even when training and resources are available, expert epileptologists often find it challenging to identify seizure types and to distinguish between specific epilepsy syndromes[5]. Information relevant to the diagnosis is present in narrative form in the medical record across several visits for an individual patient. Our ultimate goal is to create a system that will assist physicians in the diagnosis of epilepsy. This paper explores, as a baseline, text classification methods that attempt to correlate the narrative text features to the diagnosis of West syndrome (Infantile Spasms), using data from Phoenix Children's Hospital (PCH). We tested these methods against a dataset containing known (coded) diagnosis of West Syndrome, and found the best performing method to have a precision / recall / f-measure of 76.8 / 66.7 / 71.4 when evaluated with 10-fold cross validation.

Introduction

Epilepsy and misidentification of specific epileptic syndromes has negative implications to public health. Epilepsy affects children and older adults, with 1 in 26 people suffering from it at some point in their lifetime, and about 65 million people affected world-wide[6]. In the U.S., epilepsy is the fourth most common neurologic disorder with a prevalence of 2.2 million and an incidence rate of 150,000 individuals annually. Epilepsy is often poorly managed, and misdiagnosed and untreated in the worst cases[1,6,7]. These epileptic patients may experience difficulties with independent living that includes difficulties in school, uncertainties about employment possibilities, and limitations on driving. About 1 in 10,000 newly diagnosed patients suffer sudden unexpected death in epilepsy[6]. While seizures, epilepsy, and their sequelae have always been present, there have been recent advances in the knowledge and understanding of the disease, its management, and its treatment. Unfortunately, current diagnostic and treatment methods do not adequately capitalize on these advances and patients who could be helped continue to suffer needlessly.

Epilepsy is a complex neurological disorder that manifests as two or more unprovoked seizures of varying types occurring within a 24-hour period[2,8]. An epileptic seizure, also known as an ictal event, is a transient occurrence of signs and/or symptoms that are the manifestations of abnormally excessive synchronous activity of a set of neurons in the brain[7 9,10]. A specific epilepsy syndrome is characterized by a cluster of signs and symptoms that define a unique epilepsy condition and often includes various seizure types[9–12]. To make the diagnosis of epilepsy, clinicians currently utilize the 1981 and 1989 International League Against Epilepsy (ILAE) classifications and clinical case experience. The identification of a specific epileptic syndrome begins with a description of symptoms by the patient and signs by eyewitnesses and requires the inclusion of electroencephalographic recordings of the ictal events. Depending on the type of seizures a person is suffering, along with other diagnostic measures such as age, developmental history, and EEG data, different types of epilepsy syndromes can be diagnosed, each of which requires different treatments. Unfortunately, few clinicians receive the training of expert epileptologists and epilepsy remains unidentified or misdiagnosed in as many as 12 to 23% of all cases leading to a mismatch of epilepsy syndrome to appropriate treatment and management[13].

Concern over epilepsy misdiagnosis has reached a tipping point in recent years, leading to the Institute of Medicine

(IOM) issuing the report, "Epilepsy Across the Spectrum"[6]. In the report, it was recognized that the inability of physicians to make or agree upon a diagnosis consistently for a specific epilepsy syndrome is the major cause of inadequate management of the disease, which in turn, has negatively impacted public health.

As a first step towards better methods for epilepsy syndrome identification, a retrospective analysis was conducted of 27524 patient records referred to Phoenix Children's Hospital (PCH). The objective of this study is to evaluate text classification methods that do not require any additional annotations, and that can correlate the narrative text present in the records (including EEG reports) to a specific diagnosis (West Syndrome).

Background

West Syndrome, also known as infantile spasms was one of the first epilepsy syndromes discovered. It has an incidence rate of 2 to 3.5 per 10,000 live births with 90% of cases occurring in the first year[14]. The diagnosis can be made based on age, developmental history, semiology (observed signs), and EEG patterns. Spasms typically have a neonatal onset and affect boys slightly more than girls between 4 to 8 months of age, but occasionally late onset may occur[15]. Additionally, the child typically has marked developmental delay and mental retardation[14]. Semiologically, a cluster of motor seizures of two major types (spasms and tonic contractions) occur for a duration of less than 1-10 minutes[14,15]. The initial component consists of 2-100 brief epileptic spasms of variable frequency of 1-2 seconds for each spasm[14,15]. These affect primarily the axial muscles of the neck and trunk and appear as characteristic head noddings or "bobbings"[15]. As these seizures occur, various characteristic features on an EEG can be observed. These observations are typically described in detail within the narrative portions of the record and the EEG report.

The clinical features of infantile spasms are considered characteristic, and a diagnosis can be easily made by an expert epileptologist. It is one of the least misdiagnosed epilepsy syndromes, and is thus an excellent case study for the methods proposed, as we expect that most of the true positives will be correctly coded for West Syndrome in the PCH dataset. Still, manual review of false positives is ongoing.

However, it is possible that if certain features co-occur, there may be cases of West Syndrome that may have been missed in the original assessment, and there might be cases that require only one or two additional pieces of information to clinch the diagnosis. An automatic method that could notify the tending physician at the right time could help identify such cases.

Natural Language Processing based Clinical Decision Support has been discussed in Demner-Fushman et al., however Meystre et al. suggest that this area of research is relatively underdeveloped compared to other areas of BioNLP[16,17]. However, there do exist a few similar systems. Yetisgen-Yildiz et al. present a system for the identification of patients with acute lung injury from free-text chest x-ray reports[18]. This system uses a feature set based on unigrams, bigrams and trigrams as well as an assertion analysis system to classify the free-text of x-ray reports. Their best performing classifier configuration achieved a precision / recall / f-measure score of 81.70% / 75.59 % / 74.61 %.

Another system developed by Wagholikar et al. uses NLP techniques to generate cervical cancer screening guidelines from free-text Pap reports[19]. This system used hand-crafted rules to suggest cervical cancer screenings based on the text of the Pap reports, and in their evaluation they found that their system suggested the optimal screening recommendations in 73 of their 74 test cases.

A very similar area of research that has been more supported by the BioNLP community is the automatic coding of medical text. This field of study was the subject of a shared task, which challenged teams to automatically assign ICD-9-CM codes to radiology reports[20]. While a number of the top performing systems in the challenge rely on experts to manually craft rules for their systems, Farkas and Szarvas present that uses machine learning techniques to automatically generate coding rules for their system[21]. Their resulting system achieved a precision / recall / f-measure score of 87.85% / 90.04 % / 88.93 % on the shared task test set, whereas the highest scoring system from the challenged achieved an f-measure of 89.08 %.

Methods

Dataset: A retrospective analysis was conducted of 27524 patient records referred to Phoenix Children's Hospital (PCH). These records consist of patients that have been coded for epilepsy (all ICD9 345 codes) as well as those

with insufficient clinical evidence to support a specific diagnosis of an epilepsy syndrome (patients coded with ICD9 780.39 'Other Convulsions'). We divided the patient records into three groups: 1) 144 patients coded for Infantile Spasms (ICD9 codes 345.60 and 345.61); 2) 2818 patients with records that contain Infantile Spasm-related keywords ["infantile spasms", "tuberous sclerosis", "hypsarrythmia", "ACTH", "prednisolone" and "aicardi syndrome"], but are not coded for Infantile Spasms; and 3) 27524 patients with records neither coded for infantile spasms nor containing Infantile Spasm-related keywords. For our experiments, we created a corpus consisting of the records of the 144 patients coded for infantile spasms as positive examples and the records of 3600 randomly chosen patients from group three as negative examples.

Preprocessing: For this classification task, we only used the free-text from discharge summaries and EEG reports. We used a simple tokenizer to tokenized the text at whitespace and punctuation, and removed all digits and special characters. We also removed all Infantile Spasm-related keywords and English stopwords from the text.

Feature generation: We tested two different techniques for generating a feature set from the patient record free-text, TF-IDF vectors and a topic distribution based on Latent Dirichlet Allocation (LDA).

Our first feature set consisted of TF-IDF (term frequency-inverse document frequency) vectors, a popular scheme that is generally used for indexing documents for information retrieval[22]. For each term in a document, the term frequency (how many time a term appears in the document) and the inverse document frequency (a measure of how rare a term is across documents) is calculated, and these values are used to construct a term vector representation of the document.

Our second approach consisted of representing each patient as a topic distribution based on Latent Dirichlet Allocation (LDA)[23]. LDA is an unsupervised technique used for topic discovery and text classification. It assumes that each document is generated based on some topic distribution and each topic's word distribution. It then attempts to use the observed information, i.e. the words in the documents, to predict the unobserved information, i.e. the topics, the topic distribution and the word distributions within each topic. We learned an LDA topic model on the patient data consisting of 1500 topics, with the topic size chosen based on testing a subsample of the corpus, and we used the corpus to estimate the topic distribution for each record. We used the Mallet toolkit to build the LDA models[24] We used this topic distribution as a feature set, in a method that tantamount to using LDA for dimensionality reduction and is mentioned in Blei and McAuliffe[25].

Training data sampling: One issue we encountered with the dataset is the relative rarity of Infantile Spasm patients compared to the negative examples. Because of this disparity, we ran into issues of having too few positive examples to train our classifiers.

We tested two solutions to this problem, oversampling the positive classes and undersampling the negative classes. In both cases, we attempted to have a 3 to 1 ratio of negative to positive examples (compared to the 25 to 1 ratio of our corpus). In oversampling, we keep the number of negative examples the same, but give the positive examples a weight of 8.33 times the negative examples. In undersampling, we trained the model on a random subset of the data which has the 3 to 1 ratio.

Classification and Evaluation We compared two different classification algorithms for this task; a multinomial Naïve Bayes classifier and a Support Vector Machine classifier[26,27]. These classifiers were chosen for their speed and for their documented performance in text classification tasks.

We evaluated our classification models using 10-fold cross validation, and we trained and evaluated each classification model using Weka[28].

Results and Discussion

The results of our experiment can be seen in Table 1.

These results do not overtake the top performing systems, yet they are compatible to other systems in this domain, and they represent a reasonable baseline for which to continue research. Furthermore, these results show that the use of domain knowledge is not a necessary requirement to achieve reasonable results. There are also a few conclusions that can be drawn from these results. The first conclusion is that a sampling method (either over or under sampling) is a

Table 1: Classification results.

Classifier	*Precision*	*Recall*	*F-Measure*
multinomial Naïve Bayes - LDA - no sampling	16.5%	61.8%	26.1%
multinomial Naïve Bayes - TF*IDF - no sampling	19.4%	53.5%	28.5%
SVM - LDA - no sampling	45.0%	34.7%	39.2%
SVM - TF*IDF - no sampling	55.1%	29.9%	38.7%
multinomial Naïve Bayes - LDA - oversampling	47.4%	75.7%	58.3%
multinomial Naïve Bayes - TF*IDF - oversampling	77.5%	55.6%	64.7%
SVM - LDA - oversampling	81.4%	29.9%	43.7%
SVM - TF*IDF - oversampling	88.2%	29.9%	44.6%
multinomial Naïve Bayes - LDA - undersampling	50.0%	70.8%	58.6%
multinomial Naïve Bayes - TF*IDF - undersampling	59.7%	75.0%	66.5%
SVM - LDA - undersampling	70.1%	65.3%	67.6%
SVM - TF*IDF - undersampling	76.8%	66.7%	71.4%

requirement to get usable performance. However, from analyzing the results, it seems that oversampling may be over fitting the data, which causes the precision-recall difference between the over and under sampling.

We can also gather from these results that LDA is an effective dimension reduction technique for this type of text, when using a SVM classifier. Though the TF-IDF features outperform the LDA-based features, the TF-IDF vectors are 10 times as large, and represent a non-trivial computation time cost for certain classifiers. Finally, we can see that SVM consistently had high precision-low recall compared to the Naïve Bayes' comparatively low precision-high recall. Though one would ideally want both high precision and high recall, these classifier tendencies are worth keeping in mind depending on the task.

Conclusion

The use of Natural Language Processing for Clinical Decision Support is an academically underserved domain that holds a large potential. The results of our baseline system are comparable to other systems in the domain, and can identify misdiagnosed epilepsies or improve the ability for medical doctors to identify an epilepsy syndrome not previously diagnosed. This can potentially improve the match between patient and the best treatment available for a specific epilepsy syndrome, and can further the goal for better seizure control and improvement in quality of life for the patient.

References

1. Birbeck, G.L.. Revising and refining the epilepsy classification system: Priorities from a developing world perspective. Epilepsia 2012;53(s2):18–21.
2. Berg, A.T., Berkovic, S.F., Brodie, M.J., Buchhalter, J., Cross, J.H., Van Emde Boas, W., et al. Revised terminology and concepts for organization of seizures and epilepsies: report of the ilae commission on classification and terminology, 2005–2009. Epilepsia 2010;51(4):676–685.
3. Scheffer, I.E.. Epilepsy: A classification for all seasons? Epilepsia 2012;53(s2):6–9.
4. Berg, A.T., Cross, J.H.. Towards a modern classification of the epilepsies? The Lancet Neurology 2010;9(5):459–461.
5. Ottman, R., Hauser, W.A., Stallone, L.. Semistructured interview for seizure classification: agreement with physicians' diagnoses. Epilepsia 1990;31(1):110–115.
6. England, M.J., Liverman, C.T., Schultz, A.M., Strawbridge, L.M.. Epilepsy across the spectrum: Promoting

health and understanding.: A summary of the institute of medicine report. Epilepsy & Behavior 2012;25(2):266–276.

7. Thurman, D.J., Beghi, E., Begley, C.E., Berg, A.T., Buchhalter, J.R., Ding, D., et al. Standards for epidemiologic studies and surveillance of epilepsy. Epilepsia 2011;52(s7):2–26.
8. Blume, W.T., Lüders, H.O., Mizrahi, E., Tassinari, C., van Emde Boas, W., Engel, J.. Glossary of descriptive terminology for ictal semiology: report of the ilae task force on classification and terminology. Epilepsia 2001;42(9):1212–1218.
9. Engel, J.. A proposed diagnostic scheme for people with epileptic seizures and with epilepsy: report of the ilae task force on classification and terminology. Epilepsia 2001;42(6):796–803.
10. Engel Jr, J.. Ilae classification of epilepsy syndromes. Epilepsy research 2006;70:5–10.
11. Epilepsy, A.. Proposal for revised classification of epilepsies and epileptic syndromes. The treatment of epilepsy: principles & practice 2006;354.
12. Wolf, P.. Basic principles of the ilae syndrome classification. Epilepsy research 2006;70:20–26.
13. Network, S.I.G.. Diagnosis and management of epilepsies in children and young people; guideline no. 81 ed. Royal College of Physicians; 2005.
14. Pellock, J.M., Hrachovy, R., Shinnar, S., Baram, T.Z., Bettis, D., Dlugos, D.J., et al. Infantile spasms: a us consensus report. Epilepsia 2010;51(10):2175–2189.
15. Shields, W.D.. Infantile spasms: little seizures, big consequences. Epilepsy Currents 2006;6(3):63–69.
16. Demner-Fushman, D., Chapman, W.W., McDonald, C.J.. What can natural language processing do for clinical decision support? Journal of biomedical informatics 2009;42(5):760–772.
17. Meystre, S.M., Savova, G.K., Kipper-Schuler, K.C., Hurdle, J.F.. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform 2008;35:128–44.
18. Yetisgen-Yildiz, M., Bejan, C.A., Wurfel, M.M.. Identification of patients with acute lung injury from free-text chest x-ray reports ????;.
19. Wagholikar, K.B., MacLaughlin, K.L., Henry, M.R., Greenes, R.A., Hankey, R.A., Liu, H., et al. Clinical decision support with automated text processing for cervical cancer screening. Journal of the American Medical Informatics Association 2012;19(5):833–839.
20. Pestian, J.P., Brew, C., Matykiewicz, P., Hovermale, D., Johnson, N., Cohen, K.B., et al. A shared task involving multi-label classification of clinical free text. In: Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing. Association for Computational Linguistics; 2007, p. 97–104.
21. Farkas, R., Szarvas, G.. Automatic construction of rule-based icd-9-cm coding systems. BMC bioinformatics 2008;9(Suppl 3):S10.
22. Salton, G., McGill, M.J.. Introduction to moderm information retrieval. 1983.
23. Blei, D.M., Ng, A.Y., Jordan, M.I.. Latent dirichlet allocation. the Journal of machine Learning research 2003;3:993–1022.
24. McCallum, A.K.. Mallet: A machine learning for language toolkit 2002;.
25. Blei, D.M., McAuliffe, J.D.. Supervised topic models. arXiv preprint arXiv:10030783 2010;.
26. McCallum, A., Nigam, K., et al. A comparison of event models for naive bayes text classification. In: AAAI-98 workshop on learning for text categorization; vol. 752. Citeseer; 1998, p. 41–48.
27. Platt, J.C.. 12 fast training of support vector machines using sequential minimal optimization 1999;.

28. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.. The weka data mining software: an update. ACM SIGKDD Explorations Newsletter 2009;11(1):10–18.

Adding flexible temporal constraints to identify chronic comorbid conditions in ambulatory claims data

Walton Sumner, MD[1], Dustin L Stwalley, MA[1], Phillip V. Asaro, MD[1], Michael D. Hagen, MD[2], Margaret A Olsen, PhD, MPH[1]

[1]Washington University School of Medicine, St. Louis, MO;
[2]American Board of Family Medicine, Lexington, KY

Abstract

Chronic comorbid conditions are important predictors of primary care outcomes, provide context for clinical decisions, and are potential complications of diseases and treatments. Comorbidity indices and multimorbidity categorization strategies based on administrative claims data enumerate diagnostic codes in easily modifiable lists, but usually have inflexible temporal requirements, such as requiring two claims greater than 30 days apart, or three claims in three quarters. Table structures and claims data search algorithms were developed to support flexible temporal constraints. Tables of disease categories allow subgroups with different numbers of events, different times between similar claims, variable periods of interest, and specified diagnostic code substitutability. The strategy was tested on five years of private insurance claims from 2.2 million working age adults. The contrast between rarely recorded, high prevalence diagnoses (smoking and obesity) and frequently recorded but not necessarily chronic diagnoses (musculoskeletal complaints) demonstrated the advantage of flexible temporal criteria.

Introduction

Comorbid conditions are extremely important predictors of primary care outcomes[1-3]. In addition, comorbid conditions may complicate disease management, and may even result from disease management.

Comorbidity indices merge numerous clinical observations into a single score that predicts an outcome. The Charlson score, published in 1994, predicted mortality in patients undergoing elective surgery based on 19 categories of comorbid conditions[4]. The Elixhauser comorbidity index, published in 1998, defines 30 categories of comorbid conditions for use with administrative claims data[5]. These categories comprise sets of diagnosis codes from the International Classification of Diseases, version 9, Clinical Modification (ICD-9-CM). The Elixhauser index predicts mortality, charges, and length of a hospitalization. Klabunde, et al described in 2000 how outpatient records could identify comorbid conditions that were undocumented during hospitalization[6] but that would affect hospital outcomes[7].

While comorbidity indices simplify predictive analyses, the complexity of comorbidity, often called multimorbidity, is more relevant to many clinical decisions. Analysis of multimorbidity requires that thousands of very specific diagnostic codes be mapped into scores or hundreds of categories of conditions. Several reports describe categorization strategies for common comorbid conditions in claims data[8-10]. Even crude measures of multimorbidity, e.g. counts of comorbid conditions, are consistently predictive of outcomes.

Several issues distinguish the identification of comorbidities from ambulatory vs. hospital claims data. First, omissions are common in ambulatory claims. Outpatient claims headings in the USA are restricted to 4 diagnostic codes, although "line codes" may include an additional diagnostic code for each procedure claimed. Inpatient claims headings currently are restricted to 25 diagnostic codes for Medicare claims, but have been as few as 9 codes in the past. Furthermore, diagnoses such as smoking and obesity may be absent for years in outpatient claims, but then be recorded when patients have elective surgery. Undocumented diagnoses may appear to (i) be tangential or subordinate to acute problems, (ii) lack effective treatments, leading to lack of attention, (iii) lack documentation incentives, such as reimbursement or quality measures, (iv) carry a risk of stigmatizing patients, or (v) carry a risk of causing a confrontation when patients see the code descriptions on billing records. Thus, inpatient claims from one hospitalization provide a more complete picture than office claims from one visit. Practical limits on numbers of claims imply that (a) any short series of outpatient visits can fail to document relevant comorbidities and (b) diagnoses that are often missed, neglected, or ignored could appear infrequently in claims. For these reasons, long periods of surveillance may be appropriate.

Second, an outpatient claim does not indicate a definite diagnosis. One problem is that clinicians may record codes for diagnoses being ruled out. Another issue is that clinicians often select codes with little direction and only delayed feedback through reimbursement. Professional coders assign inpatient claims, following guidelines that should

increase the odds that the diagnosis is present. Thus, inpatient claims could be more accurate than office claims, although both may be biased toward codes yielding higher reimbursement[11-13]. Finally, diagnostic conclusions can be wrong[14-16]. When diagnostic errors are discovered, the incorrect diagnosis should disappear from the record. Ambulatory claims analyses therefore require some strategy for establishing validity of outpatient claims. One rule of thumb requires a diagnosis to appear twice, more than 30 days apart, to infer that an outpatient diagnosis is real[17]. A more stringent German algorithm required a diagnosis to appear once each quarter (3 months) for 3 consecutive quarters[10]. Of course, some chronic diseases can actually resolve, especially those related to weight, diet, substance abuse, and other reversible problems. The three preceding quarters defines a moving window that eventually removes incorrect and resolved diagnoses.

Third, some broad comorbidity categories used to predict hospital mortality warrant dissection for primary care quality and outcomes analyses. Distinctions between organ impairment and failure are important for renal and hepatic diseases. Patients with injured organs need their physicians to help them minimize the rate at which further injury accumulates, while patients with overt organ failure often face different risks. For instance, the risk of cardiovascular disease initially rises slowly as renal function deteriorates but becomes extreme as kidneys finally fail[18, 19]. Hospital mortality-oriented comorbidity classifications may combine diverse, common neurologic diseases that primary care analyses should subdivide, such as dementia, headaches, seizures, and stroke.

Fourth, many conditions could influence treatments, costs, and outcomes in primary care, or result from medical treatment, without affecting hospital mortality. Topics of primary care include allergies, anxiety, esophageal reflux, migraine and tension headaches, chronic musculoskeletal problems, osteoporosis, smoking, and personality disorders. For instance, osteoporosis may have important outpatient implications (such as increasing the risk of using of corticosteroids) without affecting hospital mortality. Conversely, diagnoses with dire implications for hospital mortality may be unmanageable in the outpatient context.

In order to analyze outpatient claims data describing family physicians' management of specific diseases, we needed to classify patients' comorbid conditions. Due to the issues just enumerated and other experiences with outpatient claims, we believed that the inference of comorbid conditions from outpatient data could be further refined. We particularly wanted to increase the temporal flexibility of comorbidity definitions in response to common documentation patterns. We report the development and testing of a chronic disease list with temporal criteria for analyzing outpatient claims data.

Methods

Categories of diagnosis codes

The Elixhauser [5] and German ambulatory care [10] lists of comorbid conditions were reviewed by category. Categories that included common problems with very diverse clinical implications were subdivided into categories with more uniform implications. Categories that were unlikely to occur in primary care were deleted, and categories for common and clinically influential problems were added. ICD-9-CM diagnosis codes were sought using the hierarchical structure of ICD, and by searching for specific codes and string values.

We determined that assigning codes to a single comorbidity category still limited our ability to make higher resolution inferences that would be desirable in some analyses. We defined *subgroups* within categories, including a "subgroup zero" containing codes that could match any other subgroup within a category.

After defining subgroups, we noted that requiring repeated documentation of codes in a subgroup could lead to false positive matches unless mutually substitutable codes were specified. We therefore defined *bundles* of codes that we considered mutually substitutable. For instance, the musculoskeletal group has a subgroup for joint problems. Within the joint subgroup, there are several codes for arthritis of the shoulder and others for arthritis of the hip. The shoulder codes are mutually substitutable and belong to one bundle; the hip codes are in a different bundle. Again, a "bundle zero" identified very imprecise codes that could match any bundle in the subgroup, e.g. "arthritis, NOS."

Timing of claims

Flexible temporal constraints were added to subgroup definitions. First, we provided minimum numbers of outpatient events as a subgroup specification. Fixed requirements for at least two or at least three outpatient claims over a period were judged unlikely to work consistently. More or fewer claims could be appropriate. For diseases that are rarely documented at all and normally are recorded after being "ruled in," such as smoking, obesity, and personality disorders, one appearance in an outpatient claim suggests relevance for years before and after the claim. Conversely, an acute problem recurring many times over a few years could be considered a chronic problem.

Second, we added minimum separation between outpatient events as a subgroup specification. Any fixed temporal separation may be excessively or insufficiently restrictive for some condition. For instance, a patient presenting with angina could be evaluated and receive an effective medical intervention in less than 30 days. If chest pain resolves and other risk factors command attention at subsequent visits, angina codes could disappear from claims records. Nevertheless, the patient would have evidence of coronary artery disease as a comorbid condition. Clinical scenarios requiring longer minimum intervals are rare, but consider provoked and unprovoked venous thrombosis. Provoked venous thrombosis will generate a series of claims over 3 to 6 months. Clotting disorders will generate claims over longer intervals, due to recurrence of clots. Therefore, a minimum interval of 90, 180, or even 270 days between venous thromboembolic claims could be appropriate when attempting to infer a clotting disorder.

Third, we added a maximum separation between outpatient events as a subgroup specification. Most algorithms have not specified maximum separation between codes, but some maximum separation is often justifiable. Invasive breast cancer codes may appear annually in claims for women aged 40 to 60; this probably does not indicate an active neoplasm, but the intent of the diagnostic test (to identify preclinical breast cancer). Discovery of a new neoplasm should generate a series of claims at short intervals. At the other extreme, a seizure disorder documented at annual visits could indicate a stable patient receiving annual medication refills.

Finally, we added persistence, a time to look back through for the other requirements, as a subgroup specification. This change relaxes the fixed three quarters implemented in the German ambulatory care algorithm. In a long series of claims data, some potentially chronic diseases will remit: weight loss will cure some diabetics and abstinence will cure some alcoholics' hepatitis. However, if an incurable diagnosis disappears from the series, then the diagnosis is in question. For instance, Parkinson's disease is currently incurable and debilitating: it should appear regularly in a patient's claims. If the diagnosis was not recorded in the most recent two years, then any earlier diagnosis of Parkinsons' disease is doubtful.

Table structure

We implemented these constraints in two tables, subgroups and codes. The subgroup table has these fields, with each combination of Group and Subgroup being unique:

Group	positive integer identifier
Abbrev	short name for the group
Subgroup	non-negative integer identifier
Name	short name for the subgroup
IP Count	number of inpatient claims (within a bundle) required to establish the subgroup; zero if inpatient claims never establish the subgroup being present
OP Count	number of outpatient claims (within a bundle) required to establish the subgroup
MinSep	minimum number of days between claims (within a bundle)
MaxSep	maximum number of days between claims (within a bundle)
Persist	maximum number of days to look back from a specified date

The code table has these fields:

Key	a unique row number
Group	positive integer identifier
SubGroup	non-negative integer identifier
Bundle	non-negative integer identifier
Code	ICD9-CM code
Description	description of the ICD9-CM code

General approach to specifying groups and subgroups

Given the above considerations and the lack of standardization in assignment of diagnostic codes for outpatient billing claims, we found that subgroup specifications were quite subjective, especially in regard to time intervals. We therefore developed some general principles for assigning subgroups' attributes.

- Number of instances required
    - One inpatient code may establish a clearly chronic condition
    - Inpatient codes for potentially acute or iatrogenic conditions may be ignored, so that the condition must be established by outpatient codes
    - One outpatient code may establish a clearly chronic condition that is under-documented

- Two outpatient codes are needed to infer most chronic conditions
- More than two outpatient codes may be used to infer that a normally acute condition is functionally chronic

- Separation of outpatient codes
  - As a default setting, pairs of claims are at least 30 days apart and not more than 180 days apart
  - Acute events that imply chronic problems may have shorter minimum separation requirements (strokes ad heart attacks)
  - Chronic conditions that could be stabilized and managed with annual checks were set 400 days (13 months) apart (seizure disorders)
  - Infrequent events that imply chronic problems may be set farther apart (smoking, obesity), but in the USA, private claims data may not capture two infrequent events due to limited periods of enrollment (3 years on average)
- Time since last diagnosis (persist field)
  - Incurable, confidently diagnosed, and rarely recorded diseases may persist indefinitely (strokes)
  - Potentially curable chronic diseases should be confirmed by periodic reappearance (restless leg syndrome) or change to "history of" codes (neoplasms)
  - Inexorably progressive diseases should be confirmed by periodic reappearance (degenerative neurologic diseases)
  - Rarely recorded but curable diseases should be inferred from appearance over a long but not indefinite interval (smoking, obesity)

Algorithm

These definitions reflect the possibility of evolving comorbidity status in primary care analyses of claims data. In procedural pseudo code, the logic for interpreting these data is:

```
Establish assessment date D
Get individual's claims data
For each Group G in the subgroup table
    Put claims for Group G, Subgroup zero, Bundle zero into wild card Subgroup claims
    For each Subgroup S>0 in Group G
        Put claims for Group G, Subgroup S, Bundle zero into wild card Bundle claims
        For each Bundle B>0 in Group G, Subgroup S
            Put claims for Group G, Subgroup S, Bundle B into Bundle B claims
            Merge wild card Subgroup and Bundle into the Bundle B claims list
        If Group G, Subgroup S defines Persist>0 then
                Limit Bundle B claims list to dates between D - Persist and D
            End if
            If Group G, Subgroup S defines IPCount>0 then
                If number of Inpatient claims in Bundle B claims>IPCount then
                    Patient has comorbidity Group G
                    Patient has comorbidity Subgroup S
                    Next Subgroup
                End if
            End if
            For each claim in the claims list
                Put claim into Sequence Q
                For each subsequent claim in the claims list
                    Get days between last claim in Q and this subsequent claim
                    If days >= SepMin and days <= SepMax then
                        Add claim to sequence Q
                    End if
                    If number of claims in Q > OP count for Group G, Subgroup S then
                    Patient has comorbidity Group G
                        If at least one claim in Q is in Bundle B then
                            Patient has comorbidity Subgroup S
                            Next Subgroup
```

Trial

We implemented the search algorithm and applied it to five years of private insurance claims data from three states in the Truven Health Analytics MarketScan® Commercial Claims and Encounters Databases. Medical claims were restricted to (i) persons aged 18 to 64 years and (ii) claims with a specialty code for general internists or family physicians, thus excluding facility, laboratory, and other specialty claims. Comorbidity prevalence was calculated for each category and subgroup. Patient level prevalence over the five years was calculated for each ICD-9CM claim that was not categorized as a chronic comorbidity. These codes were reviewed manually to identify (a) codes that were included in the codes table but failed temporal criteria, (b) codes that should have been included in the codes table, and (c) codes that were reasonably excluded from the codes table. After reviewing results from the preliminary draft, some comorbid conditions were added and temporal criteria were modified for others. The search algorithm was refined to obtain the form described above and applied to the revised tables to obtain the ensuing results.

Results

Categories, subgroups and bundles

We defined 50 categories with 154 specific subgroups and 13 wild card subgroups. The 50 categories included 2,733 ICD9-CM codes in 367 bundles. The most commonly assigned temporal requirement was 2 claims 30 to 400 days apart, over the past 1 or more years (67 uses), followed by 2 claims 30 to 180 days apart, over the past 1 or more years (33 uses) and 2 claims 30 to 730 days apart, over the past 2 or more years (10 uses). Persistence was about equally divided between 1, 2, 3, 5, and indefinite time spans. Three subgroups required a single claim within 4 or 5 years. Six subgroups required three or more claims.

Test run results

The test run analyzed records of up to five years' duration from 2.2 million adult patients seen by family physicians and/or general internists between 2006 and 2012. The average duration of enrollment was 2.5 years.

Table 1 summarizes the frequency with which patients met criteria for the 50 categories, sorted from most to least frequent positive matches as a percentage of the 2.2 million patients. The number of subgroups and ICD codes specified in each category is given. The "Time Criteria Fails" column lists the percentage of patients who had one of the codes in the group documented, but a time criterion was not met. The largest of these, musculoskeletal diagnoses, involves 12% of patients. An extremely low number of patients were documented to be smokers (0.83%,) or obese or morbidly obese (1%). Two thirds of the smokers and half of the obese patients had exactly one claim with the diagnosis. In contrast, the population prevalence of smoking is 5% for individuals with graduate degrees [20], and above 20% for all workers [21], and obesity affects nearly 30% of the workforce[22].

Figure 1 plots time criteria failures (Table 1, Column 5) as a function of comorbidity prevalence (Table 1, Column 2) for each category. For instance, the outlier point at (2%, 12%) is the musculoskeletal category. The point indicates that 2% of patients have chronic musculoskeletal conditions, as defined here. Another 12% have diagnostic codes included in the musculoskeletal category, but did not meet specified time constraints. For instance, a person having two claims related to back injuries occurring less than 20 days apart would fail the temporal criteria (see table 2). Other musculoskeletal subgroups require more than two claims.

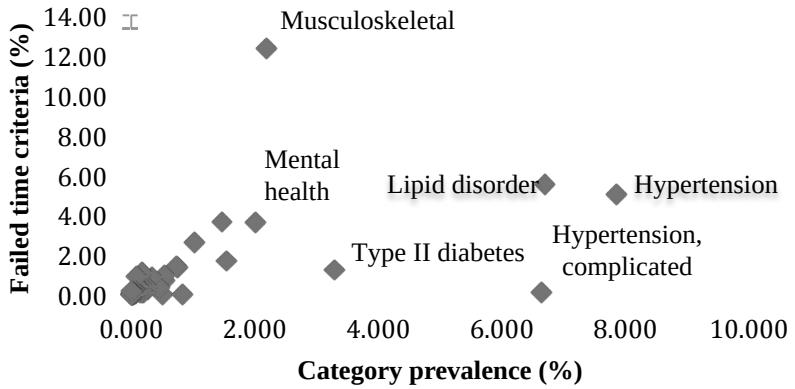
Table 2 summarizes the main subgroup's requirements in each category. Usually, the other subgroups in a category will have comparable constraints. The number of Inpatient (*IP*) and Outpatient (*OP*) events required is listed. Zero IP events means that only outpatient diagnoses are accepted. Zero OP events means that the subgroup is a wildcard. The *min*imum and *max*imum separation between events is listed, and the persistence of a diagnosis is listed in the *look back* column. The number of distinct ICD-9CM *codes* used and number of *bundles* in each main subgroup completes the table.

Claims that did not match categorization criteria involved 9,158 distinct ICD-9CM codes, of which 7,316 codes did not map to any of the 50 categories. The remaining 1,842 matched a code in a category, but were not in a series of claims that met temporal requirements. Table 3 lists the uncategorized codes affecting at least 0.5% of the 2.2 million people in the sample. Codes that are included in a comorbidity category are in bold italics.

Table 1. Categories with most frequently matched subgroups

Categories	+(%)	Sub-groups (#)	ICD9 Codes (#)	Time Criteria Fails (%)	Main subgroup	+ (%)
Hypertension	7.9	1	2	5.0	Benign	7.9
Lipid disorder	6.7	1	11	5.5	Hyperlipidemia	6.7
Htn, complicated	6.6	4	45	0.10	Other	6.4
Type II diabetes	3.3	4	28	1.2	Uncomplicated	3.1
Musculoskeletal	2.2	8	295	12.3	Back	1.1
Mental health	2	6	175	3.6	Depression	1.1
Endocrine	1.5	3	46	1.7	Hypothyroid	1.3
Unexpl. illnesses	1.5	2	31	3.6	Functional Somatic	1.2
Gastrointestinal	1	6	106	2.6	GE reflux disease	0.82
Body mass index	1	2	26	0	Obesity	0.78
Smoking	0.83	1	6	0	Tobacco	0.83
Asthma	0.75	2	14	1.4	Reactive airway dis	0.73
Sleep problems	0.73	4	73	1.4	Other	0.51
Anemia	0.54	4	36	0.73	Other	0.27
Headaches	0.53	5	101	0.95	Migraine	0.52
Coronary art dis	0.45	4	48	0.35	Ischemia	0.41
Chr. Obst. Lung	0.33	3	8	0.86	Other	0.19
Rheumatic dis.	0.26	3	44	0.25	Rheum arthritis	0.13
Rhythm disturb	0.25	5	19	0.41	Atrial fibrillation	0.13
Type I diabetes	0.19	2	14	0.12	Uncomplicated	0.16
Allergy	0.18	5	51	0.44	Contact	0.12
Infection	0.18	5	23	1.1	Sinusitis	0.17
Skin diseases	0.18	3	48	1.1	Various	0.16
Chr. kidney disease	0.17	2	61	0.08	Stage I-IV	0.16
Neoplasms	0.15	2	437	0.26	Local	0.15
Coagulopathy	0.14	3	63	0.10	VTE	0.057
Liver disease	0.13	3	40	0.35	Cirrhosis/Enceph.	0.051
Bone diseases	0.13	1	6	0.21	Osteoporosis	0.13
Sexual dysfunction	0.11	1	10	0.34	Various	0.11
Genitourinary	0.11	4	33	0.40	Calculi	0.074
Heart valve	0.093	5	40	0.17	Mitral valve	0.046
Neurology	0.086	7	42	0.92	Multiple sclerosis	0.038
Pain syndromes	0.072	2	15	0.08	Neuropathy	0.069
Pulm emboli/htn	0.071	3	8	0.03	Pulm embolism	0.058
Seizures	0.068	7	29	0.14	Traumatic seizures	0.037
Heart failure	0.063	3	33	0.04	Hypertensive HF	0.062
Periph vasc disease	0.045	7	66	0.08	Atherosclerosis	0.038
HIV	0.041	1	12	0.00	HIV	0.041
Alcohol misuse	0.033	3	26	0.07	End organ damage	0.019
Drug misuse	0.032	2	73	0.03	Dependence	0.024
Stroke	0.03	3	24	0.03	Thromboembolic	0.021
Lymphoma	0.021	3	254	0.00	Non-Hodgkins	0.015
Reproductive	0.019	1	38	0.00	Various	0.019
Paralysis	0.012	4	63	0.02	Syndromes	0.004
Heart blocks	0.008	5	21	0.04	AV block	0.002
Nutrition	0.006	2	11	0.01	Anorexia / Bulimia	0.003
Dementia	0.005	5	35	0.01	Frontal dementias	0.003
Cystic Fibrosis	0.005	1	9	0.01	Various	0.005
Restrictive pulm dis	0.002	2	15	0.01	Various	0.002
Sensory losses	0.001	1	68	0.18	Vision loss	0.001

Figure 1 Prevalence vs. time criteria failure rate for major categories



Discussion

We present a refinement in strategies for defining comorbid conditions based on provider (outpatient) rather than facility (inpatient) claims. The disease categories presented here are familiar from previous work, especially by Charlson[4], Elixhauser[5], and van den Bussche[10]. We extended these systems by adding flexible temporal constraints, including the option to require any number of instances of a diagnosis in provider claims. The value of this flexibility in required numbers was evident in the test run. Smoking and obesity documentation is so sparse that less than 5% of cases are documented during the average 2.5 years of claims data, and half of the documented cases were based on only one claim. Requiring two or three claims to confirm these diagnoses would risk eliminating nearly all documentation of these fundamental problems. Conversely, six out of seven patients with musculoskeletal claims do not generate enough claims to be categorized as chronic conditions – these problems are usually of short duration.

In addition, we have explicitly described three levels of hierarchy, starting with 50 broad categories, followed by 154 specific subgroups. The subgroups have much more uniform primary care management implications than the categories. The diagnoses in many subgroups share similar pathology, treatment strategies, complication risks, and morbidity implications. The third level of the hierarchy consists of substitutable bundles of diagnoses within subgroups. Given the frequency of claims in some categories, such as musculoskeletal, false positive categorization would occur over time because of related processes occurring at unrelated sites unless the sites' (or processes') codes are grouped in substitutable sets.

These temporal definitions are quite flexible. Transient comorbidities could be defined. For instance, respiratory tract infections in the last 30 days could be specified as a transient comorbidity in a study of asthma. A respiratory infection is a transient risk factor for asthma exacerbations, but one that persists for only a few weeks: farther removed infections are usually irrelevant.

Limitations

The categorization outlined here may improve with the addition of treatment information or inferences based on other diagnostic claims. For instance, both smoking and obesity can be inferred from the appearance of associated diseases. Patients with mental health problems (smoking risk factor) and acute bronchitis (smoking consequence) are likely to smoke. Chronic obstructive pulmonary disease or lung cancer would make a history of smoking nearly certain. Obesity is likely when recent claims include knee osteoarthritis, gastroesophageal reflux, sleep apnea, hypertension and type II diabetes. Many other diseases and risk of death can be inferred from the presence of prescription claims[23, 24]. However, pharmacy claims analysis introduces new challenges. Paradoxical relationships between claims and outcomes have been attributed to "selective under-use of drugs by elderly patients" [25]. Increasingly popular "$4 drug list" prescriptions [26] are not captured in claims data. Anonymity as well as price may motivate use of these deeply discounted pharmacies [27]. Widespread use by insured patients will limit inferences based on prescription claims [28-30].

Another issue is that longitudinal surveillance identifies patterns of claims, which may have distinct implications for outcomes [31]. The use of temporal criteria to identify comorbid conditions risks obscuring this complexity if claims patterns are not considered in analyses.

Conclusions

Temporal constraints applied to ambulatory claims may improve comorbid condition categorization. Nevertheless, incomplete documentation of relevant conditions impedes complete description of primary care patients' multimorbidity.

Table 2. Temporal definitions of the most frequently matched subgroups in each category

Categories	Main subgroup	IP;OP (#:#)	Min-Max (days)	Look Back (days)	Codes (#)	Bundles (#)
Hypertension	Benign	1;2	30-400	730	2	1
Lipid disorder	Hyperlipidemia	0;2	30-400	1095	11	1
Htn, complicated	Other	0;0	0-0	0	3	0
Type II diabetes	Uncomplicated	1;2	30-400	730	2	1
Musculoskeletal	Back	1;2	20-400	400	19	3
Mental health	Depression	1;2	30-180	1095	21	2
Endocrine	Hypothyroid	1;2	30-400	400	8	1
Unexpl. illnesses	Functional Somatic	1;2	30-400	730	6	1
Gastrointestinal	GE reflux disease	1;2	30-400	1095	10	1
Body mass index	Obesity	1;1	0-1500	1500	18	1
Smoking	Tobacco	1;1	0-1825	1825	6	1
Asthma	Reactive airway dis	1;2	30-400	1095	11	1
Sleep problems	Other	0;2	20-400	400	52	3
Anemia	Other	0;0	0-0	0	1	0
Headaches	Migraine	1;2	30-400	1825	42	1
Coronary art dis	Ischemia	1;2	14-270	0	15	1
Chr. Obst. Lung	Other	1;2	30-400	1095	1	1
Rheumatic dis.	Rheumatoid arthritis	1;2	30-730	730	12	1
Rhythm disturb	Atrial fibrillation	1;2	30-400	400	3	1
Type I diabetes	Uncomplicated	1;2	30-180	3650	2	1
Allergy	Contact	1;2	30-400	1095	22	5
Infection	Sinusitis	0;2	30-400	400	7	1
Skin diseases	Various	1;2	30-180	730	21	9
Chr. kidney disease	Stage I-IV	1;2	30-400	730	42	1
Neoplasms	Local	1;2	14-90	1825	402	38
Coagulopathy	Venous thromb	2;2	90-1500	0	13	1
Liver disease	Cirrhosis/Enceph.	1;2	30-400	400	13	4
Bone diseases	Osteoporosis	1;2	14-400	3650	6	1
Sexual dysfunction	Various	1;2	30-400	730	10	1
Genitourinary	Calculi	1;2	14-1095	1095	11	1
Heart valve	Mitral valve	1;2	30-730	1825	8	1
Neurology	Multiple sclerosis	1;2	30-400	730	5	1
Pain syndromes	Neuropathy	0;2	30-180	730	6	1
Pulm emboli/htn	Pulm. embolism	1;2	10-180	1825	3	1
Seizures	Traumatic seizures	1;2	30-400	400	2	1
Heart failure	Hypertensive HF	1;2	30-400	1825	26	1
Periph vasc disease	Atherosclerosis	1;2	30-180	0	25	1
HIV	HIV	1;2	30-400	730	12	1
Alcohol misuse	End organ damage	1;2	30-180	0	20	1
Drug misuse	Dependence	1;2	30-270	0	45	1
Stroke	Thromboembolic	1;2	10-180	0	12	1
Lymphoma	Non-Hodgkins	1;2	30-180	1825	171	1
Reproductive	Various	1;2	30-180	730	38	3
Paralysis	Syndromes	2;2	30-400	1095	31	1
Heart blocks	AV block	1;2	30-400	1825	5	1
Nutrition	Anorexia / Bulimia	1;2	20-180	400	2	2
Dementia	Frontal dementias	1;2	30-180	400	11	3
Cystic Fibrosis	Various	1;2	30-400	1095	9	2
Restrictive pulm dis	Various	1;4	30-400	1095	9	1
Sensory losses	Vision loss	1;2	30-400	730	34	1

Table 3. ICD-9CM codes from claims that did not match a category

Code	Description	Patients (#)	%
V700	Routine general medical examination	236058	5.8
4619	Acute sinusitis, unspecified	111050	2.73
4659	Acute upper respiratory infections of unspecified site	102665	2.52
462	Acute pharyngitis	90305	2.22
4660	Acute bronchitis	86781	2.13
V7231	Routine gynecological examination	80650	1.98
V0481	Vaccination against influenza	79894	1.96
2724	***Other and unspecified hyperlipidemia***	63177	1.55
4011	***Benign essential hypertension***	62957	1.55
78079	***Other malaise and fatigue***	58502	1.44
4779	Allergic rhinitis, cause unspecified	57454	1.41
4019	***Unspecified essential hypertension***	48904	1.2
7862	Cough	45206	1.11
5990	Urinary tract infection, site not specified	43967	1.08
7242	***Lumbago***	41305	1.01
78650	Chest pain, unspecified	37557	0.92
2720	***Pure hypercholesterolemia***	37220	0.91
7295	Pain in limb	37053	0.91
7840	Headache	34046	0.84
V061	Vaccination against diphtheria-tetanus-pertussis	33382	0.82
78900	Abdominal pain, unspecified site	33303	0.82
4610	Acute maxillary sinusitis	32290	0.79
6929	Contact dermatitis and other eczema, unspecified cause	31538	0.77
2449	***Unspecified acquired hypothyroidism***	29010	0.71
53081	***Esophageal reflux***	28176	0.69
71946	***Pain in joint, lower leg***	27279	0.67
311	***Depressive disorder, not elsewhere classified***	26877	0.66
30000	***Anxiety state, unspecified***	26528	0.65
7245	***Backache, unspecified***	24594	0.6
7804	Dizziness and giddiness	21465	0.53
7231	***Cervicalgia***	20879	0.51
7821	Rash and other nonspecific skin eruption	20796	0.51
71941	***Pain in joint, shoulder region***	20786	0.51
49390	***Asthma, unspecified type, unspecified***	20607	0.51
4739	***Unspecified sinusitis (chronic)***	20596	0.51
3829	Unspecified otitis media	20224	0.5

Acknowledgements

This work was funded in part by Washington University Institute of Clinical and Translational Sciences grant UL1 TR000448 from the National Center for Advancing Translational Sciences (NIH) and by grant number R24 HS19455 (PI: V. Fraser) from the Agency for Healthcare Research and Quality (AHRQ), and by the American Board of Family Medicine.

References

1. Feinstein A. The pre-therapeutic classification of co-morbidity in chronic disease. *Journal of chronic diseases.* 1970;23(7):455-468.
2. Kadam UT, Croft PR. Clinical multimorbidity and physical function in older adults: a record and health status linkage study in general practice. *Fam Pract.* Oct 2007;24(5):412-419.
3. Saver BG, Wang CY, Dobie SA, Green PK, Baldwin LM. The central role of comorbidity in predicting ambulatory care sensitive hospitalizations. *Eur J Public Health.* Feb 2014;24(1):66-72.
4. Charlson M, Szatrowski TP, Peterson J, Gold J. Validation of a combined comorbidity index. *Journal of clinical epidemiology.* Nov 1994;47(11):1245-1251.
5. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. *Medical care.* Jan 1998;36(1):8-27.

6. Klabunde CN, Potosky AL, Legler JM, Warren JL. Development of a comorbidity index using physician claims data. *Journal of clinical epidemiology.* Dec 2000;53(12):1258-1267.
7. Wang CY, Baldwin LM, Saver BG, et al. The contribution of longitudinal comorbidity measurements to survival analysis. *Medical care.* Jul 2009;47(7):813-821.
8. Huntley AL, Johnson R, Purdy S, Valderas JM, Salisbury C. Measures of multimorbidity and morbidity burden for use in primary care and community settings: a systematic review and guide. *Ann Fam Med.* Mar-Apr 2012;10(2):134-141.
9. Schafer I, Hansen H, Schon G, et al. The German MultiCare-study: Patterns of multimorbidity in primary health care - protocol of a prospective cohort study. *BMC Health Serv Res.* 2009;9:145.
10. van den Bussche H, Schon G, Kolonko T, et al. Patterns of ambulatory medical care utilization in elderly patients with special reference to chronic diseases and multimorbidity--results from a claims data based observational study in Germany. *BMC Geriatr.* 2011;11:54.
11. Bowden K. Managing up to maximize medicare reimbursement for outpatient care. *Am J Health Syst Pharm.* Oct 1 2001;58 Suppl 1:S14-16.
12. Kesselheim AS, Brennan TA. Overbilling vs. downcoding--the battle between physicians and insurers. *The New England journal of medicine.* Mar 3 2005;352(9):855-857.
13. Holmberg S, Rothstein B. Dying of corruption. *Health Econ Policy Law.* Oct 2011;6(4):529-547.
14. Thammasitboon S, Singhal G. Diagnosing diagnostic error. *Curr Probl Pediatr Adolesc Health Care.* Oct 2013;43(9):227-231.
15. Ely JW, Kaldjian LC, D'Alessandro DM. Diagnostic errors in primary care: lessons learned. *J Am Board Fam Med.* Jan-Feb 2013;25(1):87-97.
16. Singh H, Giardina TD, Forjuoh SN, et al. Electronic health record-based surveillance of diagnostic errors in primary care. *BMJ Qual Saf.* Feb 2013;21(2):93-100.
17. Baldwin LM, Klabunde CN, Green P, Barlow W, Wright G. In search of the perfect comorbidity measure for use with administrative claims data: does it exist? *Medical care.* Aug 2006;44(8):745-753.
18. Ninomiya T, Perkovic V, Turnbull F, et al. Blood pressure lowering and major cardiovascular events in people with and without chronic kidney disease: meta-analysis of randomised controlled trials. *BMJ.* 2013;347:f5680.
19. Marenzi G, Cabiati A, Assanelli E. Chronic kidney disease in acute coronary syndromes. *World J Nephrol.* Oct 6 2013;1(5):134-145.
20. Cigarette smoking among adults and trends in smoking cessation - United States, 2008. *MMWR Morb Mortal Wkly Rep.* Nov 13 2009;58(44):1227-1232.
21. Lee DJ, Fleming LE, Arheart KL, et al. Smoking rate trends in U.S. occupational groups: the 1987 to 2004 National Health Interview Survey. *J Occup Environ Med.* Jan 2007;49(1):75-81.
22. Hertz RP, Unger AN, McDonald M, Lustik MB, Biddulph-Krentar J. The impact of obesity on work limitations and cardiovascular risk factors in the U.S. workforce. *J Occup Environ Med.* Dec 2004;46(12):1196-1203.
23. Clark DO, Von Korff M, Saunders K, Baluch WM, Simon GE. A chronic disease score with empirically derived weights. *Medical care.* Aug 1995;33(8):783-795.
24. Von Korff M, Wagner EH, Saunders K. A chronic disease score from automated pharmacy data. *Journal of clinical epidemiology.* Feb 1992;45(2):197-203.
25. Glynn RJ, Knight EL, Levin R, Avorn J. Paradoxical relations of drug treatment with mortality in older persons. *Epidemiology (Cambridge, Mass.* Nov 2001;12(6):682-689.
26. Gatwood J, Tungol A, Truong C, Kucukarslan SN, Erickson SR. Prevalence and predictors of utilization of community pharmacy generic drug discount programs. *J Manag Care Pharm.* Jul-Aug 2011;17(6):449-55.
27. Rucker NL. $4 generics: How low, how broad, and why patient engagement is priceless. *J Am Pharm Assoc.* Nov-Dec 2010;50(6):761-763.
28. Czechowski JL, Tjia J, Triller DM. Deeply discounted medications: Implications of generic prescription drug wars. *J Am Pharm Assoc.* Nov-Dec 2010;50(6):752-757.
29. Tungol A, Starner CI, Gunderson BW, Schafer JA, Qiu Y, Gleason PP. Generic drug discount programs: are prescriptions being submitted for pharmacy benefit adjudication? *J Manag Care Pharm.* Nov-Dec 2012;18(9):690-700.
30. Omojasola A, Hernandez M, Sansgiry S, Paxton R, Jones L. Predictors of $4 generic prescription drug discount programs use in the low-income population. *Res Social Adm Pharm.* Jan-Feb 2014;10(1):141-148.
31. Freund T, Kunz CU, Ose D, Szecsenyi J, Peters-Klimm F. Patterns of multimorbidity in primary care patients at high risk of future hospitalization. *Popul Health Manag.* Apr 2012;15(2):119-124.

Designing a Clinical Dashboard to Fill Information Gaps in the Emergency Department

Jordan L. Swartz, MD, MA[1]; James J. Cimino, MD[2]; Matthew R. Fred, MD[3]; Robert A. Green, MD, MPH, MA[4]; David K. Vawdrey, PhD[1,3]

[1]Department of Biomedical Informatics, Columbia University
[2]Laboratory for Informatics Development, National Institutes of Health
[3]Information Systems, NewYork-Presbyterian Hospital
[4]Department of Medicine, Columbia University
New York, NY

Abstract

Data fragmentation within electronic health records causes gaps in the information readily available to clinicians. We investigated the information needs of emergency medicine clinicians in order to design an electronic dashboard to fill information gaps in the emergency department. An online survey was distributed to all emergency medicine physicians at a large, urban academic medical center. The survey response rate was 48% (52/109). The clinical information items reported to be most helpful while caring for patients in the emergency department were vital signs, electrocardiogram (ECG) reports, previous discharge summaries, and previous lab results. Brief structured interviews were also conducted with 18 clinicians during their shifts in the emergency department. From the interviews, three themes emerged: 1) difficulty accessing vital signs, 2) difficulty accessing point-of-care tests, and 3) difficulty comparing the current ECG with the previous ECG. An emergency medicine clinical dashboard was developed to address these difficulties.

Introduction

Fragmentation of patient information is a common problem in healthcare. Health information is seldom shared among competing healthcare delivery organizations, and even within a single organization, it is common for data to be isolated within various information 'silos.'[1] Adoption of electronic health records (EHRs) may help address some aspects of information fragmentation, but EHR systems themselves are fragmented, making it difficult for clinicians to easily review data that "belong together." For example, in our commercial EHR, laboratory results and medication orders are accessed through different modules that are not visible on the computer screen at the same time, even though reviewing these two types of data together makes clinical sense. This fragmented model for displaying patient information requires increased cognitive effort to obtain a holistic understanding of a patient. In turn, increased cognitive effort can lead to a higher rate of medical error.[2]

Related to the challenge of information fragmentation is the increasing awareness of gaps in clinicians' information needs. Stiell and colleagues found that physicians reported information gaps in one-third of patients presenting to the emergency department.[3] Of these information gaps, half were felt to be either very important or essential to patient care, they were found more commonly in sicker patients, and they were independently associated with a prolonged length of stay in the emergency department. Historical information (e.g. previous visits, past medical history) was the most common information gap among the studied emergency physicians.

Even when historical information is available to emergency physicians, they do not always access it,[4] especially if it is difficult or time-consuming to find.[5] Effective cognitive support—providing information in an optimal format to clinicians when and where it is needed—should be a foundational principle for designing any clinical information system. The importance of cognitive support was highlighted in the 2009 report of the U.S. National Research Council, "Computational Technology for Effective Health Care: Immediate Steps and Strategic Directions."[6] The report noted that current healthcare information systems force clinicians to "devote precious cognitive resources to the details of data," explaining that "without an underlying representation of a conceptual model for the patient showing how data fit together and which are important,…understanding of the patient can be lost."[6]

Dashboards have been defined as "a visual display of the most important information needed to achieve one or more objectives that has been consolidated on a single computer screen so it can be monitored at a glance."[7] Dashboards have been employed in a wide range of medical settings, including emergency medicine, otolaryngology, nursing care, and maternity care.[8-11] While clinical dashboards are typically used to summarize the status of a cohort of patients (such as an emergency department tracking board), we believe there is a pressing need for dashboards to

organize and efficiently display data for individual patients. We hypothesized that emergency medicine clinicians have specific information needs, that these needs are not adequately addressed by our existing electronic health record, and that a clinical dashboard could be created to fill the information gaps.

Methods

This IRB-approved study was performed at the Columbia University Medical Center emergency department, which has 126,000 annual visits and serves an urban, low-income population. A survey was used to elicit the general information needs of emergency medicine clinicians. Structured interviews were conducted to supplement the survey data and to identify situation-specific information gaps.

Survey

After a review of the survey methodology literature, a preliminary survey was developed. This online instrument adhered to the tenets of proper survey design, including limiting forced responses and allowing for free-text answer choices.[12] The survey questions were designed to clarify the clinical information items most important to emergency medicine clinicians. Questions were answered via a slider that could be moved continuously from 0 to a maximum score of 100; the default location of the sliders was set to 50.

The preliminary survey was iteratively refined based on feedback from an advisory committee composed of emergency medicine physicians and experts in qualitative research methods. The result of this process was a six-question survey instrument that was emailed to all emergency department attending physicians, fellows, and resident physicians using the Qualtrics Survey software (Qualtrics LLC, Provo, UT). A subsequent reminder email was sent approximately two weeks after the initial survey email.

The survey asked clinicians to consider their information needs while caring for a typical patient in the emergency department. The survey also inquired about which patient data items should be included in a clinical dashboard. To orient the survey respondents to the definition of a "dashboard" and the layout of a potential emergency medicine dashboard at our institution, a sample image of an ambulatory medicine dashboard was included in the survey. To avoid biasing respondents, the sample dashboard included information such as preventive care recommendations that were not especially relevant to the emergency environment.

Interviews

Structured interviews were conducted with emergency department faculty, residents, and midlevel practitioners while they were working in the emergency department. The interviewer (JS) followed an interview script that focused the conversation on information gaps that were present in the emergency department, specifically with respect to the hospital's information systems. Clinicians were asked about their "pain points" during an average shift and how the EHR could be modified to reduce frustration and improve efficiency. The interviews were coded using the grounded theory method, an inductive approach in which interview and observation data are coded and then organized into themes.[13] Data collection and analysis were performed concurrently; the interviews were conducted until theme saturation was achieved.

Dashboard Development

Based on the results of the survey and interviews, a dashboard display was designed and created. Our institution uses a commercial EHR product, Allscripts Sunrise (Allscripts Corp., Chicago, IL), in the emergency and acute care settings. A locally developed system called iNYP integrates with the EHR and provides advanced data review capabilities. iNYP is a Java-based service-oriented web application that builds on Columbia University's 25-year history of clinical information system innovation.[14-16] iNYP is available as a custom tab within the commercial EHR (supplementing the native results review capabilities) and can also be accessed from a web browser or a mobile device. At the time of the study, approximately 8,000 clinicians used iNYP alongside the commercial EHR each month. During the study, a new architecture based on HTML5 and JavaScript was added to iNYP to enable the creation of clinical dashboards. The dashboard architecture facilitated the display of data originating from disparate EHRs and from different locations within the same EHR. A comprehensive evaluation of the clinical use of the dashboard was outside the scope of the current study.

Results

Survey

Of 109 emergency department attending physicians, fellows, and resident physicians, 52 (48%) completed the online survey. The majority of respondents (62%) were attending physicians.

The clinical information items reported to be most helpful while caring for patients in the emergency department were vital signs, electrocardiogram (ECG) reports, previous discharge summaries, and previous laboratory test results, as shown in Table 1.

Table 1. Helpfulness while seeing average patient in ED

Survey Item	Average Score
Vital Signs	94.35
Previous ECG	92.04
Prior discharge summary	89.59
Prior lab results	84.73
Prior ED note	84.16
Something else*	74.06
Triage note	69.75
Guidelines	60.85
Immunizations	26.93

*Something else

Medication list - 4 respondents
Imaging results - 3 respondents
PMD phone # - 3 respondents
Ambulance note - 2 respondents
Allergies - 2 respondents
ED visits - 2 respondents
Clinic notes - 1 respondent

Table 2. Helpfulness of *historical* data in dashboard

Survey Item	Average Score
Something else*	92.2
Previous ECG	91.8
Past medical history	89.8
Prior discharge summary	87
Prior lab results	85
Prior ED note	79.3
Prior imaging	78.6
Immunizations	32

*Something else:

Medication list – 3 respondents
Vital signs – 2 respondents
Lab results – 1 respondents
Microbiology results – 1 responden

Table 3. Helpfulness of *active* data in dashboard

Survey Item	Average Score
Vital signs	94.35
Lab results	92.04
Current imaging results	89.59
Something else*	84.73
PMD phone #	84.16
Triage note	74.06
Reference Material	69.75

*Something else:

Don't show labs – 1 respondent
Meds given in ED – 1 respondent
Triage level – 1 respondent
Pending results – 1 respondent
Medication list – 1 respondent
Ambulance note – 1 respondent

The list of a patient's medications was the most frequently requested "Something else?" free-response item. Survey respondents identified triage notes, evidence based-guidelines, and immunization histories as less important.

After showing respondents an image of a sample ambulatory medicine dashboard, they were asked to identify the clinical information items that they thought would be most helpful for an emergency medicine dashboard. In terms of historical clinical information items, as shown in Table 2, the results suggested that previous ECG, past medical history, most recent inpatient discharge summary, and prior lab results would be most helpful.

For active clinical information items (i.e., from the current visit), as shown in Table 3, respondents reported that it would be most helpful to have vital signs, lab results, and imaging results in the dashboard. In contrast, information such as the private medical doctor's contact information, the triage note, and reference material were felt to be less helpful. In the "Something else" free-response component of the survey, one respondent requested that lab results not be shown in the dashboard because they could already be found elsewhere.

The final question allowed respondents to answer in a free-response manner the single feature they would most want to add to the existing EHR. The majority of answers were similar to those found in the other parts of the survey (e.g. medication list, previous results). However, there were two novel answers: 1) patient photograph and 2) a way for outside physicians referring their patients to the emergency department to have a means of communicating referral information directly into the EHR.

Interviews

Of the 18 interviews conducted, 8 were with attending physicians, 7 were with resident physicians, and 3 were with midlevel providers. The interviews lasted approximately 10-20 minutes. There were three key themes that emerged: 1) difficulty accessing current vital signs, 2) difficulty accessing current point-of-care tests, and 3) difficulty comparing the current ECG with the previous ECG.

The interviewees noted that although vital signs were very important to the emergency medicine clinician, their display in the EHR was both tedious to access and in a format that was time-consuming to comprehend. A similar sentiment was expressed regarding the results of the point-of-care tests, such as urine pregnancy. Because these results are hand-entered by nurses, they are found in a section separate from other laboratory test results. Clinicians complained that these results were "buried at the bottom of a long flowsheet," which was cumbersome and time-consuming to access.

Despite the fact that all ECGs were accessible in the EHR, it was not possible to compare two ECGs on the computer screen at the same time. Instead, clinicians were obliged to print one of the ECGs and compare it to the other on the screen, or more often, to print both of the ECGs and compare them side-by-side.

Dashboard Development

Based on the knowledge obtained from the survey and interviews, we designed a dashboard to address the information gaps of the emergency medicine clinicians. The dashboard provided summary "tiles" of five of the most-requested clinical information items: vital signs, point-of-care tests, ECGs, previous notes/discharge summaries, and medications. In designing the tiles, we considered the best practices of information visualization, such as adhering to location-based emphasis (i.e., placing the most important information in the upper left) and minimizing the amount of non-data pixel use.[17]

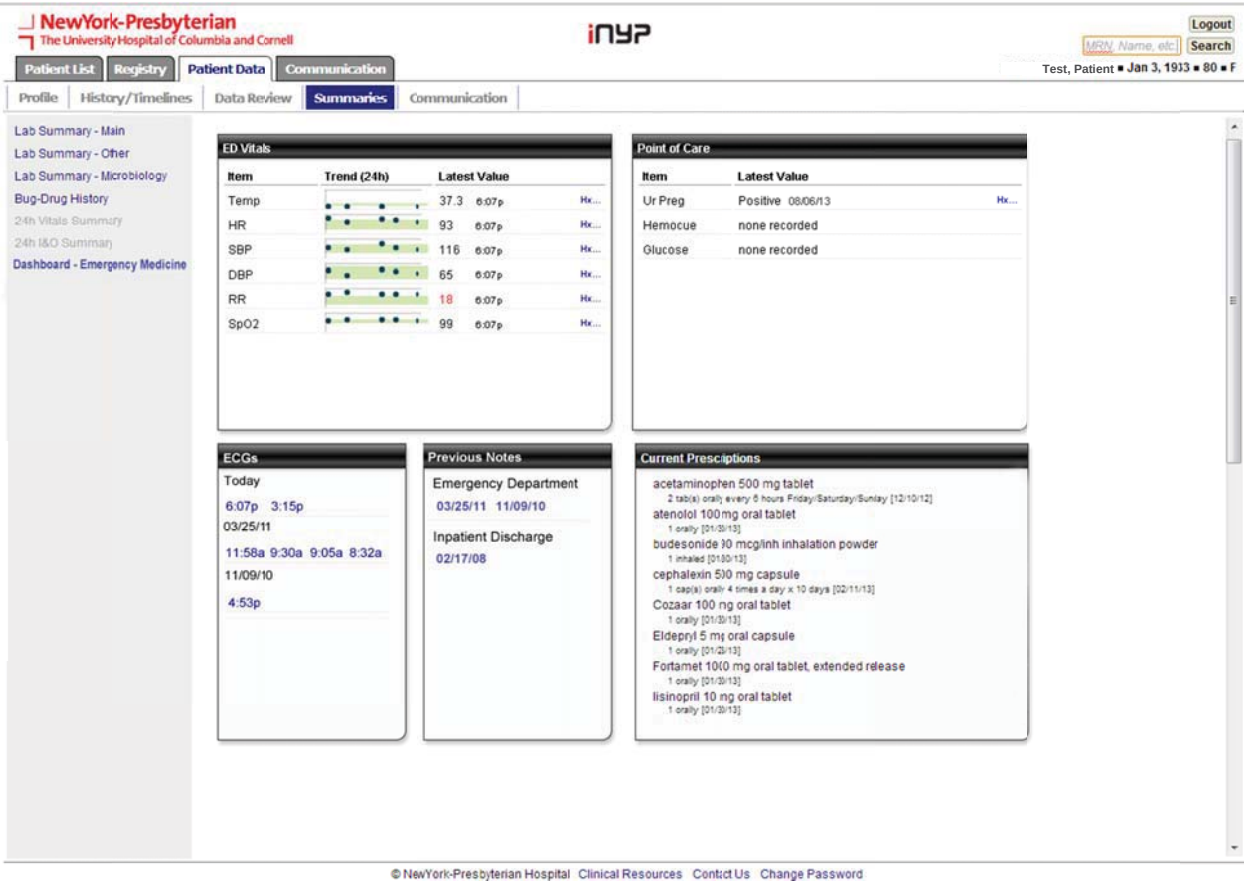


Figure 1. Screenshot of the Emergency Medicine Dashboard

The dashboard display is shown in Figure 1. The "ED Vitals" tile presents the patient's vital signs in graphical and tabular formats. The sparkline graphic provides a quick way to identify trends during a patient's stay in the emergency department. The shaded band indicates the normal range; data points that are abnormal (and therefore fall outside the band) are readily apparent. Alternatively, a more precise display of the same data can be found by reviewing the "Latest Value" column and clicking the history ("Hx") link, which pops up a display of the previous vital sign measurements. Abnormal values are marked in red font.

The "Point-of-Care" tile shows the results of the point-of-care tests performed at our institution: urine pregnancy, urine dipstick, finger-stick glucose, and finger-stick hemoglobin. If no results are available, "none recorded" is displayed. Similar to the ED Vitals tile, the history ("Hx") link displays any earlier point-of-care test result. This display is useful when clinicians need to follow a patient's finger-stick hemoglobin or glucose test during their stay in the emergency department.

The "ECG" tile divides electrocardiogram results by date. Clicking one of the hyperlinks launches a pop-up ECG viewer, which graphically displays the tracing in PDF format, along with the cardiologist's report (if available). Multiple ECGs can be viewed simultaneously by clicking on additional hyperlinks.

Similarly, the "Previous Notes" tile organizes discharge summaries and previous emergency department visit notes by date. Clicking the associated hyperlink launches a note viewer. Finally, the "Current Prescriptions" tile shows the name, dosage, and dosing schedule for the patient's home medications.

Discussion

In order to design clinical information systems that minimize data fragmentation, it is necessary to know which clinical information items are most important to the users of the system. Information needs of clinicians do not follow a "one-size-fits-all" model: a given specialty cohort of clinicians may use information systems in a much different way than their colleagues who work in a different setting or come from another specialty. Once information needs of a specific group are known, extra care can be given to the design of clinical information systems to ensure that various workflows are supported and data are presented in a way that makes "clinical sense."

Our qualitative investigation of the information needs of emergency medicine clinicians demonstrated both the information needs of the average emergency medicine clinician as well as institution-specific information gaps in our EHR system. We found that vital signs, current and previous ECG, previous discharge summary, lab results, and medication list were particularly important; this finding is probably applicable to most emergency care settings. Our investigation also showed that at our institution, there were information gaps related to the inefficient display of these clinical information items.

Our institution is not unique; many (if not all) clinical information systems have been designed with inadequate usability testing and apparent lack of clinical input.[18] Our study provides a methodology by which the information needs of a specialty-specific group of clinicians can be assessed. In turn, this can inform the development of specialty-specific dashboards that fill information gaps.

The clinical dashboard we created was designed based on the feedback elicited from clinicians who used the information systems regularly. The information gaps that were identified were related to vital signs, lab results, ECGs, and discharge summaries, all of which were felt to be among the most important clinical information items by the clinicians. In order to fill these gaps, individual tiles were created within the clinical dashboard to enable at-a-glance monitoring and access to these items.

In the future, we envision "smart dashboards," which dynamically change based on the chief complaint of the patient. For example, a patient who presents to the emergency department with a chief complaint of "laceration" would have his tetanus status displayed, whereas different information might be surfaced for someone presenting with chest pain. This in turn might inform the development of universal rules of clinical data display. For example, a patient's creatinine level and pregnancy status (if appropriate) should always be shown when ordering a computed tomography (CT) scan. Similarly, a previous ECG, when available, should always be shown next to the current one. We believe that improved information displays will better support the cognitive tasks of clinicians.

Our study has several limitations. First, the survey response rate was 48%, reflecting a possible bias because of differences in the types of subjects who completed the survey versus those who did not. In any case, this study at least reflects almost half of the relevant clinicians. Second, a single investigator conducted the structured interviews. While an interview script was used, it is possible that the way in which questions were asked biased the respondents in their answers. Third, because our study relied on survey and interview, it is subject to recall bias. Observation

could be used in future studies to confirm the interview and survey results. Fourth, we have not yet implemented the dashboard within the emergency department. Fifth, the study was performed at only one emergency department with only one electronic health record. Our findings may not generalize to other environments or EHR systems. Nevertheless, we believe our results can help inform EHR vendors about the information needs of their users and also encourage the investigation of clinical information needs in diverse settings.

Conclusion

Electronic health records suffer from data fragmentation, which adversely affects patient care. Our study presented a methodology by which the information needs of a specialty cohort of clinicians can be studied. We demonstrated how a better understanding of clinicians' information needs can inform the development of specialty-specific clinical dashboards that provide cognitive support and improve efficiency.

Acknowledgements

Dr. Cimino was supported in part by research funds from the National Library of Medicine and the NIH Clinical Center.

References

1. Mohler MJ. Collaboration across clinical silos. Frontiers of health services management. 2013;29(4):36-44.
2. Horsky J, Allen MB, Wilcox AR, Pollard SE, Neri P, Pallin DJ, et al. Analysis of user behavior in accessing electronic medical record systems in emergency departments. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2010;2010:311-5.
3. Stiell A, Forster AJ, Stiell IG, van Walraven C. Prevalence of information gaps in the emergency department and the effect on patient outcomes. CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne. 2003;169(10):1023-8.
4. Shapiro JS, Kuperman G, Kushniruk AW, Kannry J. Survey of emergency physicians to determine requirements for a regional health information exchange network. AMIA Spring Congress. 2006 May 16-18.
5. Hripcsak G, Sengupta S, Wilcox A, Green RA. Emergency department access to a longitudinal medical record. Journal of the American Medical Informatics Association : JAMIA. 2007;14(2):235-8.
6. In: Stead WW, Lin HS, editors. Computational Technology for Effective Health Care: Immediate Steps and Strategic Directions. The National Academies Collection: Reports funded by National Institutes of Health. Washington (DC)2009.
7. Few S. Information dashboard design : the effective visual communication of data. 1st ed. Beijing ; Cambride MA: O'Reilly; 2006. viii, 211 p. p.
8. Khemani S, Patel P, Singh A, Kalan A, Cumberworth V. Clinical dashboards in otolaryngology. Clinical otolaryngology : official journal of ENT-UK ; official journal of Netherlands Society for Oto-Rhino-Laryngology & Cervico-Facial Surgery. 2010;35(3):251-3.
9. Stone-Griffith S, Englebright JD, Cheung D, Korwek KM, Perlin JB. Data-driven process and operational improvement in the emergency department: the ED Dashboard and Reporting Application. Journal of healthcare management / American College of Healthcare Executives. 2012;57(3):167-80; discussion 80-1.
10. Simms RA, Ping H, Yelland A, Beringer AJ, Fox R, Draycott TJ. Development of maternity dashboards across a UK health region; current practice, continuing problems. European journal of obstetrics, gynecology, and reproductive biology. 2013.
11. Tan YM, Hii J, Chan K, Sardual R, Mah B. An electronic dashboard to improve nursing care. Studies in health technology and informatics. 2013;192:190-4.
12. Schleyer TK, Forrest JL. Methods for the design and administration of web-based surveys. Journal of the American Medical Informatics Association : JAMIA. 2000;7(4):416-25.
13. B K, J M. Qualitative research methods for evaluating computer information systems. Evaluating the Organizational Impact of Healthcare Information Systems. 2005:30-55.
14. Hripcsak G, Cimino JJ, Sengupta S. WebCIS: large scale deployment of a Web-based clinical information system. Proceedings / AMIA Annual Symposium AMIA Symposium. 1999:804-8.
15. Hendrickson G, Anderson RK, Clayton PD, Cimino J, Hripcsak GM, Johnson SB, et al. The integrated academic information management system at Columbia-Presbyterian Medical Center. MD computing : computers in medical practice. 1992;9(1):35-42.

16. Johnson S, Friedman C, Cimino JJ, Clark T, Hripcsak G, Clayton PD. Conceptual data model for a central patient database. Proceedings / the Annual Symposium on Computer Application [sic] in Medical Care Symposium on Computer Applications in Medical Care. 1991:381-5.
17. Tufte ER. The visual display of quantitative information. 2nd ed. Cheshire, Conn.: Graphics Press; 2001. 197 p.
18. Middleton B, Bloomrosen M, Dente MA, Hashmat B, Koppel R, Overhage JM, et al. Enhancing patient safety and quality of care by improving the usability of electronic health record systems: recommendations from AMIA. Journal of the American Medical Informatics Association : JAMIA. 2013;20(e1):e2-8.

SOEMPI: A Secure Open Enterprise Master Patient Index Software Toolkit for Private Record Linkage

Csaba Toth[1], MS, Elizabeth Durham[1], PhD, Murat Kantarcioglu[2], PhD, Yuan Xue[3], PhD, Bradley Malin[1,3], PhD
[1]Dept. of Biomedical Informatics, Vanderbilt University; [2] Dept. of Computer Science, University of Texas at Dallas, Richardson, TX; [3]Dept. of Electrical Engineering & Computer Science, Vanderbilt University, Nashville, TN

Abstract

To mitigate bias in multi-institutional research studies, healthcare organizations need to integrate patient records. However, this process must be accomplished without disclosing the identities of the corresponding patients. Various private record linkage (PRL) techniques have been proposed, but there is a lack of translation into practice because no software suite supports the entire PRL lifecycle. This paper addresses this issue with the introduction of the Secure Open Enterprise Master Patient Index (SOEMPI). We show how SOEMPI covers the PRL lifecycle, illustrate the implementation of several PRL protocols, and provide a runtime analysis for the integration of two datasets consisting of 10,000 records. While the PRL process is slower than a non-secure setting, our analysis shows the majority of processes in a PRL protocol require several seconds or less and that SOEMPI completes the process in approximately two minutes, which is a practical amount of time for integration.

Introduction

Healthcare organizations (HCOs) are encouraged, or are required, to share data to support various endeavors, such as post-market surveillance and biomedical research. To mitigate bias in investigations, it is important to resolve when a patient's data resides in multiple resources. This process, called record linkage, is non-trivial because a patient's record often contains typographical and semantic errors.[1] Sophisticated record linkage strategies have been proposed to resolve these problems[2], but their application is hampered by policies or regulations, such as the HIPAA Privacy Rule[3], which limit the sharing of identifiers that facilitate the linkage process (e.g., personal name and Social Security number). To overcome this barrier, a growing list of techniques has been proposed to support private record linkage (PRL).[4]

From a high level, the PRL process has a lifecycle that entails (but is not necessarily limited to) the following steps:

1. Generation and storage of keys for cryptosystems, or salt values for hash functions, invoked in a PRL protocol;
2. Communication of keys and salt to the entities encoding the records upon request;
3. Transformation of identifiers into their protected form as specified by the protocol;
4. Execution of the record linkage framework (e.g., feature weighting, blocking, and comparison of record pairs to predict which correspond to the same individual); and
5. Transfer of records and parameters related to the linkage protocol (i.e., all communication between parties).

Unfortunately, there has been a lack of transition of sophisticated PRL techniques into practice. We suspect this is due, in part, to factors associated with: i) complexity in the description and design of protocols, ii) availability of working software, and iii) coverage of the data lifecycle.

Regarding the first factor, a wide variety of methods exist to support PRL. For instance, let us focus on one aspect of step four of the lifecycle for a moment. There are many techniques that have been proposed for computing the similarity of strings in a privacy preserving manner. Some of these are based on secure multiparty computation (SMC) protocols (e.g., [5]). Others are based on less cryptographically intense approaches, such as mapping the identifiers to Bloom filter encodings (BFEs)[6-8]. Hybrid strategies selectively reveal information (e.g., patient's age as a 5-year range) to speed up the SMC process[9]. These methods vary considerably in their effectiveness (e.g., precision and recall), efficiency (e.g., computational runtime, memory required, and bandwidth), and security[11].

Second, even when PRL techniques appear to balance these factors, they may not be adopted because it is argued that they require more engineering than a simple protocol[12]. In many respects, this is a valid argument because the majority of published PRL methods are often limited to software implementations developed for experimental

analysis. HCOs that wish to use such strategies would be hindered from doing so without significant additional engineering. This is problematic because most HCOs have neither the time nor resources to interpret and implement cryptographic procedures into an easy to use software product. To the best of our knowledge, there are only two tools[13,14] (as reviewed below) which explicitly provide readily available software that supports PRL methodologies.

Third, to the best of our knowledge, all published PRL methods, as well as the software disseminated in support, tend to focus on the linkage stage of lifecycle only. In other words, they assume that HCOs have the ability to manage patients' records, establish communication links, and execute linkage procedures. Consider, a certain subset of PRL methods require HCOs to work with a third party to more appropriately balance efficiency and security.[4] Yet, this assumption does not hold true and, thus, the PRL methods in the literature, as well as the aforementioned software tools, fail to cover the lifecycle. This, again, requires HCOs (or research data managers) to piece together or augment certain software tools.

Given the aforementioned challenges, we developed an open source software toolkit to support the PRL lifecycle. Thus, there are several contributions of this paper:

1. **Open source software framework for PRL:** We extend an existing open source master patient indexing tool[15] to handle the record linkage process in a manner consistent with accepted frameworks. The resulting software, called the Secure Open Enterprise Master Patient Index (SOEMPI), entails an innovative architecture to tailor the specification of PRL protocols to an HCO's needs. In doing so, SOEMPI manages the communication between disparate data providers, as well as the third parties, involved in the PRL process. Furthermore, the architecture of SOEMPI is implemented in a component-based manner and is readily extensible to PRL approaches that follow the aforementioned lifecycle.
2. **Case Studies in PRL Protocol Implementation:** We implement several PRL protocols to illustrate the capability of SOEMPI. All protocols are based on a Bloom filter-based record transformation method proposed in the literature. In these protocols, HCOs hash the patient identifiers in their respective medical record systems into Bloom filters encodings (BFEs). Based on a recently published protocol, each variable is encoded in its own filter, which are subsequently combined into a single composite filter, with bit positions weighted to optimize linkage performance.[16] These encodings are subsequently sent to a third party for linkage.
3. **Runtime Analysis:** To demonstrate the feasibility of SOEMPI, we perform an analysis over the PRL lifecycle. We compare the running time between SOEMPI and OpenEMPI and show the communication and transformation required for PRL is relatively short, such that it can be completed in a practical amount of time. Given that the accuracy of such strategies has been addressed[6], this paper focuses on an evaluation of runtime.

PRL Participants and Protocols

This section begins with a review of record linkage and PRL. We then proceed into describing common participants of the linkage schemes and finally describe two actual protocols.

Record linkage is a data management process. At its core, it consists of two or more data managing HCOs who wish to integrate disparate collections of records. The process of record linkage requires the transfer of data from one HCO to another, who performs a data comparison and integration procedure. This procedure may be deterministic and rule-driven[17] or probabilistic and supported by robust statistical inference methods (e.g., the expectation-maximization implementation of the Fellegi-Sunter (FS) algorithm[18]). In PRL, the set of entities may be expanded to include one or more third parties (i.e., non-contributors of data). As we explain in further depth below, these parties can take on a variety of responsibilities, ranging from the generation and communication of cryptographic keys for the HCOs to the execution of the integration process on behalf of the HCOs in a manner that prevents the inference of patients' identifiers. Thus, before proceeding into the details of existing systems, we take a moment to review the classes of participating entities described in various PRL systems and their responsibilities.

Data Providers (DPs) are the HCOs who manage the identified patient records that will be linked. Without loss of generality, Let us focus on two disparate patient data holder HCOs, Alice and Bob. They engage in record linkage sessions with other HCOs through the third parties.

Key Server (KS) provides "salt" and/or cryptographic key values needed by the DPs, so that they can perform data transformations (e.g., generation of BFEs) of their data fields in a consistent manner.

Data INtegrator (DAN) performs the integration component of the PRL process for the HCOs. DAN may accept encoded values (e.g., BFEs) from the DPs and performs record integration in a privacy preserving manner.

PArameter Manager (PAM) receives sample data from the DPs to determine the parameters of the linkage protocol (e.g., number of bits in a Bloom filter). It coordinates the communication of such parameters to the DPs.

A. Three-Party PRL Protocol

As an example, let us consider a PRL protocol with one third party as shown in the blue-shaded section of Figure 1. In this protocol, DAN requires Alice and Bob to encode their patients' identifiers using a consistent set of salted (i.e., keyed) HMAC (Hash-based Message Authentication Code) functions. As such, it is often asked, "Where do these keys come from?" In the first variation of the protocol illustrated, we integrate an independent, semi-trusted authority in the form of key server KS to generate the salts for Alice and Bob. In Figure 1, the sections shaded in blue depict a sequence diagram that summarizes the series of steps and service calls. After receiving the salts, Alice and Bob encode their records into BFEs based on a salted hash function. The resulting encodings are then submitted to DAN, who waits until both datasets and match request tickets (explained below) have been received. Upon reception, DAN runs the requested matching procedure.

B. Four-Party PRL Protocol

Recent research has shown that if the BFEs submitted to DAN are not tuned properly, they are vulnerable to cryptanalysis and leakage (e.g., mapping of an encoded value to a patient's real name).[7] To mitigate such an attack, Alice and Bob should parameterize their BFE strategy in a manner that minimizes cryptanalysis, but maximizes accuracy.[16] Due to privacy concerns, Alice and Bob cannot exchange patient identifiers directly, but may wish to employ the assistance of an additional third party to provide feedback on how best to setup the system (e.g., determine weights for each field, such as forename, surname, and Social Security number). This is where a parameter manager, PAM, can be of assistance.

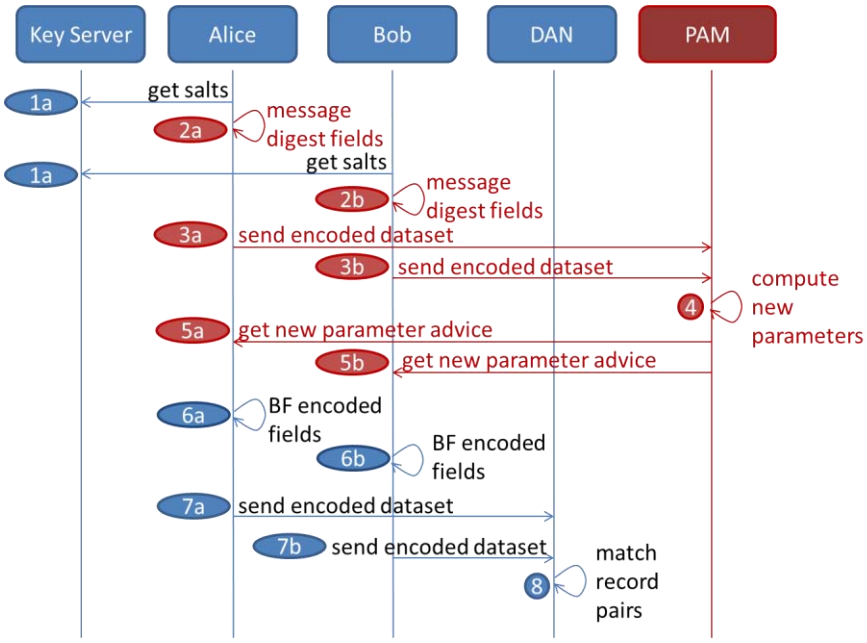


Figure 1. An illustration of the flow for a private record linkage protocol that incorporates multiple third parties. The blue sections of the flow correspond to a third-party protocol, while the incorporation of the red-shaded sections convert it into a four-party protocol. In this diagram, the variables *a* and *b* are used to denote the steps associated with Alice's and Bob's datasets, respectively).

In Figure 1, the sections shaded in red provide additional support for such a process. In this case, Alice and Bob request and receive salts from KS as before. However, instead of sending data to DAN, they provide a small sample of their HMAC encoded records to PAM. This entity compares the datasets (as described elsewhere[16]) and responds to Alice and Bob with the appropriate set of parameter values (e.g., size of Bloom filter, number of hash functions, size of the *n*-grams into which patient identifiers should be split, and random order of bits into which the hash functions will be mapped). The protocol then continues in the same manner as described in the third party protocol.

Background: the State of PRL Software

This section provides a systematic analysis of existing free software solutions for master patient indexes (MPI), record linkage, and PRL. Table 1 summarizes various aspects of software tools for record linkage and PRL that are readily available. The majority of the software suites in the table are limited in that they were developed to facilitate the integration of specific datasets in a non-coordinated manner. Moreover, they are limited to the user interface of a local machine and, thus, their main deployment is for desktop-level usage. While it is easier to install desktop

software than to deploy it as a server application, it significantly limits its scalability. If such software was installed on a server, for instance, interaction would only be possible through a remote desktop protocol. Because this functions in a stand-alone manner, it is beyond its scope to support any protocol in an automatic fashion that involves multiple participants.

Table 1. A systematic comparison of existing generic and privacy preserving record linkage software tools.

Tool	PRL	Free	Open Source	Extensible	Communication[a]	GUI[k]
Link King[21]	No	No[c]	Yes	Limited[e]	Manual	Desktop[b]
Link Plus[20]	No	Yes	No	No	Manual	Desktop[b]
FEBRL[26]	No[d]	Yes	Yes	Yes	Manual	Desktop[b]
FRIL[19] + *LinkIt*[29]	Yes	Yes	Yes	Yes	Manual	Desktop[b]
MTB[28]	Yes	Yes	No[f]	Yes	Manual	Desktop[b]
OpenEMPI[15]	No	Yes	Yes	Yes[g]	Yes[h]	Web[j]
OpenMRS[23]	No	Yes	Yes	Yes[g]	Yes[h]	Web[j]
RECLINK[22]	No	No[c]	Yes	Limited[e]	Manual	Desktop[b]
SOEMPI	Yes	Yes	Yes	Yes[g]	Yes[i]	Web[j]

[a] Communication with other entities.
[b] Requires graphical desktop sharing solution on server and client side to view the graphical user interface of the server locally.
[c] The script is free in itself, but requires additional SAS or Stata license to run.
[d] Proposed, but not implemented.
[e] The software is not free, and requires specific programmer knowledge.
[f] BloomEncode and SafeLink sourcecode is available only for research projects.
[g] SOA software by nature designed for extensibility, but also require programmer knowledge.
[h] With standard HCO actors, but not in a complex record linkage protocol.
[i] With other SOEMPI instances for record linkage.
[j] Users can view graphical interface of web applications remotely with a browser easily by nature
[k] Graphical User Interface

Most of these software tools enable a data cleaning process and support the tuning of record linkage parameters. They are all capable of performing FS-style probabilistic record linkage and most have rich and detailed user interfaces. The Fine-grained Record Integration and Linkage (FRIL) tool even aids in the execution of several consecutive match runs, where the outcome of each run iteratively helps refine parameters.[19] Link Plus is free, however, it is closed source and only available for Windows.[20] Link King[21] and RECLINK[22] are both based on statistical software suites (SAS and Stata, respectively) which require licenses, but they are free and open source additions. However, this means that their extensibility requires scripting knowledge of the specific statistical package.

A. Open Source Master Patient Indexing

Distributed healthcare systems require systematic approaches for coordination, integration, and management of linked records. One way this has been accomplished is through master patient indexing (MPI) initiatives, which have been integrated into health information exchange systems. Many of the resulting technologies have been implemented as open source software tools. OpenEMPI[15] is an open source MPI project with ongoing development. It is capable of performing deterministic and probabilistic matching and supports MPI. Every aspect of the record linkage process is configurable. It can interface with various health information systems in a standardized manner. The software is based on a service-oriented architecture (SOA) design and a component framework, which makes it extensible. OpenMRS[23] was designed to support the delivery of healthcare in developing countries. It has a matching module that is configurable and capable of performing probabilistic matching. Other open source MPI systems, such as the OpenEMed[24] and OpenHRE[25], have not been supported for years and, thus they were not considered further. It is worth noting that these solutions only address record linkage and patient indexing and do not support PRL.

B. Open Source Private Record Linkage

In comparison to the various options for record linkage tools, the PRL landscape is more barren. The developers of the Freely Extensible Biomedical Record Linkage (FEBRL)[26] tool, for instance, planned to implement PRL methods based on *n*-gram hashes[27], but this has yet to be realized. To the best of our knowledge, there exist only two actual PRL implementations. The first corresponds to the BloomEncode and SafeLink companion tools for the Merge ToolBox (MTB) software.[28] To facilitate PRL, the BloomEncode tool is invoked to transform patient identifiers into BFEs. These are then manually passed to a third party who runs the MTB system and performs record linkage. The other is the LinkIt companion tool of FRIL.[29] We wish to highlight, however, that to realize a fully automated PRL lifecycle, MTB or FRIL/LinkIt would need communication protocols to interact with the companion tools at disparate organizations securely.

Methods

In this section, we describe the PRL lifecycle and how it is supported by the SOEMPI architecture.

A. Requirements

In preparation for this project, we performed a requirements analysis to identify the factors in an existing record linkage software that support the PRL lifecycle. From the outset, it was determined that such a system should have the following properties:

1. Coverage of the entire PRL process;
2. Configurable to support various parameterizations of data schemas, record linkage methods, as well as encoding functions, blocking strategies, and record comparison / matching algorithms;
3. Flexible to facilitate various data schemas at the data providers and the design of different PRL protocols;
4. Enterprise environment capable of enabling the scalable deployment of the software on a server or in a data warehousing environment; and
5. Open source technology to allow for extension and revision by a community;
6. Use of industry-wide accepted technologies for easier communications with possible non-PRL HCO software.

OpenEMPI was selected as a code base because it satisfies most of these properties and is currently maintained.

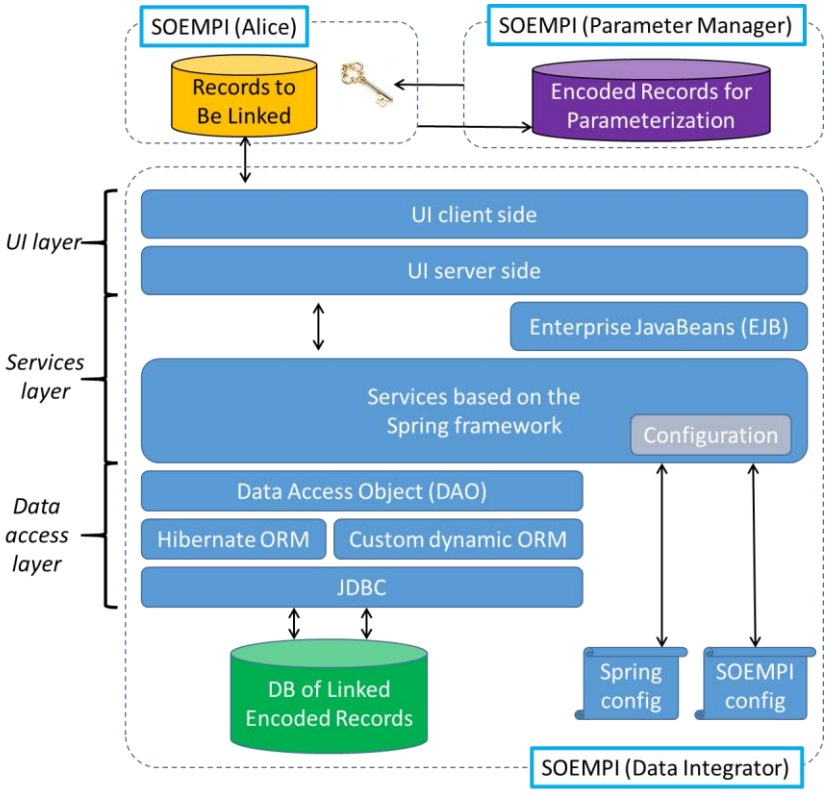


Figure 2. A high-level architecture of the SOEMPI software

B. Software and Communications Architecture

This section begins with a high-level overview of the SOEMPI design, composition, and main building blocks. Next, we provide an overview of the pluggable architecture associated with SOEMPI, which allows for flexible management of the lifecycle. Then, we take a closer look at certain key functionalities in the software design. Finally, we discuss how the various participating entities and data schemas are supported.

1. Architecture of a SOEMPI Instance

Figure 2 provides a high-level view of a SOEMPI instance. This view is based on an *n*-tier design, which facilitates responsible isolation and decoupling of processes. Specifically, the user interface layer (both client- and server-side) at the top interacts with the middle services tier below, which connects to the data access layer, and finally connects to the underlying database. Under the hood, an SOA is applied to achieve a flexible and pluggable architecture.

While many record linkage protocols involve several parties, SOEMPI serves as a universal actor because it can be instantiated according to each of the possible roles described earlier. The actual role played can be selected through either a configuration file or during the login process. According to the selected role, SOEMPI displays only the relevant features on the user interface of the particular role and is restricted to perform only the appropriate operations (e.g., when acting as a Key Server, it can only generate and distribute salts). If additional SOEMPI instances are involved in the protocol (e.g., a third party), the programs communicate with the help of Java EJB remote calls, a method inherited from OpenEMPI.

2. Pluggable Services

While the complete details of SOEMPI are beyond the scope of this paper, we wish to highlight the pluggable aspect of the components in the PRL process. This is one of the key contributions because it enables a modular approach to PRL protocol design and deployment. Figure 3 depicts the steps of the PRL process and how each step offers various encapsulated method options. For instance, SOEMPI offers a variety of record comparison functions, blocking algorithms, and probabilistic matching methods. The composition of a method-chain is possible through specification in an XML configuration file or runtime user interface interaction.

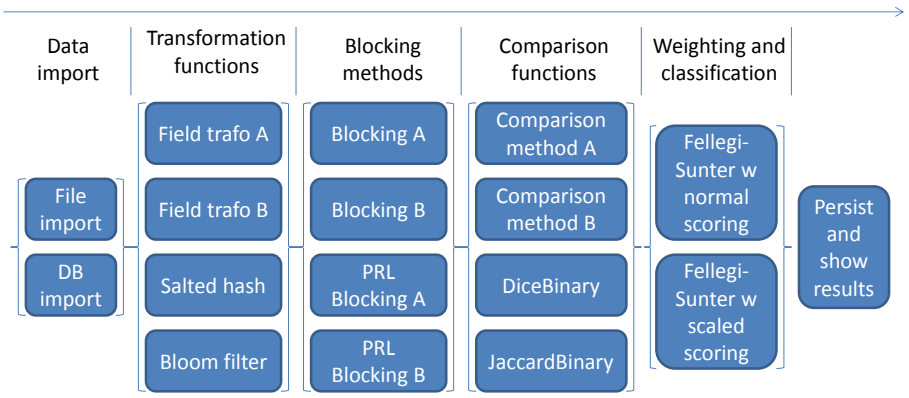


Figure 3. A depiction of the pluggable service architecture for the record encoding and matching functionality of the SOEMPI system.

3. Dynamic Data Schemas

Given that we cannot predict which patient-specific attributes will be utilized for record linkage purposes, SOEMPI should be able to store any type of data and database tables with diverse schemas. We designed SOEMPI to store multiple instances of flexible typed datasets (any number and type of fields). This requirement derives from real-world scenarios where, for example, mother's weight and newborn's weight are required (both are floating point data types). The software was designed such that each imported dataset has its own table in the underlying database, as well as an entry in a registry table which keeps track of the uploaded datasets. Likewise, when SOEMPI links records, it creates a new table for the resulting join, and updates a registry documenting which tables were joined. SOEMPI still leverages a conventional persistence layer (i.e., object relational mapping tools) whenever possible (i.e., user management, sessions, etc.), but uses a custom flexible persistence layer when necessary.

Performance Analysis

In this section, we compare the running time of OpenEMPI to SOEMPI with respect to a certain PRL protocol to illustrate how time and computation are influenced by the privacy preserving procedures.

A. Experimental Design

To allow for reproducibility, we performed our experiments with records from the publicly available North Carolina Voter Registration (NCVR) database, which, at the time of this study, consisted of 6,190,504 records. We generated 10 datasets, each of which consists of 10,000 randomly selected records over the following fields: *Forename*, *Surname*, *City (of Residence)*, *Street Name (of Residence)*, *Gender*, and *Ethnicity*. From each dataset, we created a

corresponding "corrupted" dataset for matching purposes as described by Durham et al.[16] In doing so, every field of each record is subject to a corruption procedure, a part of the SOEMPI toolkit, that extends the strategy implemented in FEBRL[26] (which is based on research by Christen and Pudjijono[31]). This procedure introduced *character-level* errors at rates consistent with those reported in practice. These included optical character recognition (OCR) errors (e.g., S swapped for 8), phonetic errors (e.g., ph swapped for f), and typographic errors (e.g., insertions and transpositions). In our experiments, these corruptions occurred with the following probabilities: insertions and deletions: 0.15, phonetic errors: 0.03, OCR errors: 0.01, substitutions: 0.35, and transpositions: 0.05. To further simulate record linkage challenges faced by HCOs, our extensions introduced errors at the *value-level*, such as the use of a nickname or a change in residential address. First, nicknames based on the Massmind database[32] were substituted for Forenames with probability 0.15. Second, surnames were substituted, using the U.S. Census names data[33], for females (e.g., changes due to marriage) with probability 0.1, for males with probability 0.01, and were hyphenated with probability 0.01. Third, street names were changed with probability 0.1 using a random selection from the entire NCVR dataset.

For OpenEMPI, we considered a record linkage framework that involved three parties: Data Providers Alice and Bob and a Data Integrator DAN. The Data Providers transformed the Forename and Surname fields using double metaphone (a phonetic encoding algorithm designed to mitigate noise) during the import of the datasets. These are appended as additional fields to the datasets, which are then passed onto DAN, where a blocking procedure in the form of the sorted neighborhood algorithm is performed in two rounds over these fields. The matching protocol consisted of an Expectation Maximization-based (EM-based) FS algorithm.

For PRL, we used the Four-Party protocol described above. During the Parameter Manager phase, Alice and Bob transferred 1,000 random samples to PAM in an HMAC-encoded form. PAM then performed an EM-based FS algorithm to determine the parameters of the Bloom filter representations of the records as described elsewhere.[16] After Alice and Bob transferred their BFEs to DAN, blocking is performed through three rounds of a clustering processing based on locality sensitive hashing (LSH), where each seed of a cluster corresponds to 10 bit positions randomly sampled from the Bloom filter schema. Finally, DAN measured the similarity of record pairs (one BFE contributed by Alice and one contributed by Bob) in each cluster (using a Dice similarity function) in the same blocks to link records to their best match.

For each experiment, we matched the clean version of a given dataset to its own corrupted counterpart and reported on the average (and standard deviation) time for the 10 runs. We measured the time required to 1) import the data, 2) exchange data, and 3) execute various aspects of the linkage algorithm. All experiments were run on a single Quad-core Intel i7-2670QM processor @ 2.2GHz with 12 GB system memory. The majority of the computations were measured natively on the operating system. The data transfer measurements were performed between one SOEMPI instance running on the native OS and one running in a virtual box. Figure 1 depicts the specific steps in the process that were measured.

B. Findings

Our results focus on i) the bandwidth required to transfer and manage datasets for record linkage and ii) the time required to complete the record linkage protocols.

Bandwidth Required

In the non-privacy preserving environment, datasets transferred for record linkage required 550 KB. The size of the privacy preserving BFE datasets depend on their length of the Bloom filters, which itself is dependent on records involved in the linkage process. The length of the BFE is determined by PAM during the Parameter Manager phase. Over the 10 runs of our protocol, the average recommended size of the Bloom filter was 9,217 bits, with a standard deviation of 2,197 bits. However, size ranged from a minimum of 7,133 to a maximum of 12,904, such that the size of the datasets transferred for private record linkage to DAN for integration ranged from approximately 8.9 MB to 16.1 MB. It should also be noted that the runtime of the PRL matching process is greatly influenced by the number of bits used in the LSH-based blocking process, as well as the number of rounds of blocking. Since the goal of this paper is to report on how the communication protocol influences the time necessary to perform PRL, we note that such a sensitivity analysis is beyond the scope of this paper and refer the reader elsewhere for a discussion on Bloom Filter record linkage[8] and LSH blocking accuracy[34].

Time to Complete Protocol

Before comparing the record linkage protocols, it is important to note that for each step in SOEMPI in which a participant initiates communication with another participant the first time, a one-time authentication procedure must take place. This authentication necessary to establish a secure connection and incurs a fixed cost of 8-9 seconds.

Turning our attention to a comparison of the protocols, Table 2 provides a summary of the results and a breakdown by process. Here, there are several notable findings to highlight. First, the Four-Party protocol incurs three additional categories of cost in comparison to the conventional record linkage method. The first corresponds to the request and dissemination of salt values from the Key Server, which required approximately 9.5 seconds. The majority of this time, however, is spent in authentication and is a fixed cost, which will decrease in its relative contribution to the overall runtime as datasets grow in size. The second corresponds to another fixed cost associated with the FS-based matching step performed at PAM. It can be seen that the Parameter Manager phase is quick. This is due, in part, to the fact that PAM performs an exact matching protocol (as opposed to a similarity comparison) because each record is represented by a single HMAC only. The cost is kept relatively low also because PAM performs this step over a subsample (1,000 records) of the datasets. The third additional cost corresponds to Bloom filter generation. This cost comes from the PRL match step, which is more variable in terms of running time.

Overall, the non-secure process requires around one minute, while the secure process requires about one and a half minutes. Though the secure process is roughly 1.5X slower, it should be recognized this is a relatively fixed cost. The majority of this increase is due to parameterization of the record encoding process. The record linkage process (step 8) itself is only 1.2X slower, which derives from the fact that it takes more time to compare two Bloom filters or several thousand bits (see bandwidth findings) than comparison over each person-level field used in this study.

Table 2. Average time (+/- standard deviation) in seconds for 10 runs across each step of the linkage protocols. Private record linkage (PRL) corresponds to the Four-Party protocol in Figure 1 via SOEMPI, while non-private record linkage (non-PRL) is the standard record linkage protocol in OpenEMPI. The steps seen here correspond to the protocol steps depicted in Figure 1. (a+b) refers to operations done both on Alice's and Bob's dataset (Figure 1).

Linkage Step	PRL	Non-PRL
0: Import datasets	6.70 (0.59)	5.58 (0.73)
1 (a+b): Obtain salts	0.09 (0.02)	-
2 (a+b): Message digest fields	0.04 (0.01)	-
3 (a+b): Send encoded datasets	2.31 (0.08)	-
4: Compute new parameters	5.32 (0.77)	-
5 (a+b): Obtain parameter advice	<0.01 (<0.01)	-
6 and 7: Create BFEs and send data	20.11 (3.36)	5.95 (0.38)
8: Block and match record pairs	58.74 (42.70)	47.86 (3.61)
Total Time required	**86.13 (43.52)**	**59.78 (4.12)**

Discussion

The findings from our experimental investigation illustrate that the lifecycle of private record linkage (using BFE encodings and a Four-Party protocol) is slower and requires more bandwidth than record linkage over identifiable patient records. However, at the same time, our results show that the costs do not incur drastic loss in speed or increase in memory footprint and that the PRL process lifecycle can be completed in a practical amount of time. At the same time, though SOEMPI was designed to be flexible, configurable, and enterprise capable, there are certain limitations to our current implementation that we wish to highlight, which can be enhanced in the future.

First, it should be recognized that the evaluation was a pilot study only. As such, the empirical analysis was performed over a dataset of 10,000 records only with a specific blocking and record linkage algorithm. Many of the authentication steps and transmission of key / salt processes will have negligible changes as the various parameters of the PRL process are changed. However, the speed of the system will vary with the size of the dataset, length of the Bloom Filter, and number of bits sampled for LSH change scalability of the system. It is thus recommended that a more comprehensive scalability assessment be performed before applying SOEMPI in larger record linkage frameworks.

At the same time, we wish to point out that the majority of SOEMPI operations are engineered to run in single-threaded processes. However, many of the procedures can be translated into multi-threaded versions, particularly LSH-based blocking[35] and record matching[36] to take advantage of modern parallel computing frameworks, such as MapReduce.

Second, from a PRL perspective, though SOEMPI can handle communication between the various participating entities, it does not implement any of the cryptographic primitives or protocols that have been proposed in PRL protocols based on secure multiparty computation (SMC). However, SOEMPI can readily incorporate various crypto-toolkits that have been implemented in Java (e.g., the UTD Paillier toolkit[37]). We believe SOEMPI can be an environment for integrating and managing various PRL protocols in a plug-and-play manner. Other Bloom filter-based solutions could be implemented with small effort in SOEMPI.

Third, from a technical stance, there are several aspects of the system that can be improved. Notably, EJB may not be the best technology for remote communication in the case of record linkage. While it is ideal for serving concurrent and independent web queries, as well as exchanging individual patient records, PRL has different needs. In particular, certain PRL protocols may require longer runtimes, as well as concurrent computations where threads should exchange information. To address these needs, special care will be required to increase the timeout values of the system and achieve synchronization. Moreover, our customized persistence technology supports only the PostgreSQL database management system (DBMS). However, SOEMPI uses standard database connectivity and thus is readily extensible to other DBMS technologies.

Finally, the virtual machine's JBoss application server is not fully secured and a SSL communication layer must be configured. As such, SOEMPI will need to undergo some additional security hardening when applied in the real world settings.

Conclusions

This paper introduced an open source software suite to support the private record linkage (PRL) lifecycle. SOEMPI's root goes back to extensively tested and proven existing Open Master Patient Index (OpenEMPI) reference implementation software to manage, transfer, and perform record linkage over encoded patient identifiers, but it is modified and enhanced in several key areas. We performed a meta-analysis to compare and contrast the proposed software toolkit with various freely available record linkage and PRL software toolkits that have been available to the biomedical research community. In addition to describing the software architecture and curtails of the specific technologies leveraged to realize the SOEMPI in working code, we provided a high-level depiction of how to build several PRL protocols. These protocols demonstrate how multiple third-parties, as well as data providers, can be integrated through a common communication and software system to cover the lifecycle. We also provided a runtime analysis of a Bloom filter-based PRL protocol that incorporates the lifecycle of encoding, blocking, and matching, and showed that such a procedure can complete in a practical amount of time.

Acknowledgements

This research was supported by grants CCF-0424422, CNS-1016343, and CNS-0964350 from the U.S. National Science Foundation and R01-LM009989, UL1-TR000135 from the U.S. National Institutes of Health. The authors would like to thank Steve Nyemba, MS from Vanderbilt University, Doug Bell, MD, PhD from the University of California at Los Angeles, Abel Kho, MD from Northwestern University, and Peter Christen, PhD from Australian National University for helpful discussions during the development of this software and for evaluating prototypes during their development.

References

1. Hernandez M, Stolfo S. Real-world data is dirty: data cleansing and the merge/purge problem. Data Mining and Knowledge Discovery 1998; 2: 9-37.
2. Christen P. Data matching: concepts and techniques for record linkage and duplicate detection. Springer. 2012.
3. U.S. Dept. of Health and Human Services. Standards for privacy of individually identifiable health information, final rule. Federal Register, 20 Feb 2003; 45 CFR: Pt 164.
4. Vatsalan D, Christen P, Verykios. A taxonomy of privacy-preserving record linkage techniques. Information Systems 2013; 38: 946-69.
5. Atallah M, Kerschbaum F, Du W. Secure and private sequence comparisons. Proc ACM Workshop on Privacy in the Electronic Society 2003: 39-44.
6. Schnell R, Bachteler T, Reiher J. Privacy-preserving record linkage using Bloom filters. BMC Med Inform Decis Mak 2009; 9: 41.
7. Kuzu M, Kantarcioglu M, Durham E, Toth C, Malin B. A practical approach to achieve private medical record linkage in light of public resources. J Am Med Inform Assoc 2013; 20: 285-92.

8. Randall SE, Ferrante A, Boyd JH, Bauer J, Semmens JB. Privacy-preserving record linkage on large real world data sets. J Biomed Inform 2014: in press.
9. Inan A, Kantarcioglu M, Bertino E, Scannapieco M. A hybrid approach to private record linkage. Proc IEEE International Conference on Data Engineering 2008: 496-505.
10. Kum HC, Krishnamurthy A, Machanavajjhala A, Reiter MK, Ahalt S. Privacy preserving interactive record linkage (PPIRL). J Am Med Inform Assoc 2014; 21: 212-20.
11. Durham E, Kantarcioglu M, Xue Y, Malin B. Quantifying the correctness, computational complexity, and security of privacy-preserving string comparators for record linkage. Inf Fusion 2012; 13: 245-59.
12. Weber SC, Lowe H, Das A, and Ferris T. A simple heuristic for blindfolded record linkage. J Am Med Inform Assoc 2012; 19: e157-61.
13. Schnell R, Bachteler T, and Bender S. A toolbox for record linkage. Austrian J Statistics 2004; 33: 125-33.
14. Bonomi L, Xiong L, Lu J. LinkIT: privacy preserving record linkage and integration via transformations. Proc ACM International Conference on Management of Data 2013 1029-32.
15. Pentakalos O and Xie Y. An extensible open source enterprise master patient index. Poster Presentation - AMIA Annu Symp Proc 2009. Software available at: http://www.openempi.org/.
16. Durham E, Kantarcioglu M, Xue Y, Toth C, Kuzu M, Malin B. Composite Bloom filters for secure record linkage. IEEE Transactions on Knowledge and Data Engineering. In press. DOI: 10.1109/TKDE.2013.91.
17. Grannis SJ, Overhage JM, and McDonald CJ. Analysis of identifier performance using a deterministic linkage algorithm. Proc AMIA Symp 2002: 305-9.
18. Grannis SJ, Overhage JM, Hui S, McDonald CJ. Analysis of a probabilistic record linkage technique without human review. AMIA Annu Symp Proc. 2003: 259-63.
19. Jurczyk P, Lu JJ, Xiong L, Cragan JD, Correa A. FRIL: a tool for comparative record linkage. AMIA Annu Symp Proc 2008: 440-4. Software online at: http://fril.sourceforge.net/.
20. Thoburn KK, Gu D, and Rawson T. Link Plus: Probabilistic record linkage software. Probabilistic Record Linkage Conference Call 2007.
21. Campbell KM. Rule your data with the Link King (a SAS/AF application for record linkage and unduplication). 30th SAS User Group International Meeting 2005.
22. Blasnik M. RECLINK: Stata module to probabilistically match records. Boston College Dept of Economics. 2010. Available online at http://ideas.repec.org/c/boc/bocode/s456876.html.
23. Wolfe BA, Mamlin BW, Biondich PG, et al. The OpenMRS system: collaborating toward an open source EMR for developing countries. AMIA Annu Symp Proc. 2006: 1146.
24. Available online at: http://openemed.org/
25. Available online at: http://www.openhre.org/
26. Christen P. FEBRL – a freely available record linkage system with a graphical user interface. Proc Australasian Workshop on Health Data and Knowledge Management 2008: 17-25. Software available online at: http://datamining.anu.edu.au/projects/linkage.html#prototype_software
27. Churches T and Christen P. Some methods for blindfolded record linkage. BMC Med Inform Decis Mak. 2004; 4: 9.
28. Available online at: http://www.record-linkage.de/
29. Bonomi L, Xiong L, Lu JJ. LinkIT: privacy preserving record linkage and integration via transformations. SIGMOD Conference 2013: 1029-32.
30. Available online at: ftp://www.app.sboe.state.nc.us/data
31. Christen P, Pudjijono, A. Accurate synthetic generation of realistic personal information. Proc Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining 2009: 507-14.
32. Massmind Nicknames Database. Online at: http://techref.massmind.org/techref/ecommerce/nicknames.htm.
33. U.S. Census Bureau, Population Division. Online at: http://www.census.gov/genealogy/names/namesfiles.html.
34. Kim H, Lee D. HARRA: a faster iterative hashed record linkage for large-scale data collections. Proc International Conference on Extending Database Technology 2010: 525-36.
35. Karapiperis D, Verykios V. A distributed framework for scaling up LSH-based computations in privacy preserving record linkage. Proc Balkan Conference on Informatics 2013: 102-9.
36. Yan W, Xue Y, Malin B. Scalable load balancing for MapReduce-based record linkage. Proc IEEE International Performance Computing and Communications Conference 2013: 1-10.
37. Available online at http://www.utdallas.edu/~mxk093120/cgi-bin/paillier/index.php.

An Algorithm Using Twelve Properties of Antibiotics to Find the Recommended Antibiotics, as in CPGs

R. Tsopra, MD[123], A. Venot, MD, PhD[123], C. Duclos, PharmD, PhD[123]
[1]INSERM, U1142, LIMICS, F-75006, Paris;
[2]Université Paris 13, Sorbonne Paris Cité, F-93000, Bobigny;
[3]Sorbonne Universités, Univ Paris 06, F-75006, Paris, France.

Abstract

Background. Clinical Decision Support Systems (CDSS) incorporating justifications, updating and adjustable recommendations can considerably improve the quality of healthcare. We propose a new approach to the design of CDSS for empiric antibiotic prescription, based on implementation of the deeper medical reasoning used by experts in the development of clinical practice guidelines (CPGs), to deduce the recommended antibiotics. Methods. We investigated two methods ("exclusion" versus "scoring") for reproducing this reasoning based on antibiotic properties. Results. The "exclusion" method reproduced expert reasoning the more accurately, retrieving the full list of recommended antibiotics for almost all clinical situations. Discussion. This approach has several advantages: (i) it provides convincing explanations for physicians; (ii) updating could easily be incorporated into the CDSS; (iii) it can provide recommendations for clinical situations missing from CPGs.

Introduction

The first Clinical Decision Support Systems (CDSS) were developed in the 1960s. Many were expert systems, designed to provide support for diagnosis and/or treatment decisions in a particular medical domain. Their development required collaboration between a medical expert and a computer scientist[1], with the medical knowledge and reasoning of the expert captured and implemented by the computer scientist. For example, INTERNIST-I[2] was a rule-based expert system providing support for multiple, complex diagnoses in general internal medicine. MYCIN[1] was a rule-based expert system developed for the diagnosis and treatment of infectious diseases.

In the 1990s, the concept of "Evidence-Based Medicine" was introduced and defined as "the integration of best research evidence with clinical expertise and patient values"[3]. This new paradigm led to the production and diffusion of Clinical Practice Guidelines (CPGs) by national health authorities[4]. CPGs are documents written by a group of experts for a particular domain, recommending diagnostic and therapeutic strategies on the basis of a systematic review of the available clinical evidence. However, CPGs are long, complex, textual documents that are difficult to use in daily clinical practice[5]. Many CDSSs were then designed to implement CPGs[6], rather than the medical knowledge of a single medical expert, to overcome these limitations and to take the concept of "evidence-based medicine" into account. Several formalisms were developed to facilitate and standardize the implementation of CPGs (e.g. Arden Syntax, EON, GLIF). These formalisms made it possible to associate "actions" (e.g. amoxicillin should be prescribed) to "patient conditions" (e.g. patient allergic to penicillin) in different ways: as Medical Logic Modules (MLMs) in Arden Syntax[7], an arborescence of the different clinical situations for a particular disease in Decision trees[8], a graph focusing on patient states in EON[9], or a flowchart of structured steps in GLIF[10].

However, there are two main problems associated with the implementation of CPGs in CDSSs. First, many clinical situations are not described in CPGs and are therefore not considered[11,12]. For instance, in some CPGs , experts give recommendations for uncomplicated cystitis, but not for cystitis in a woman with renal impairment. Second, CPG updating lags behind advances in medical knowledge[13], because updating takes time and it can be difficult to determine when an update is actually required[14]. For instance, in France, general practitioners have continued to prescribe amoxicillin-potassium clavulanate combinations for adult patients with sinusitis, as recommended in the CPGs written in 2005, but the frequency of acquired resistance to amoxicillin in *Haemophilus influenzae* has actually decreased considerably, making it possible to prescribe amoxicillin alone rather than the amoxicillin-potassium clavulanate combination. Physicians did not receive this information until 2011.

It may be possible to overcome these limitations by implementing the medical reasoning used to deduce medical "actions" from patient "conditions", rather than implementing "conditions"-"actions" combinations. In this

approach, the CDSS should be able to retrieve the recommended drugs without expert intervention. For example, for women with cystitis, CPGs recommend fosfomycin trometamol treatment.

(i) The usual approach is based on superficial associative medical reasoning, often involving the implementation of expert conclusions: i.e. the association "woman cystitis → fosfomycin trometamol" (e.g.: "if cystitis in a woman, then prescribe fosfomycin trometamol");

(ii) In our approach, we use a deeper reasoning: we try to implement the arguments used by the experts who wrote the CPGs to recommend one antibiotic, i.e. "In cystitis, the recommended antibiotic should have the following properties: it must be naturally active against *E. coli*, it must reach sufficiently high concentrations in the bladder, it must not be contraindicated in the patient, etc.". By taking these properties into account, the CDSS should be able to deduce that fosfomycin trometamol should be preferred over other choices for women with cystitis.

The use of this deeper medical reasoning should make it possible to retrieve the recommended treatment from patient and disease conditions. This should make it easier to cover a larger number of clinical situations, and should facilitate updating of the knowledge base.

We used the empiric prescription of antibiotics in primary care as a case study. In this domain, it is particularly important to update recommendations frequently, in accordance with advances in medical knowledge (e.g. the frequency of acquired resistance), because of the risk of bacterial resistance emerging[15]. For the testing of our approach, we needed to understand the deeper medical reasoning used by the experts to recommend a particular antibiotic over others in CPGs for a given clinical situation. We carried out a literature review, but found no accurate description of this deeper medical reasoning. However, we hypothesized that such medical reasoning could be extracted from CPGs and formalized.

The goals of our study were:

(i) To extract from CPGs the deeper medical reasoning on which experts based their recommendations concerning the antibiotics suitable for given clinical situations;

(ii) To implement this reasoning by two different methods and then to select the method giving the best automatic retrieval of the antibiotics recommended in CPGs.

We will first describe the two methods for reproducing the deeper medical reasoning of experts. We then present the study design for their evaluation and the results obtained.

Methods

An analysis of CPGs showed that the deeper medical reasoning used by the experts to recommend a particular antibiotic over others in CPGs was based on the use of the antibiotic properties. We began by identifying these properties and then investigated two methods for reproducing the medical reasoning taking these properties into account. Finally, we implemented and tested the two methods.

Extraction of the properties of antibiotics on which expert recommendations are based

We first extracted from the CPGs the properties of antibiotics used by the experts writing these CPGs to argument the recommendation of a particular antibiotic over others .

We analyzed seven CPGs: five were provided by French health authorities, one by the European Society of Clinical Microbiology and Infectious Diseases and one by both the Infectious Diseases Society of America and the European Society of Clinical Microbiology. They concerned 21 clinical situations relating to various diseases (cystitis, pyelonephritis, prostatitis, pharyngitis, otitis, sinusitis and pneumonia). We manually extracted all the expressions linked to the properties of antibiotics used to argument the preference of one antibiotic over others. Similar expressions were then grouped into categories of properties. For example, "natural sensitivity" and "natural activity" were grouped into the category "natural activity".

For each category of properties, we then added a question to determine whether the antibiotic considered displayed the property concerned. For example, the property "natural activity" was associated with the question "Does the antibiotic have sufficient microbiological activity against wild-type strains of the causal bacterium?" The response to the question, obtained from CPGs, was used to determine whether a given antibiotic had the property considered: if the response to the question was "yes" or, "no", then the considered antibiotic was considered to "have" or "not have" the property concerned, respectively, and if the response was "not available", then we considered that there was "no information available".

We then differentiated between:

(i) The "necessary" properties that an antibiotic must have to be usable in a patient, and to treat the infection. These properties ensure that an antibiotic is both safe for the patient, and able to cure the infection. These properties were used to obtain a list of appropriate antibiotics;

(ii) The "preference" properties that an appropriate antibiotic must have for that antibiotic to be preferred from a list of appropriate antibiotics, in a given clinical situation. These properties make it possible to choose one antibiotic through a list of antibiotics that could be prescribed to cure a patient. These properties were used to generate a list of recommended antibiotics.

Use of antibiotic properties to reproduce the deeper medical reasoning used by experts to generate a list of recommended antibiotics

We then tried to reproduce the deeper medical reasoning used by the experts writing CPGs, to generate a list of appropriate and recommended antibiotics, taking into account the properties of these drugs. We investigated two methods.

Constructing a list of appropriate antibiotics

For each clinical situation, we began with an initial list of antibiotics, all of which were potential candidates for recommendation. This initial list of candidate antibiotics differs between clinical situations and corresponds to all antibiotics described for the situation concerned in CPGs.

For each antibiotic on the list, we searched for answers to questions about necessary properties in CPGs. We excluded from the list all antibiotics for which there was at least one "no" or "not available" answer to the questions about necessary properties. The remaining antibiotics were considered to be appropriate.

Generating a list of recommended antibiotics by method 1

Method 1 involved calculating a score expressing the extent to which a particularly antibiotic satisfied the preference properties. The antibiotics with the highest scores were identified as those to be recommended by this method.

For each antibiotic from the list of appropriate antibiotics, we attributed a value according responses to questions about preference properties: if the response to the question was "yes", "no" or "not available", we attributed scores of "1", "-1" or "0", respectively, for the property concerned.

For each appropriate antibiotic, we then calculated the sum of the values attributed for all the preference properties. We retained the antibiotics with the highest scores and discarded the others from the list. The final list obtained with method 1 contained the antibiotics with the highest scores and should correspond to the list of antibiotics recommended in CPGs.

Generating a list of recommended antibiotics by method 2

In method 2, we excluded an antibiotic as soon as a property for preference was not satisfied. The antibiotics remaining in the list after the series of questions depended on the order of the questions in the sequence. The antibiotics remaining in the list, or if none remained, those excluded in response to the last question, were considered to be the recommended antibiotics according to this method.

For each successive question relating to preference properties, we excluded the antibiotic from the list if the answer was "no", but retained the antibiotic in the list if the response to the question was "yes" or "not available".

The list was, thus, progressively reduced after each question. The final list of antibiotics to be recommended corresponded to the antibiotics remaining in the list, or if there were no antibiotics remaining, those excluded by the last question. The final list should correspond to the list of antibiotics recommended in the CPGs.

As the final list depends on the order in which the questions are asked, we tested all possible sequences of questions for the 21 clinical situations, and selected the sequences that retrieved the list of antibiotics recommended in the CPGs for the largest number of clinical situations.

Implementation and evaluation of the two methods for reproducing the deeper medical reasoning used by experts in CPGs

We compared the two methods, by creating a database containing all the clinical situations and antibiotics described in CPGs (34 substances and 11 classes of antibiotics), the properties of which were extracted from CPGs. We implemented both methods in PHP.

We then tested each method as follows. First, for each clinical situation, we established:

- An initial list of candidate antibiotics for the testing of both methods, corresponding to all antibiotics described for the situation in CPGs (about 13 antibiotics per clinical situation);
- A list of the antibiotics recommended in CPGs, which we took as the gold standard.

For each clinical situation, we then applied the method to the initial list of candidate antibiotics and obtained a final list of antibiotics. This final list was then compared with the gold standard. If the final list of antibiotics obtained with the method corresponded exactly to the full list of antibiotics recommended in CPGs, then the method was considered "satisfactory" for the clinical situation.

We then calculated the total number of clinical situations for which the method was considered "satisfactory".

Finally, we compared the numbers of clinical situations for which a "satisfactory" result was obtained between the two methods. The method with the largest number of situations for which a "satisfactory" result was obtained was considered to be the best method for reproducing the deeper medical reasoning of experts for the empiric prescription of antibiotics.

Results

Properties of antibiotics used by the experts to formulate recommendations

Twelve antibiotic properties were retrieved in CPGs, in one or more clinical situations (e.g.: "natural activity" was retrieved for all clinical situations, whereas "availability" was retrieved for only one clinical situation). Two kinds of properties could be distinguished:

(i) "Necessary" properties (A to F, Table 1). Two of these properties related to the use of the antibiotic ("availability", "contraindication"). Four related to its potential efficacy ("natural activity", "likely activity", "concentration", "evidence of clinical efficacy"). Any antibiotic with all six necessary properties was considered "appropriate". For instance, amoxicillin, ampicillin, and penicillin V were all considered appropriate for pharyngitis;

(ii) "Preference" properties (G to L, Table 2). These properties related to the efficacy of the antibiotic ("level of efficacy", "protocol characteristics"), tolerance ("side effects") or ecological risk ("class characteristics", "spectrum of activity", "ecological adverse effects"). For instance, amoxicillin is recommended over ampicillin and penicillin for pharyngitis, because of the characteristics of its treatment protocol (shorter duration of treatment, favoring compliance).

Information about the properties described in CPGs was obtained from various resources:
(a) Results of clinical trials (properties: "evidence of clinical efficacy", "protocol characteristics", "level of efficacy");
(b) Clinical data (property: "contraindication");
(c) Microbiological data (properties: "natural activity", "likely activity", "spectrum of activity");
(c) Pharmacokinetics data (properties: "concentration in the infected organ");
(d) Pharmacovigilance (properties: "side effects");
(e) Drug marketing (property: "availability);
(f) Expert knowledge (properties: "class characteristics", "ecological adverse effects").

Table 1. Necessary properties used in expert medical reasoning as the basis for recommendations concerning antibiotic use. The frequency of use of the properties is the number of clinical situations in which the property is mentioned.

	Property	Questions relating to the property concerned, with examples of responses indicated in italics	Frequency of use (%)
A	Market availability	Is the antibiotic commercially available in the country? *No, pivmecillinam is not available in North America*	1/21 (5)
B	Natural activity against etiologic bacteria	Does the antibiotic have sufficient microbiological activity against wild-type strains of the causal bacterium? *Yes, amoxicillin is naturally active against Streptococcus pyogenes*	21/21 (100)
C	Concentration in the infected organ	Does the antibiotic reach sufficiently high concentration in the infected organ? *No, nitrofurantoin does not reach high concentrations in the kidney*	16/21 (76)
D	Evidence of clinical efficacy	Has clinical efficacy been proven in the clinical situation? *Yes, fluoroquinolones have been shown to be effective for acute pyelonephritis in a randomized controlled trial*	20/21 (95)
E	Likely activity against etiologic bacteria	Is the frequency of acquired resistance to the antibiotic low in the etiologic bacterium? *Yes, the frequency of acquired resistance to amoxicillin in Streptococcus pyogenes is less than 10%*	21/21 (100)
F	Contraindication in the patient	Is the antibiotic not contraindicated in the patient? *No, telithromycin is contraindicated in children under the age of 12 years*	17/21 (81)

Table 2. Preference properties used in the deeper medical reasoning of experts for antibiotic recommendations. The frequency of use of the properties corresponds to the number of clinical situations for which the property is mentioned.

	Property	Questions relating to the properties, with examples of responses given in italics	Frequency of use (%)
G	Protocol characteristics	Does the protocol for the use of the antibiotic favor compliance? *Yes, fosfomycin trometamol is prescribed as a single dose for uncomplicated cystitis*	17/21 (81)
H	Class characteristics	Does the antibiotic belong to a class that is not precious? *No, levofloxacin belongs to a precious class of antibiotics that should be reserved for serious indications*	6/21 (29)
I	Side effects	Is the antibiotic known to have (no serious or no frequent) side effects? *No, cefixime is associated with a high risk of pseudomembranous colitis caused by Clostridium difficile*	13/21 (62)
J	Level of efficacy (high, middle, low)	Is the antibiotic very effective? *Yes ciprofloxacin is highly effective in women with uncomplicated cystitis (clinical cure: 90% [85;98])*	17/21 (81)
K	Activity spectrum	Is the spectrum of activity of the antibiotic narrow? *No, levofloxacin has a broad spectrum of activity*	4/21 (19)
L	Ecological adverse effects	Does the antibiotic have a low risk of collateral damage? *No, first-generation quinolones promote the emergence of bacterial resistance*	14/21 (67)

Number of clinical situations for which a satisfactory result was obtained with each method

Method 1: Attribution of a relative score to antibiotics (illustration, Table 3)

With method 1, we obtained satisfactory results for 16 of a total of 21 situations (Table 4). The five clinical situations for which satisfactory results were not obtained were "uncomplicated pyelonephritis in women", "uncomplicated cystitis in women", "pharyngitis in adults without allergy", "maxillary sinusitis in adults without allergy" and "pneumonia in adults".

Method 2: Exclusion of antibiotics through a sequence of questions (illustration, Figure 1)

Permuting the questions about the six preference properties resulted in 720 sequences of questions (6!=720). We tested all 720 sequences for the 21 clinical situations, and tried to identify the most generic sequences (those giving satisfactory results in the largest number of clinical situations). Ten such sequences were identified:

"G, H, I, J, L, K" / "G, H, J, I, L, K" / "G, I, H, J, L, K"
"H, I, J, G, L, K" / "H, I, G, J, L, K" / "H, G, I, J, L, K" / "H, G, J, I, L, K"
"I, G, H, J, L, K" / "I, H, G, J, L, K" / "I, H, J, G, L, K"

The qualitative analysis of the 10 sequences showed that:

- In the 1st and 2nd positions of the sequence, we always found questions relating to properties "G" (Protocol characteristics), "H" (Class characteristics) or "I" (Side effects);
- In the 3rd position of the sequence, we always found questions relating to the properties "G" (Protocol characteristics), "H" (Class characteristics), "I" (Side effects) or "J" (Level of efficacy);
- In the 4th position of the sequence, we always found questions about property "G" (Protocol characteristics), "I" (Side effects) or "J" (Level of efficacy);
- In the 5th position of the sequence, we always found questions about property "L" (Ecological adverse effects);
- In the 6th position of the sequence, we always found questions about property "K" (Activity spectrum).

Method 2 gave a satisfactory response in 20 of the 21 clinical situations (Table 4), for these 10 sequences of questions. The only clinical situation for which a satisfactory result was not obtained was "uncomplicated cystitis in women".

Table 3. Method 1 – Attributing a relative score to antibiotics. (See the correspondence of properties A to L in Tables 1 and 2). Example of seven antibiotics, for pharyngitis in adults with penicillin allergy and without a contraindication for beta-lactams. All seven antibiotics were present in the initial list of candidate antibiotics. Two antibiotics did not have all the necessary properties, and were therefore excluded from the list ("amoxicillin" for property F, and "azithromycin" for property E). The five remaining antibiotics were considered appropriate. For each of these antibiotics, we attributed a relative value to each preference property. The sum of these values was maximal for three antibiotics ("cefuroxime axetil', "cefotiam hexetil", and "cefpodoxime proxetil"), which were therefore considered to be recommended by method 1. As they corresponded to the gold standard (i.e. the list of antibiotics recommended in CPGs), method 1 was considered "satisfactory" for this clinical situation.

	Necessary properties Responses to questions *(Y: Yes, N: No)*						**Preference properties** Attribution of a score						**Sum for G to L**	**Conclusion**
	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**		
Amoxicillin	Y	Y	Y	Y	Y	N	-	-	-	-	-	-	-	Inappropriate
Azithromycin	Y	Y	Y	Y	N	Y	-	-	-	-	-	-	-	Inappropriate
Cefaclor	Y	Y	Y	Y	Y	Y	-1	0	0	0	0	-1	-2	Appropriate
Cefuroxime axetil	Y	Y	Y	Y	Y	Y	1	0	0	0	0	-1	**0**	Recommended
Cefotiam hexetil	Y	Y	Y	Y	Y	Y	1	0	0	0	0	-1	**0**	Recommended
Cefpodoxime proxetil	Y	Y	Y	Y	Y	Y	1	0	0	0	0	-1	**0**	Recommended
Pristinamycin	Y	Y	Y	Y	Y	Y	0	0	0	-1	0	0	-1	Appropriate

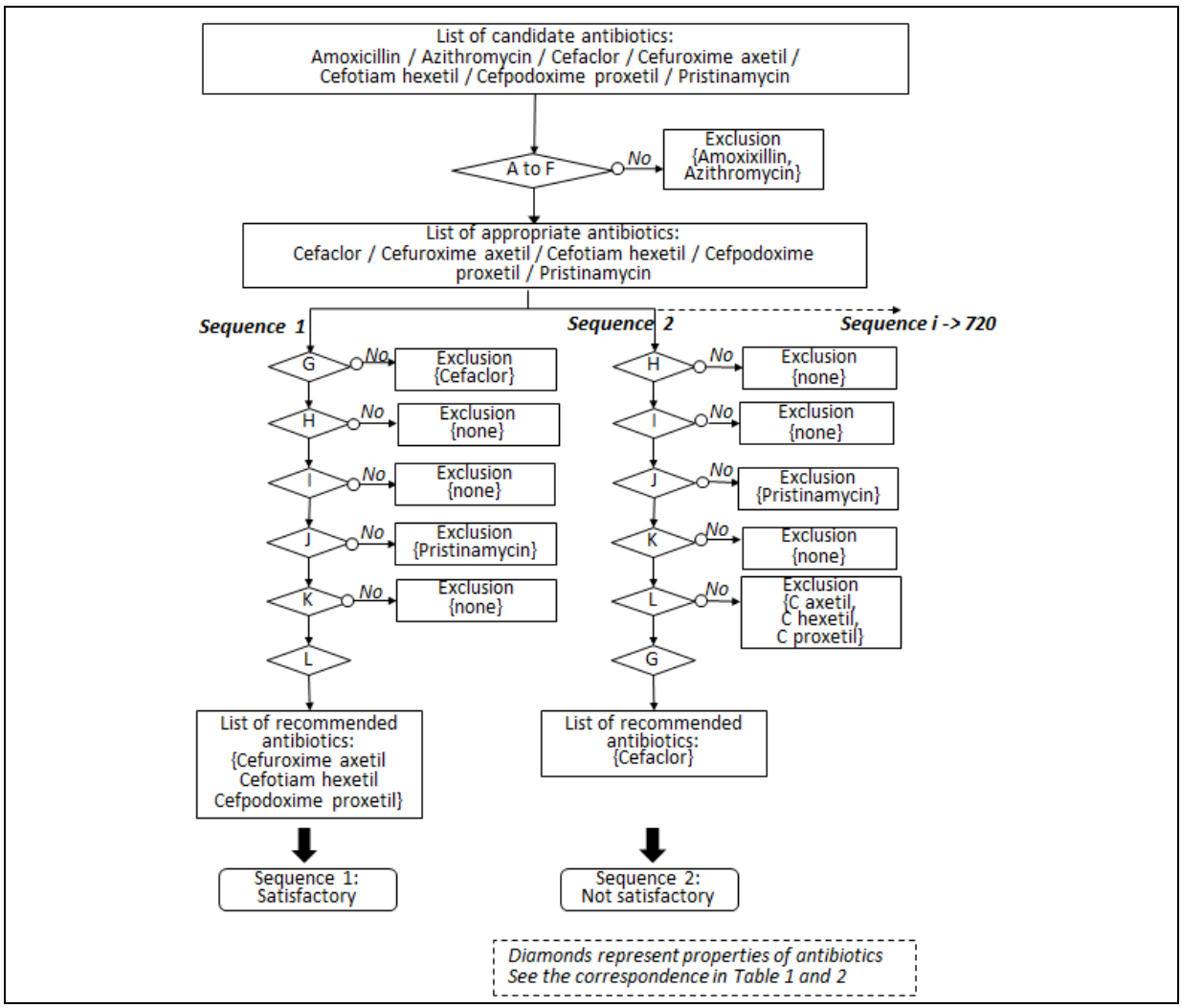


Figure 1. Method 2 – Exclusion of antibiotics according to a sequence of questions (see the correspondence of properties A to L in Tables 1 and 2). Example of seven antibiotics, for pharyngitis in adults with penicillin allergy and without a contraindication for beta-lactams. All seven antibiotics were present in the initial list of candidate antibiotics. Two antibiotics did not have all the necessary properties and were excluded from the list ("amoxicillin" for property F, and "azithromycin" for property E). The other five antibiotics were considered appropriate. For the six preference properties, 720 sequences of questions are possible by permutation of the properties. The final list of antibiotics obtained depends on the order of questions in the sequence. We illustrate the results for two sequences: sequence 1 generated a list of three antibiotics ("cefuroxime axetil', "cefotiam hexetil", and "cefpodoxime proxetil"), whereas sequence 2 yielded one antibiotic ("cefaclor"). As the list obtained with sequence 1 corresponds to the gold standard (i.e. the list of antibiotics recommended in CPGs), sequence 1 of method 2 is considered "satisfactory" for this clinical situation, whereas sequence 2 is not.

Table 4. Comparison of methods 1 and 2. Method 2 gives satisfactory results for a larger number of clinical situations than method 1.

	Method 1 – Attribution of a relative score to antibiotics	Method 2 – Exclusion of antibiotics through a sequence of questions
Number of clinical situations for which the result was considered "satisfactory"	16	20
Number of clinical situations for which the result was considered "not satisfactory"	5	1
Total	21	21

Discussion and Conclusion

We extracted the deeper medical reasoning underlying the CPGs, from the arguments used by the experts formulating recommendations concerning antibiotic use. We used this reasoning in two ways: "attribution of a relative score to antibiotics" and "exclusion of antibiotics through a sequence of questions". The "exclusion" method reproduced expert reasoning more accurately, as it retrieved the full list of recommended antibiotics for nearly all clinical situations (20 clinical situations, versus 16 for the "scoring" method). Furthermore, the only situation for which the "exclusion" method did not give a satisfactory result (uncomplicated cystitis) cannot really be considered a failure. In this situation, the experts produced a CPG that can be divided into several nation-specific recommendations, with a broad list of antibiotics that could be recommended in various countries, the final choice depending on local levels of acquired resistance. We gave equal weighting to all the properties of antibiotics considered. This approach yielded a highly satisfactory score, and we would have been unlikely to obtain a better result by weighting the properties differently.

Our approach is different from that of MYCIN[1] because (i) we used the arguments of a group of experts based on evidence-based medicine rather than the knowledge of a single expert; (ii) our method did not require interaction with the clinician; (iii) our method is simple and can generate recommendations very rapidly; (iv) our method can provide clinicians with an overview of the deeper medical reasoning underlying recommendations. This is not the case for MYCIN, for which the underlying reasoning is too complex to be presented in full to clinicians.

Our approach can be used to design a CDSS reproducing the deeper medical reasoning used by the experts writing CPGs. CDSSs generally implement the conclusions of the medical reasoning, i.e. the "actions" recommended for a particular clinical situation (e.g.: "*Amoxicillin is recommended for childhood pharyngitis*"). In our approach, we tried to implement the arguments underpinning the reasoning behind the recommendation, to make it possible to deduce the recommended antibiotics automatically (e.g.: "*for prescription for childhood pharyngitis, an antibiotic should have properties A to F, then, successively, properties G, H, I, J, L and K (see the correspondence in table 1 and 2)*"). The implementation of deeper medical reasoning, rather than its conclusions, has several advantages:
Recommendations can be justified and explained to physicians. As the deeper medical reasoning is based on the properties of the antibiotics, it can easily be understood by physicians. For example, a physician can easily understand that an antibiotic that has proved to be clinical effective and well tolerated by patients is preferred over an antibiotic that is effective but poorly tolerated. The provision of convincing and understandable explanations to physicians should increase their confidence in the CDSS, increasing the chances of its adoption[16]. Such explanations also help to provide physicians with up-to-date knowledge[17] and to develop their critical analysis capacities[17].
Recommendations may be easier to update. As the deeper medical reasoning is separate from the knowledge base containing the properties of antibiotics, it should be easier to update recommendations, and this process should be instantaneous[18]. The properties of antibiotics could be updated through various resources. For example, microbiological properties ("natural activity", "likely activity"; "activity spectrum") could be extracted from microbiological observatories. Properties relating to clinical data ("contraindication"), pharmacokinetics ("concentration"), or market availability ("availability") could be recuperated from drug databases[19]. The "side effects" property could be updated from pharmacovigilance databases. Properties relating to expert knowledge ("class characteristics", "ecological adverse effects") could be extracted from reference sources in infectious diseases. Similarly, evidence-based properties ("evidence of clinical efficacy", "protocol characteristics", "level of

efficacy") could be extracted from previous publications (e.g.: Medline). The incorporation of evidence-based medicine into CDSSs can considerably improve healthcare quality[14,20].

Recommendations could be given for clinical situations not described in CPGs. For example, the clinical situations described in CPGs for pharyngitis are: {adult; child < 6; child 6-12; child > 12} AND {without beta-lactam allergy; with penicillin allergy without cephalosporin contraindication; with beta-lactam contraindication}. The clinical situation "pharyngitis in pregnant women" is not described. With our approach, the system could deduce the antibiotics that should be recommended, by excluding all antibiotics contraindicated in pregnant women. Furthermore, the list of appropriate antibiotics could be used by physicians when they do not wish to prescribe the recommended antibiotics. For example, in uncomplicated cystitis in women, if the physician prefers not to prescribe the recommended antibiotic (fosfomycin trometamol) because it has been poorly tolerated by the patient in the past, he or she can select an alternative from the list of appropriate antibiotics. The provision of recommendations that can be adjusted to any clinical situation is likely to increase the compliance of physicians with recommendations[21].

This work now needs to be taken forward in several ways:

Confirmation of the robustness of our approach by expanding the evaluation to other clinical situations and to all the antibiotics available on the market. Both methods were assessed for urinary and respiratory infections, and only for the antibiotics described in CPGs. These methods should now be tested in other clinical situations (e.g. sexually transmitted infections, clinical situations specific to hospitals, etc.) and for all the antibiotics available on the market.

Confirmation of the validity of the recommendations generated by this method for clinical situations not described in CPGs, and for the updating of CPGs. Our approach was tested only for clinical situations described in CPGs, and not for other clinical situations. It will be necessary to test this method for clinical situations not described in CPGs, by taking the opinions of experts specializing in antibiotic treatment as the gold standard (because these situations are not described in CPGs).

Checking of the completeness of the list of the properties of the antibiotics identified in CPGs. We extracted, from CPGs, the properties of antibiotics most important for medical reasoning. It would be useful to collect the opinions of clinicians specialized in the domain of infectious diseases, to ensure that this list is complete.

Evaluation of the extent to which our approach could be extrapolated to other medical domains. The use of a deeper reasoning based on the arguments of the experts writing in CPGs, is particularly appropriate in the domain of antibiotic treatment, because arguments are explicit and related to the properties of drugs, including patient safety (contraindication), efficacy for curing the disease, and pharmaceutical properties (e.g. side effects). In other domains, other arguments would probably need to be taken into account, relating to temporal reasoning in chronic diseases, or to combinations of drugs. The possible extrapolation of this approach to other medical domains should therefore be investigated.

In conclusion, we propose a method for reproducing the deeper medical reasoning used by experts drawing up CPGs and underpinning the arguments used to justify the choices made. The robustness of this method should be assessed in a larger study before its implementation in a CDSS[22,23], to assist physicians in the empiric prescription of antibiotics in primary care. This CDSS will be assessed in clinical practice.

References

1. Shortliffe EH. Mycin: A Knowledge-based computer program applied to infectious diseases. Proc Annu Symp Comput Appl Med Care. 1977;66‑69.
2. Miller RA, Pople HE Jr, Myers JD. Internist-1, an experimental computer-based diagnostic consultant for general internal medicine. N Engl J Med. 1982;307(8):468‑476.
3. Sackett D. L, Straus S. E, Richardson W. S., Rosenberg W., Haynes R. B. Evidence-based medicine: How to practice and teach EBM. Edinburgh: Elsevier/Churchill Livingstone; 2000.
4. Woolf SH, Grol R, Hutchinson A, Eccles M, Grimshaw J. Potential benefits, limitations, and harms of clinical guidelines. BMJ. 1999;318(7182):527‑530.
5. Cabana MD, Rand CS, Powe NR, et al. Why don't physicians follow clinical practice guidelines? A framework for improvement. JAMA J Am Med Assoc. 1999;282(15):1458‑1465.
6. Lamy J-B, Ebrahiminia V, Riou C, et al. How to translate therapeutic recommendations in clinical practice guidelines into rules for critiquing physician prescriptions? Methods and application to five guidelines. BMC Med Inform Decis Mak. 2010;10:31.

7. Samwald M, Fehre K, de Bruin J, Adlassnig K-P. The Arden Syntax standard for clinical decision support: experiences and directions. J Biomed Inform. 2012;45(4):711‑718.
8. Séroussi B, Bouaud J, Antoine EC. ONCODOC: a successful experiment of computer-supported guideline development and implementation in the treatment of breast cancer. Artif Intell Med. 2001;22(1):43‑64.
9. Musen MA, Tu SW, Das AK, Shahar Y. EON: A component-based approach to automation of protocol-directed therapy. J Am Med Inform Assoc JAMIA. 1996;3(6):367‑388.
10. Boxwala AA, Peleg M, Tu S, Ogunyemi O, Zeng QT, Wang D, et al. GLIF3: a representation format for sharable computer-interpretable clinical practice guidelines. J Biomed Inform. 2004;37(3):147‑161.
11. Lugtenberg M, Zegers-van Schaick JM, Westert GP, Burgers JS. Why don't physicians adhere to guideline recommendations in practice? An analysis of barriers among Dutch general practitioners. Implement Sci IS. 2009;4:54.
12. Séroussi B, Laouénan C, Gligorov J, Uzan S, Mentré F, Bouaud J. Which breast cancer decisions remain non-compliant with guidelines despite the use of computerised decision support? Br J Cancer. 2013;109(5):1147‑1156.
13. Shekelle PG, Ortiz E, Rhodes S, et al. Validity of the agency for healthcare research and quality clinical practice guidelines: how quickly do guidelines become outdated? JAMA J Am Med Assoc. 2001;286(12):1461‑1467.
14. Takwoingi Y, Hopewell S, Tovey D, Sutton AJ. A multicomponent decision tool for prioritising the updating of systematic reviews. BMJ. 2013;347:f7191.
15. Davies J, Davies D. Origins and Evolution of Antibiotic Resistance. Microbiol Mol Biol Rev MMBR. 2010;74(3):417‑433.
16. Richard Ye L. The value of explanation in expert systems for auditing: An experimental investigation. Expert Syst Appl. 1995;9(4):543‑556.
17. Shankar RD, Martins SB, Tu SW, Goldstein MK, Musen MA. Building an explanation function for a hypertension decision-support system. Stud Health Technol Inform. 2001;84(Pt 1):538‑542.
18. Goldstein MK, Hoffman BB, Coleman RW, et al. Implementing clinical practice guidelines while taking account of changing evidence: ATHENA DSS, an easily modifiable decision-support system for managing hypertension in primary care. Proc AMIA Annu Symp AMIA Symp. 2000;300‑304.
19. Duclos C, Cartolano GL, Ghez M, Venot A. Structured representation of the pharmacodynamics section of the summary of product characteristics for antibiotics: application for automated extraction and visualization of their antimicrobial activity spectra. J Am Med Inform Assoc JAMIA. 2004;11(4):285‑293.
20. Sim I, Gorman P, Greenes RA, et al. Clinical decision support systems for the practice of evidence-based medicine. J Am Med Inform Assoc JAMIA. 2001;8(6):527‑534.
21. Lugtenberg M, Burgers JS, Besters CF, Han D, Westert GP. Perceived barriers to guideline adherence: A survey among general practitioners. BMC Fam Pract. 2011;12(1):98.
22. Tsopra R, Lamy J-B, Venot A, Duclos C. Design of an original interface that facilitates the use of clinical practice guidelines of infection by physicians in primary care. Stud Health Technol Inform. 2012;180:93‑97.
23. Tsopra R, Jais J-P, Venot A, Duclos C. Comparison of two kinds of interface, based on guided navigation or usability principles, for improving the adoption of computerized decision support systems: application to the prescription of antibiotics. J Am Med Inform Assoc JAMIA. 2014;21(e1):e107‑116.

Patient-Centered Appointment Scheduling Using Agent-Based Simulation

Ayten Turkcan[1], PhD, Tammy Toscos, PhD[2], Brad N. Doebbeling, MD, MSc[3]
[1]Department of Mechanical and Industrial Engineering, Northeastern University, Boston, MA; [2] Department of Nursing, Indiana University Purdue University, Fort Wayne, IN; [3]School of Informatics and Computing, Indiana University Purdue University, Indianapolis, IN

Abstract

Enhanced access and continuity are key components of patient-centered care. Existing studies show that several interventions such as providing same day appointments, walk-in services, after-hours care, and group appointments, have been used to redesign the healthcare systems for improved access to primary care. However, an intervention focusing on a single component of care delivery (i.e. improving access to acute care) might have a negative impact other components of the system (i.e. reduced continuity of care for chronic patients). Therefore, primary care clinics should consider implementing multiple interventions tailored for their patient population needs. We collected rapid ethnography and observations to better understand clinic workflow and key constraints. We then developed an agent-based simulation model that includes all access modalities (appointments, walk-ins, and after-hours access), incorporate resources and key constraints and determine the best appointment scheduling method that improves access and continuity of care. This paper demonstrates the value of simulation models to test a variety of alternative strategies to improve access to care through scheduling.

Introduction

Primary care clinics are under great pressure to improve access, health outcomes, quality and efficiency of care with limited availability of resources, especially in the current funding environment. The importance of providing patient-centered care with enhanced access and continuity has been emphasized in several studies.[1, 2] Interventions that vary according to the domain of care (acute care, preventive care, or chronic care) have been used to enhance primary care[3]. Same-day appointments, telephone triaging, walk-in centers, and after-hour services are used to improve access to acute care.[3] Multidisciplinary care teams, disease specific clinics, group appointments, registries, information and decision support systems, patient education and workforce development are used to improve access to care for chronic disease.[3] Community and population based programs to increase awareness, support systems for compliance and reminder systems are used to improve access to preventive care.[3] However, interventions that focus on a single component of care delivery are shown to have negative or no effect on other components of the system. For example, one study identified challenges in securing ongoing appointments for chronic care patients, due to increased availability of same-day appointments for acute care.[4] Another study evaluated the introduction of walk-in primary care clinic and reported good use by patients; however, there was no reduction in use of pre-existing services.[5] In another study, after-hours service did not show improvement in access to after-hours care, due to misconceptions about how to access the new system.[6] Therefore, multiple strategies targeting different levels or components of the health care system should be used to achieve the best performance.[3]

There are several access barriers to care including financial and non-financial barriers. While financial access barriers were found to be a primary problem experienced by 18% of US adults surveyed, 21% experienced non-financial barriers that led to unmet needs or delayed care[7]. The most commonly cited non-financial barrier was accommodation (challenges making appointments and ability to see a provider during limited hours)[8]. A systematic literature review revealed challenges for patients in obtaining appointments, longer waiting times, short business hours at the primary care clinics, unavailability of a regular physician or clinic, and low socioeconomic status to be associated with inappropriate use of emergency services.[9] Effective scheduling could address accommodation barriers; however, health centers struggle to implement novel scheduling methods due to lack of decision support tools that can be used to determine the best scheduling method tailored according to the patient population needs[10]. The number of strategies in the literature, the conflicting results of the existing studies, and the limited clinic resources make it difficult for clinic managers to determine the best set of interventions for their patient population and setting. In a recent study, two-thirds of the surveyed safety-net health centers did not have a process for same-day scheduling or had a process that needed improvement.[1] The aim of this study is to model the patient population

and primary care delivery system using simulation and then use the resultant model to determine optimal scheduling methods for improved access to care and continuity of care.

Background

In a project involving four community health centers (CHCs) in Indiana, we identified several challenges for patients in obtaining appointments including long waiting times, short business hours at the primary care clinics, and unavailability of a regular physician or clinic. The most commonly used scheduling method observed was a hybrid of traditional and advanced access scheduling. Patients with perceived need for routine follow-up (e.g. chronic conditions, routine visits for prenatal and well child check, and lab tests) are often scheduled in advance with a lead-time of 2 weeks to 9 months, depending on the condition, clinic, and provider. All clinics had adopted one or more strategies to provide some same-day access for acute care. Triage appointments (a number of appointment slots are kept open for acute problems) and walk-in hours are the strategies used to provide same-day access for acute demand. Overbooking was allowed by permission of the providers or, in urgent cases, patients were referred to local emergency rooms for care. Open access scheduling was used in one clinic where 80% of the appointments were kept open for same-day access. If a patient cannot be scheduled on the same day, he/she was asked to call back next day. Another clinic had a nurse practitioner (NP) whose schedule was explicitly kept open to provide same-day access to patients with acute care needs. Some clinics had walk-in hours on certain evenings and/or on Saturdays for established patients; an effective way of providing care at more convenient times for working patients.

Patient no-show was a major challenge for all clinics. Missed appointments reduce the continuity of care for no-show patients, reduce timely access to care for patients who cannot get an appointment, waste provider resources and can negatively impact health outcomes. While appointment reminders are often used to reduce no-shows, the missed-appointment rates were still higher than desired, about 25% for the clinics using traditional scheduling with triage appointments. No-shows worsen as the length of time between making the appointment and the actual visit date increased.

Implementation of open access scheduling may hold considerable potential in clinics where there is a high rate of patients who end-up going to emergency departments, even for non-urgent problems, because they cannot get a same-day appointment. The successful implementation of advanced access scheduling is reported to reduce no-shows, improve provider utilization and patient satisfaction. However, successful implementation requires careful analysis of clinic capacity and patient demand. Inappropriate proportions of capacity allocated for open access typically causes mismatch between capacity and demand, and result in implementation failure.[11] In our study, we noted that open access scheduling faces resistance from providers and staff due to uncertainties regarding its implementation and how it would actually work.

Previous studies have cited several reasons for no-shows including patient-related factors, scheduling system problems, and environmental and financial factors. Several interventions including appointment reminders, patient education, follow-up after a missed appointment, and open access scheduling have been used to reduce no-shows.[12] However, no-shows could not be eliminated completely due to several factors. We developed a logistic regression model that includes age, lead time (time between the appointment is made and the actual appointment time), prior no-show behavior, provider type, insurance type, and appointment type, as predictors of patient no-show.[12] Most of the existing literature ends with reporting the predictors of no-shows. Here, our aim is to use no-show prediction models to estimate actual demand and develop advanced scheduling methods considering individual no-show probabilities.

Many studies in the operations research literature propose appointment scheduling methods with the goal of improving clinic accessibility and efficiency. Cayirli and Veral[13] provided an extensive literature review with eighty papers in 2003. Since 2003, more than 300 papers cited the literature review paper of Cayirli and Veral[13], showing the significant growth of research on appointment scheduling. Existing studies consider appointment scheduling in different settings including primary care, specialty care and surgical departments. Most of the appointment scheduling studies focus on single stage, single resource environments (i.e. primary care or specialty care where only doctor appointments are considered). Earlier analytical studies used queuing theory and mathematical programming methods with simplifying assumptions to determine appointment schedules[13]. Most of the analytical studies could not be validated in real environments due to unrealistic assumptions. Simulation studies included

complex environmental factors such as unpunctuality, no-shows, walk-ins, etc. to find the best appointment scheduling rule. However, these studies were not easily implementable in other environments due to extensive data collection requirements. In recent years, analytical studies made more realistic assumptions including no-shows, cancellations, walk-ins, different patient types, priorities, and preferences. However, these advanced scheduling methods are rarely applied in clinical practice. In response to this gap, we sought to develop a simulation based tool that can be used in clinics to determine the best scheduling policy tailored to their patient population needs.

Over the past two decades, modeling and simulation have come of age as tools to help teams and managers support different cognitive and group processes. Simulation has been widely used in modeling health care systems in several settings, including outpatient departments.[14] Most health care simulation applications in the literature aim to provide better operational decision-making and planning tools.[15, 16] Simulation studies that model outpatient clinics focus on scheduling and capacity planning. Most of the studies model patient flow within the clinic and analyze the impact of scheduling methods on in-clinic patient waiting times and provider idle times.[16] However, the waiting time for an appointment (time between the time appointment is scheduled and actual appointment time) is also very important for clinics aiming to provide timely access to care.[10, 16] In this study, we focus on scheduling practices that directly accommodate the needs of patients while still meeting the efficiency needs of the CHCs and providers. We use agent-based simulation to model the care delivery system to incorporate population characteristics, care needs (demand), and common services provided within the health care system.[17]

Methods

We use simulation modeling, an operations research method, to model the scheduling process for provider appointments in CHCs. Rapid ethnography and observations are the key approaches used to understand the scheduling process and collect data for the simulation model. Rapid ethnography is a collection of methods used to understand the activities of users given significant time pressures and limited time in the field. The core elements of rapid ethnography include limiting or constraining the research focus and scope, using key informants, capturing rich field data by using multiple observers and interactive observation techniques, and collaborative qualitative data analysis.[18] For rapid ethnography and observations, we identify key informants including: 1) call center (or telephone room) and front office personnel as they relate to patient contact and scheduling, 2) enrollment specialists as they relate to new patients, 3) triage nurses as they relate to scheduling of acute appointments, and 4) quality assurance/information technology (QA/IT) personnel as they relate to the retrieval of appointment scheduling and provider capacity data. A team of two project members conducted interviews and obtained workflow summaries, policies, procedures, and artifacts used in the clinics. Based on the interviews performed, workflow diagrams are developed and then verified by the key informants. The simulation model is used to identify the impact of different scheduling and resource allocation strategies on key performance measures such as patient waiting times, provider productivity, patient no-shows, and continuity of care. The modeling tool is designed to be flexible so that it can be used by several clinics. The flexibility is achieved using input data files and graphical user interface. The inputs to the simulation model are:

- Provider capacity (capacity allocated for walk-ins, acute, non-acute and follow-up appointments)
- Patient population characteristics (age, gender, insurance, health status)
- Patient demand for care (demand for acute, non-acute and follow-up appointments)
- Scheduling method (traditional scheduling with triage appointments and open access scheduling)
- Scheduling horizon for each appointment type (acute, non-acute, follow-up)
- Maximum panel size for each provider

A data collection tool is developed to collect provider capacity, patient population characteristics, patient demand, no-show and cancellation data. The required variables are determined together with the QA/IT personnel, who retrieved the data from the electronic medical record system and sent the data for further analysis. The patient population characteristics including age, gender, insurance, and health status are provided as percentages with respect to the total number of patients. The provider template schedules are used to fill out the provider capacity data. The average number of annual visits per patient, no-shows and cancellations are calculated according to the appointment scheduling data provided by the QA/IT person. The data analysis results are discussed with the QA/IT personnel, COO and CEO of each clinic to validate the input data for the simulation model.

Figure 1 shows the general structure of the simulation model where the simulation model is initialized with clinic specific data (i.e. provider capacity and patient population characteristics), patient demand for care is generated

based on care needs of the patient population, and appointments are scheduled according to provider availability, capacity allocated for each appointment type, and appointment scheduling method.

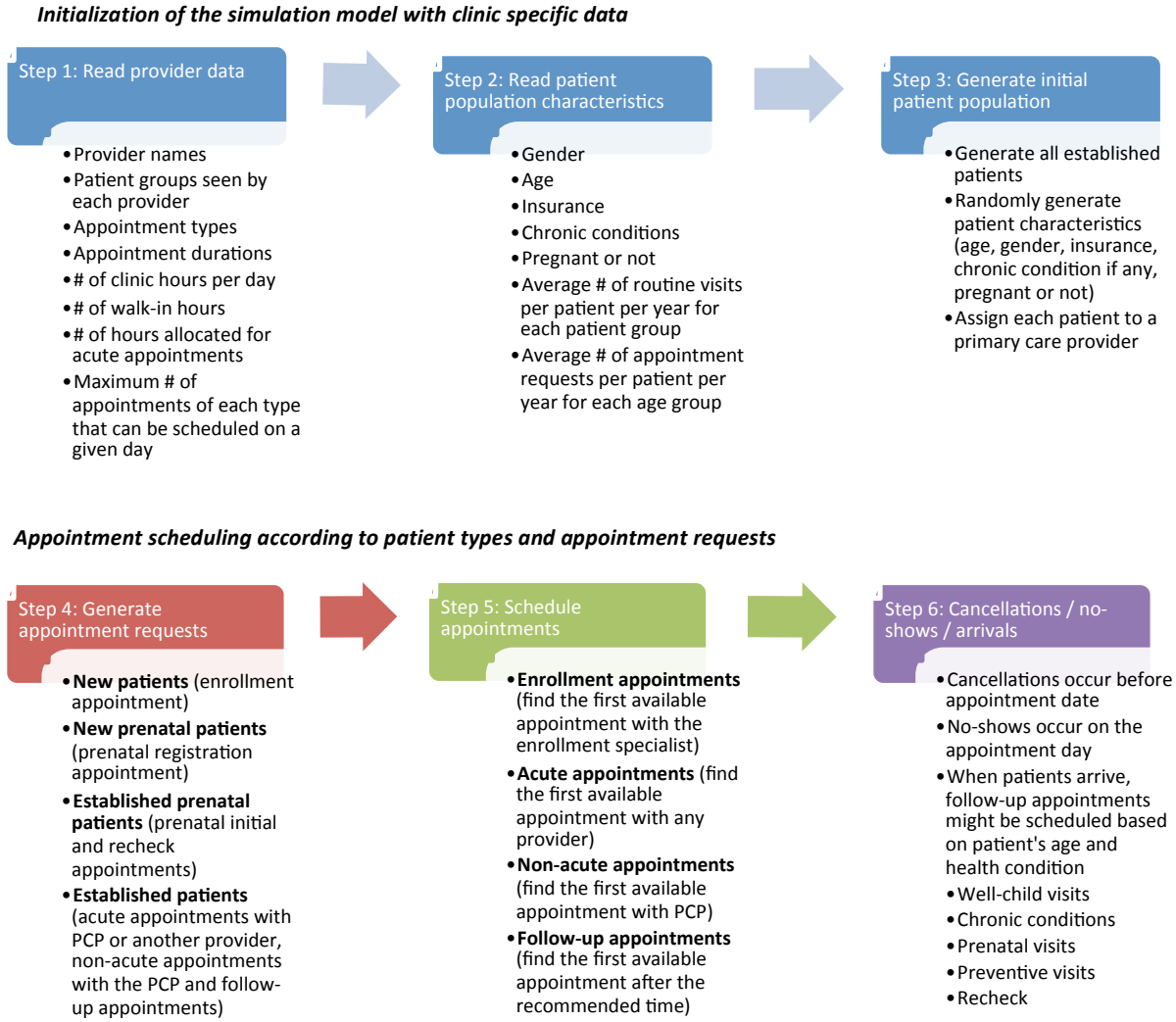


Figure 1. Simulation model

The simulation model starts with reading the provider and patient population data from an Excel file (Steps 1 and 2) and generating the initial patient population based on the number of unique patients served by the clinic (Step 3). Age, gender, insurance, and health condition (pregnancy and chronic condition) of each patient is generated based on patient demographics data. Each patient is assigned to a primary care provider based on available provider capacity and maximum panel size. The next step is generating the appointment requests for each patient (Step 4). We divided the patients into four groups based on their needs for different appointment types and resources. Once an appointment is requested, the appointment type (i.e. new, acute, routine, complex, well-child, newborn, etc.) is determined and scheduled according to patient type, resource availability, type of appointment request, and the scheduling policy used in the clinic (i.e. traditional approach with triage appointments, open-access scheduling) (Step 5). We included cancellations, arrivals and no-shows in our simulation model to calculate the access, operational, and quality measures (Step 6).

Initialization of the simulation model with clinic specific data

Step 1: The provider data required for the simulation model includes provider names, patient groups seen by each provider (i.e. pediatrics, adult, family medicine, women), number of working hours per day, number of hours allocated for acute/same-day appointments, appointment durations, and maximum number of appointments that can be scheduled. The providers allocate different durations for different types of appointments. For example, a provider might allocate 30-minute appointments for new and chronic patients, and 15-minute appointments for acute patients. The providers might have a different working schedule during the week. There might be restrictions on the number of appointments that can be scheduled on a given day (i.e. only two new patient appointments per day). An Excel file is prepared with the required information and the QA/IT person filled out that information by using the providers' template schedules. Tables 1 and 2 show sample provider data.

Table 1 – Appointment types and durations in minutes for each provider (sample data – does not include all possible appointment types)

Provider no	Providers	Patient group	New	Acute	Physical	Chronic/ complex	Well-child	Routine
1	MD1	Pediatrics	15	15	0	15	15	15
2	MD2	Adult	15	15	15	30		15
3	NP2	Adult	30	30	15	30		15
4	MD3	Women	45	30	45	45		30
5	NP2	Family medicine	30	15	30	30	30	30
6	MD4	Family medicine	30	15	15	30	15	15

Table 2 – Available provider capacity, number of hours allocated for same-day appointments, and maximum number of appointments of each type that can be scheduled on a given day (sample data – does not include all possible appointment types)

Providers	Day	Number of clinic hours	Number of walk-in hours	Number of hours allocated for acute	Number of hours allocated for non-acute	Maximum number of appointments that can be scheduled			
						New	Acute	Well child	Routine
MD1	Mon	7	4		3	2		11	
MD1	Wed	7		2	5	2		11	
MD1	Thu	7	4		3	2		11	
MD1	Fri	7		2	5	2		11	
MD2	Mon	7		2	5				
MD2	Tue	7		1	6				
MD2	Wed	7		1	6				
MD2	Thu	7		1	6				
MD2	Fri	7		2	5				

Steps 2-3: In operations research literature, simulation models, which consider modeling of care processes, patient flows and available resources, use historical data to express the demand in terms of a probability distribution function of the quantity or arrival time. In economic studies, which aim to evaluate the economic impact of diseases and health policies, demand is expressed as a function of prices, supplies, age, education, etc.[19-21] As Charfeddine and Montreuil[22] mentioned, demand for healthcare can be expressed in a better way through stochastic modeling of disease progression of each person in a patient population. Even though clinical studies model the disease progression and health status, their focus is to analyze the impact of medical interventions rather than determining the demand[23]. Healthcare demand is a function of multiple factors such as population characteristics, patient health status, treatment guidelines, adherence behavior, etc. We assume age, gender, insurance, and patient's condition

(chronic conditions, pregnancy) are predictors of number of visits. For example, the frequency and number of well-child visits change according to the age of patient. The frequency of prenatal visits change according to the stage of pregnancy. The number of routine visits change according to the health status of patients with chronic conditions. Agent-based simulation models can incorporate all these factors using agents (patient, provider, scheduler, etc.) to provide more realistic estimates of the care needs (e.g., regular provider visits, acute care visits). In the simulation model, we represent each patient as an agent. The simulation model generates the initial set of established patients at the initialization phase. The number of established patients is equal to the unique patients seen in the clinic in one year. The characteristics of each patient are generated based on initial patient demographics data. When the initial patient population is generated, each patient is assigned to a primary care provider (PCP).

Appointment scheduling according to patient types and appointment requests

Step 4: We divided patients into four groups according to two factors (new or established, and prenatal or not) due to the need for different appointment types and resources as shown in Figure 1. Once an appointment is requested, the appointment type (i.e. new, acute, routine, complex, well-child, newborn, etc.) is determined and scheduled according to resource availability, type of appointment request, and the scheduling policy used in the clinic. Figure 2 shows the overall summary of the scheduling process based on appointment types. This flowchart is prepared based on the workflow diagrams developed for the scheduling process through the key informant interviews.

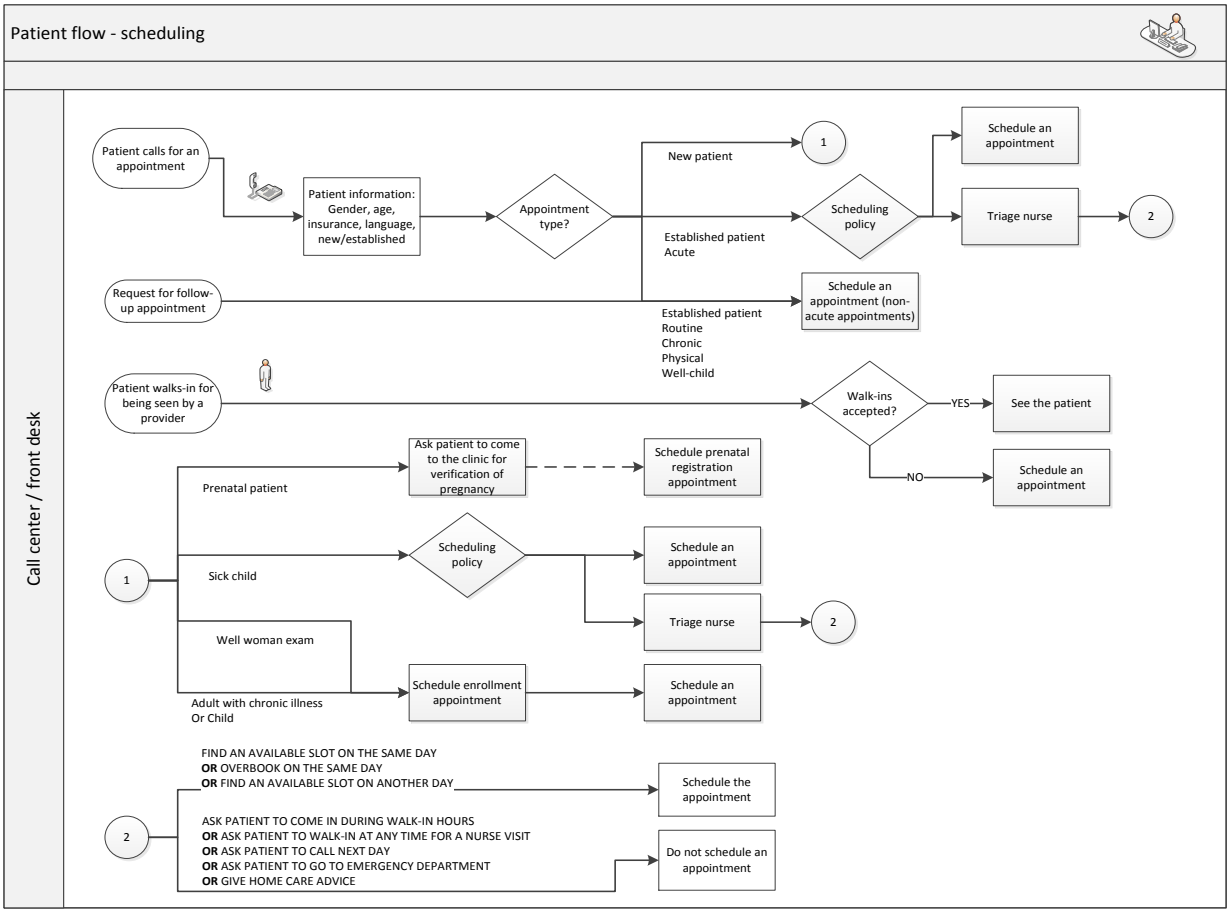


Figure 2 – Flowchart scheduling process according to appointment type and scheduling method.

Step 5: In the simulation model, we considered two types of scheduling methods (traditional scheduling with triage appointments, and open access scheduling). New patients should be scheduled for an enrollment appointment before an appointment can be scheduled with a provider. In traditional scheduling, acute appointments are scheduled on the same day or a few days in advance of triage appointment slots. Non-acute appointments and follow-up appointments

are scheduled to other slots several days in advance. In open access scheduling, most of the appointment slots are open for acute and non-acute appointments and can only be scheduled on the same day (or within a few days in advance). Follow-up appointments can be scheduled several days in advance, but the slots allocated for these appointments are limited. We used different objectives for scheduling of acute, non-acute and follow-up appointments. For acute appointments, our main objective was to minimize patient waiting time. Therefore, the available slots of PCP and other providers are searched to find an appointment time as soon as possible. For non-acute appointments, we sought to maximize continuity of care and minimize patient waiting time. Therefore, we first searched the PCP schedule and then other providers' schedules to find an available slot with a reasonable waiting time. For follow-up appointments, we maximized continuity of care by scheduling the appointment with the PCP.

Step 6: We included cancellations, arrivals and no-shows in our simulation model to determine the actual number of visits. Based on our prior research and existing literature[12, 24, 25], we assumed that the no-show probability is a function of age, insurance, lead time and patient's previous no-show behavior. In the simulation model, we used the regression equation that is determined based on clinic data to calculate the no-show probabilities. We would like to note that even though age, insurance and previous attendance rates cannot be controlled, they can be used to determine the no-show rate for each individual patient. The lead time can be controlled by changing the scheduling horizon (threshold for lead time) for each type of appointment. In our simulation model, scheduling horizon was an input parameter that could be controlled by the user or that determined by the model to minimize no-show rates and maximize continuity of care.

Results

We collected data from the clinics to validate the simulation model based on current practice in the clinics. Since an agent-based simulation model is used with several assumptions related to estimation of no-shows and cancellations, the validation of the input data is important. We compared the number of appointment requests, no-shows, cancellations, and provider utilizations of the simulation model with actual data to make sure that the demand was generated correctly. The simulation model was able to generate appointment requests, no-shows, cancellations and number of visits that were close to the observed data. For example, the average number of visits per week was 573 for the 2-week data provided by one clinic and the simulation model generated an average of 557 visits per week over a 5-year simulation run. The waiting times, percentage of appointments with the PCP, and no-show rates are used as performance measures since they are related to access and continuity of care. Table 3 shows a sample comparison between two scheduling methods. The waiting times for appointments are reduced in open access scheduling. The continuity of care did not change much. The no-show rates are reduced for non-acute appointments due to lower waiting times.

Table 3 – Comparison of traditional and open access scheduling methods

Performance measure	Traditional scheduling with triage appointments Planning horizon: (30, 90, 180)	Open access scheduling Planning horizon: (2,5,180)
Waiting time for acute appointments	Average: 0.5 day Std. dev.: 0.8 day	Average: 0.8 day Std. dev.: 0.4 day
Waiting time for non-acute appointments	Average: 51 days Std. dev.: 38 days	Average: 0.8 day Std. dev.: 0.4 day
Waiting time for follow-up appointments	Average: 48 days Std. dev.: 25 days	Average: 1 day Std. dev.: 1.2 days
Continuity of care (non-acute)	66%	65%
Continuity of care (acute)	57%	60%
No-show percentage (non-acute and follow-up)	23%	10%
No-show percentage (acute)	10%	9.6%

In this study, we could not perform statistical analysis for comparing and validating the simulation model, because the real data was not available for a longer time period. However, we are planning to address this issue in the

continuation of this project, which includes collection of 2-year appointment data from the clinics. We are also planning to use expert opinion for the validation of the simulation results.

Limitations

The proposed simulation model is developed based on current practice in two clinics that use traditional scheduling method. When open access scheduling is chosen, the simulation model reduces the no-show rates due to shorter waiting times. But the simulation model assumes that the patients will continue requesting appointments at the same rate. That is why more visits are scheduled and more arrivals occur when open access scheduling method is used. We believe the behavior of providers and patients will change when open access scheduling is implemented. For example, the providers may not ask the patient to schedule a follow-up appointment after the clinic visit unless it is really necessary. Due to limited number of follow-up appointment slots, the provider might ask the patient to come for a visit in 3-months. But the patient might forget to call to make the next appointment when 3-months pass. These two factors would reduce the number of appointment requests unless the clinics do not implement other interventions such as using provider and patient reminders to schedule the next appointment when the appropriate time comes. Currently, the simulation model does not incorporate these possible changes in behavior. As we develop this model further, we will work with clinics that implemented open access scheduling to be able to include the change in demand.

Next Steps

We are building upon this work with funding from Patient-Centered Outcomes Research Institute (PCORI). We will work with seven CHCs in Indiana to incorporate patients' perspectives to improve access to care for underserved populations. We will conduct rapid ethnography and workflow observation and modeling to identify the current barriers to access and use quantitative approaches to predict no-shows and determine the best scheduling policies that are determined based on patient population characteristics and available provider capacity. In that 3-year project, we will work with the clinics and further develop the proposed simulation model to include workflow and use the model and expert and patient panels to optimally determine the patient-centered interventions tailored for each clinic's patient population.

Conclusions

We used agent-based simulation to model the patient flow and appointment scheduling process in community health centers. The simulation model was designed to be flexible in terms of usability by different clinics. The inputs are entered through Excel files or graphical user interface, which makes the model useful for any clinic. This paper demonstrates the value of simulation models to test a variety of alternative strategies to improve access to care through scheduling.

Acknowledgments

This project was funded by MDwise, an Indiana not-for-profit health insurance company. We greatly appreciate our partnerships with CHCs, their management, staff, providers and patients in conducting this research. We also would like to acknowledge the funding received from Patient-Centered Outcomes Research Institute (PCORI) to develop patient-centered approaches to improve access to care for underserved populations.

References

1. Coleman K, Phillips K. Providing underserved patients with medical homes: assessing the readiness of safety-net health centers. Issue Brief (Commonw Fund). 2010 May;85:1-14.
2. Patient Protection and Affordable Care Act. Pub. L. No. 111-148, §2702, 124 Stat. 119, 318-319 2010.
3. Comino EJ, Davies GP, Krastev Y, Haas M, Christl B, Furler J, et al. A systematic review of interventions to enhance access to best practice primary health care for chronic disease management, prevention and episodic care. BMC Health Serv Res. 2012 Nov 21;12(1):415.
4. Subramanian U, Ackermann RT, Brizendine EJ, Saha C, Rosenman MB, Willis DR, et al. Effect of advanced access scheduling on processes and intermediate outcomes of diabetes care and utilization. J Gen Intern Med. 2009 Mar;24(3):327-33.
5. O'Cathain A, Knowles E, Munro J, Nicholl J. Exploring the effect of changes to service provision on the use of unscheduled care in England: population surveys. BMC Health Serv Res. 2007;7:61.

6. Dunt D, Day S, van Dort P. After Hours: Primary Medical Care Trials National Evaluation Report. Canberra, Australia: Commonwealth Australia, 2002.
7. Kullgren JT, McLaughlin CG, Mitra N, Armstrong K. Nonfinancial barriers and access to care for U.S. adults. Health Serv Res. 2012 Feb;47(1 Pt 2):462-85.
8. Kullgren JT, McLaughlin CG. Beyond affordability: the impact of nonfinancial barriers on access for uninsured adults in three diverse communities. J Community Health. 2010 Jun;35(3):240-8.
9. Carret ML, Fassa AC, Domingues MR. Inappropriate use of emergency services: a systematic review of prevalence and associated factors. Cad Saude Publica. 2009;25(1):7-28.
10. Rose KD, Ross JS, Horwitz LI. Advanced access scheduling outcomes: a systematic review. Arch Intern Med. 2011 Jul 11;171(13):1150-9.
11. Qu X, Rardin RL, Williams JAS, Willis DR. Matching daily healthcare provider capacity to demand in advanced access scheduling systems. European Journal of Operational Research. 2007;183(2):812-26.
12. Turkcan A, Nuti L, DeLaurentis P-C, Tian Z, Daggy J, Zhang L, et al. No-show modeling for adult ambulatory clinics. In: Denton B, editor. Healthcare Operations Management: A Handbook of Methods and Applications: Springer; 2013.
13. Cayirli T, Veral E. Outpatient scheduling in health care: a review of literature. Production and Operations Management. 2003;12(4):519-49.
14. Thorwarth M, Arisha A. Application of discrete-event simulation in health care: A review. 2010.
15. Fone D, Hollinghurst S, Temple M, Round A, Lester N, Weightman A, et al. Systematic review of the use and value of computer simulation modelling in population health and health care delivery. J Public Health Med. 2003 Dec;25(4):325-35.
16. Gunal MM, Pidd M. Discrete event simulation for performance modelling in health care: a review of the literature. Journal of Simulation. 2010;4:42-51.
17. Laffel G, Blumenthal D. The case for using industrial quality management science in health care organizations. Journal of American Medical Association. 1989;262(20):2689-873.
18. Millen DR. Rapid Ethnography: Time Deepening Strategies for HCI Field Research. In: Boyarski D, Kellogg WA, editors. Proceedings of the 3rd Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques. New York: ACM; 2000. p. 280-6.
19. Borg S, Ericsson A, Wedzicha J, Gulsvik A, Lundback B, Donaldson GC, et al. A computer simulation model of the natural history and economic impact of chronic obstructive pulmonary disease. Value Health. 2004 Mar-Apr;7(2):153-67.
20. Hoogendoorn M, Rutten-van Molken MP, Hoogenveen RT, van Genugten ML, Buist AS, Wouters EF, et al. A dynamic population model of disease progression in COPD. Eur Respir J. 2005 Aug;26(2):223-33.
21. Rutten-van Molken M, Lee TA. Economic modeling in chronic obstructive pulmonary disease. Proc Am Thorac Soc. 2006 Sep;3(7):630-4.
22. Charfeddine M, Montreuil B, editors. Integrated agent-oriented modeling and simulation of popula- tion and healthcare delivery network: Application to COPD chronic disease in a Canadian region. Proceedings of the 2010 Winter Simulation Conference; 2010.
23. Dev P, Heinrichs WL, Youngblood P, Kung S, Cheng R, Kusumoto L, et al. Virtual patient model for multi-person virtual medical environments. AMIA Annu Symp Proc. 2007:181-5.
24. Daggy J, Lawley MA, Willis DR, Thayer D, Suelzer C, DeLaurentis P-C, et al. Using no-show modeling to improve clinic performance. Health Informatics Journal. 2010;16(4):246-59.
25. Norris JB, Kumar C, Chand S, Moskowitz H, Shade SA, Willis DR. An empirical investigation into factors affecting patient cancellations and no-shows at outpatient clinics. Decision Support Systems. 2014;57:428-43.

Confidence and Information Access in Clinical Decision-Making: An Examination of the Cognitive Processes that affect the Information-seeking Behavior of Physicians

Raymonde Charles Uy, MD, MBA[1], Raymond Francis Sarmiento, MD[1], Alex Gavino, MD[1], Paul Fontelo, MD, MPH[1]
[1]National Library of Medicine, Bethesda, MD 20894

Abstract

Clinical decision-making involves the interplay between cognitive processes and physicians' perceptions of confidence in the context of their information-seeking behavior. The objectives of the study are: to examine how these concepts interact, to determine whether physician confidence, defined in relation to information need, affects clinical decision-making, and if information access improves decision accuracy. We analyzed previously collected data about resident physicians' perceptions of information need from a study comparing abstracts and full-text articles in clinical decision accuracy. We found that there is a significant relation between confidence and accuracy (φ=0.164, $p<0.01$). We also found various differences in the alignment of confidence and accuracy, demonstrating the concepts of underconfidence and overconfidence across years of clinical experience. Access to online literature also has a significant effect on accuracy ($p<0.001$). These results highlight possible CDSS strategies to reduce medical errors.

Introduction

Physicians are faced with many clinical decisions of varying complexity in their everyday practice. Most of these decisions revolve around questions on the diagnosis and therapeutic management of patients. The frequency of clinical questions varies from 0.16 to 1.27 per patient[1], depending on specialty, and up to 5 questions in the inpatient setting[2]. In the process of decision-making, physicians generally use their 'accumulated clinical knowledge', defined as a physician's personal knowledge base accumulated through years of formal education, medical training, research and clinical experience,[3, 4] to answer clinical questions. Accumulated clinical knowledge is traditionally used by physicians as it is the most convenient resource of information in the healthcare setting. This is especially highlighted in critical situations, such as the emergency department where there may be an immediate need for a diagnosis and management. However, reliance on this knowledge alone may lead to medical errors when these clinical questions remain unanswered or unsupported by more recent medical literature.

More physicians are integrating evidence-based medicine in their clinical practice. The conscientious, explicit and judicious use of current best evidence in clinical decision-making is the fundamental principle in the practice of evidence-based medicine (EBM)[5]. However, physicians are faced with numerous challenges that prevent the utilization of clinical decision support systems (CDSS), which are information systems designed to assist in clinical decisions by providing alerts, notifications, and links to important information personalized to patients, and the wider use of online evidence-based literature. Lack of time is seen to be one of major reasons that prevent physicians from accessing online literature to answer their clinical questions. In a previous study, 60% of physicians expressed that time was an issue in seeking information while in another study, 40% of clinical questions remained unresolved because of time constraints[6, 7].

Another hindrance that may be intrinsic to the medical profession is confidence in their own clinical knowledge. Physicians who do not perceive the need to access information when faced with a clinical question, may err in their clinical decision. In some cases, a physician may be unaware that there is a gap in their knowledge, thus making no attempt in remedying it[8]. Moreover, previous research demonstrated how some physicians did not even attempt to resolve clinical questions that they acknowledged might have been important in the management of their patients[9-11]. Some researchers proposed that the sociology of being a medical professional, which holds physicians that have expert knowledge and clinical competence in high regard, might explain this behavior[12, 13]. The act and perception of needing to seek knowledge may reflect negatively on a physician's competence and professionalism[14, 15].

In examining the process of how physicians make clinical decisions, an understanding of the cognitive processes of physicians must be considered. Clinical reasoning, which is a necessary cognitive process used to evaluate and manage a patient's medical problem[16], is a major process that defines clinical competence. This competency allows physicians to arrive at their diagnoses and treatment plans[17].

In our understanding of cognitive processes that complicate how physicians arrive at their decisions, we can see the role of confidence in clinical reasoning and information seeking behavior. Physicians who believe that their accumulated clinical knowledge is all they need to reach a correct decision will not be motivated to access external resources and may possibly be inflexible to change a planned management course when information in the EBM literature that may be contrary to their belief is presented[18]. Friedman, *et al.,* who first attempted to examine the relationship between confidence and clinical diagnosis accuracy, proposed four states of concordance or discordance that depend on the alignment of confidence and correctness[18, 19].

In states of concordance, physicians who are confident about their decision are indeed proven correct in their patient management. Conversely, those who are diffident, or are not confident, and are incorrect are also in a state of concordance. Physicians who fall in either of these two states of concordance exhibit appropriate confidence and their beliefs align with reality. The ideal situation would be physicians who are appropriately confident, while those who are appropriately diffident are likely to be receptive to accessing external sources of information. In contrast, states of discordance can be understood as underconfidence and overconfidence. Physicians who are not confident but are correct in their clinical decisions are deemed underconfident. Physicians in this state are more likely to seek information to confirm their decisions, however, it is still possible that doing so may steer them away from their originally correct answer[20]. The state of discordance that is of greatest concern is overconfidence. Physicians in this category are confident in their clinical decisions but in reality are incorrect and are therefore prone to committing medical errors. These physicians are less likely to pursue additional information that could correct their flawed decision.

Motivated by the complex interactions of confidence and decision-making in clinical reasoning, we aimed to determine whether there was a significant relationship between physicians' confidence in their accumulated knowledge in answering clinical cases and the accuracy of these decisions. This study attempted to address the following questions:

1. Is there a relationship between residency year level and clinical decision accuracy?
2. Is there a relationship between confidence and clinical decision accuracy?
3. Does the relationship between confidence and clinical decision accuracy depend per specialty or residency year levels?
4. Does accessing online literature affect clinical decision accuracy?
5. Does the effect of accessing online literature in clinical decision accuracy depend on specialty and residency year level?

Methods

Previous study

With permission from the authors, we reexamined the data collected by Marcelo *et al*. in their study where they compared the effects of using full-text articles against journal abstracts alone in clinical decision accuracy[3]. In that study, an attending physician from four different specialty departments (Family Medicine, Internal Medicine, Emergency Medicine, and Surgery) each prepared five clinical scenarios, specific to their specialty, and with varying complexities. The clinical cases were based on a PubMed search of recent journal articles that contained relevant clinical scenarios with their appropriate diagnostic and treatment recommendations. Resident physicians from each department answered five written clinical cases prepared by their respective attending physicians without access to online resources. After answering the case simulations, each resident was asked to indicate the clinical cases they thought could be accurately answered by accumulated knowledge alone and if they considered that a literature search was needed to arrive at the correct answer. After answering the written cases, each resident answered an online version of the same set of clinical cases, but was given the option to access relevant information either through journal abstracts or full-text articles. The attending physicians who prepared the cases evaluated the accuracy of each resident's answers in both the written (pre-intervention) and online (post-intervention) versions of the clinical cases.

Study design

In this study, we looked at the data on the resident physicians' perceptions on their need for information and their accuracy in answering clinical cases. In our analysis, the participants were additionally stratified based on specialty and year level. We examined the relationship between the residents' confidence and clinical decision accuracy. The data reviewed consisted of 322 cases. Excluded from this review were items where the resident did not indicate their initial perception (n=63).

Responses wherein the resident indicated that their accumulated knowledge was adequate to make a correct clinical decision were categorized to be confident. Responses where the residents expressed the need for accessing online PubMed sources were considered to be diffident. The residents' judgments of their confidence were completed without feedback on whether they were correct or incorrect. Clinical decision accuracy of each case was coded as correct or incorrect.

Data Analysis

In this study, each case paired a resident's subjective assessment of their confidence based on their need for an online literature search with an objective evaluation of their clinical decision accuracy based on correctness on the clinical cases. Chi-square analysis was done to determine if there is a relationship between confidence and clinical decision accuracy across all cases, and stratified based on specialty and residency year level. Cramer's V (V), Phi (φ) and Kendall's Tau-c (τ_c) values were used as symmetric measures to examine the strength of the association when the chi-square values were found to be statistically significant. McNemar's test was done to determine marginal homogeneity in matched pairs that examines the effect of online literature access by comparing clinical decision accuracy at pre- and post-intervention stages. Statistical analyses were performed using IBM SPSS Statistics Software Version 22 software.

Results

Irrespective of physician confidence, we found a small but significant relationship between residency year level and clinical decision accuracy (χ^2=8.077, df=3, p<0.05, τ_c=0.111). First year residents had more incorrect answers (63%) than their more experienced colleagues (45%, 53%, and 50% of 2nd, 3rd, and 4th year residents, respectively). In 228 of 322 cases (71%), residents expressed the need for a literature search to arrive at the correct clinical decision.

We found a small but statistically significant relationship between confidence and clinical decision accuracy in the pre-intervention scores across all cases (χ^2=8.697, df=1, p<0.01, φ=0.164,). Resident physicians were more incorrect (55%) than correct (45%), and more diffident (71%) than confident (29%) regardless of confidence. In total, 60% of the cases demonstrated appropriate confidence or diffidence where the residents' perceived need for information access was aligned with the accuracy of their answers. Calculating the risk ratios of physician confidence and clinical decision accuracy, a diffident physician was 1.42 times more likely to be incorrect than a confident physician. Resident physicians were more often underconfident (28%) than overconfident (12%). Residents were also more often appropriately diffident (43%) than they were appropriately confident (17%). (Table 1) shows a summary of findings.

Table 1. Contingency table on the Alignment of Confidence and Clinical Decision Accuracy

Physician Confidence	**Clinical Decision**		**Total**
	Correct	**Incorrect**	
Confident	17% Appropriately Confident	12% Overconfident	29% Confident
Diffident	28% Underconfident	43% Appropriately Diffident	71% Diffident
Total	45% Correct	55% Incorrect	100%

A separate analysis stratifying cases in their respective specialties found no statistically significant relationship between confidence and clinical decision accuracy.

When stratified based on residency year level, physician confidence had a small effect on clinical decision accuracy among first years (χ^2=4.97, df=1, p<0.05, *V*=0.199) and a medium effect among fourth years (χ^2=4.196, df=1, P<0.05, *V*=0.418). We could see a trend of increasing confidence as resident physicians progress in year levels. First year residents were the least confident (23%) compared to fourth year residents who were confident (46%) on the adequacy of their accumulated knowledge in answering the clinical cases.

Fourth year residents were more appropriately confident (33%) than first year residents (13%). First year residents were more appropriately diffident (52%) than their more senior residents (32%, 40%, and 37% for 2nd, 3rd, and 4th year residents, respectively). Underconfidence decreases when residents reach their 4th year (25, 38, 28, and 17% for 1st, 2nd, 3rd, and 4th year residents respectively) while overconfidence (10,12, 15, and 13% of 1st, 2nd, 3rd, and 4th year residents, respectively) remained consistent.

Association between online literature access and clinical decision accuracy

McNemar's test demonstrated a significant effect of online literature access in clinical decision accuracy (p<0.001) across all cases. Clinical decision accuracy increased from 45% in the pre-test to 72% in the post-test. When clustered by specialty, the effect is statistically significant at p<0.001 except in the Emergency Medicine specialty (p=0.108). Separately for each residency year level, literature access is seen to have a statistically significant effect on clinical decision accuracy except among forth year residents (p=0.180).

Discussion

Overall, we found a small but statistically significant relationship between residency year level and clinical decision accuracy. First year residents had more incorrect clinical decisions than their seniors. This may imply that CDSS might be beneficial when introduced earlier in medical training and education .This is consistent with Friedman's 2001 and 2004 papers that discussed how medical students had more incorrect diagnosis than residents and attending physicians.

In this study, a small but significant relationship between physicians' confidence and their accuracy in clinical decisions was found. Residents were more incorrect (55%) than correct (45%). This highlights the problem of how physicians underestimate their own error rates[25]. When made more aware of these errors, physicians may default to practicing defensive medicine, wherein their lack of confidence leads to unnecessary tests and studies, which are additional burdens to patients[21, 22]. Furthermore, as there was a larger number of instances where residents were not confident, which makes them 1.42 times more likely to be incorrect than they were if they were confident, strategies that increase their clinical competency and in effect, increase their confidence is needed.

Resident physicians were generally more concordant (60%), a setting where their confidence is aligned with correctness, than discordant (40%). In clinical practice, a concordance of confidence and clinical accuracy or correctness is favored more than in cases where there is misalignment. Within concordant cases, residents were also more often appropriately diffident (72%) than confident (28%). This demonstrates that there is a greater likelihood to be incorrect when lacking confidence compared to a marginal chance of being correct when feeling confident. Within cases of discordance, residents were more often underconfident (70%) than overconfident (30%). This finding is similar to the previous study's results that also demonstrated a tendency for clinicians towards underconfidence[32]. This has optimistic implications in the use of CDSS as these physicians are likely to confirm their decisions using external resources.

Additional findings demonstrate that the relationship of confidence and decision accuracy does not depend on specialty. Specialties were equally prone to errors, irrespective of confidence. When clustered by residency year level, a significant relationship was seen among first and fourth year residents, with the latter group having a stronger effect. First year residents were less confident than fourth years. This may be due to a perceived lack of clinical experience, or because they are overmatched by the difficulty of managing patients they see. This is further demonstrated by the fact that first years are more appropriately diffident than fourth years. Fourth year residents on the other hand were more appropriately confident than first year residents. Fourth year residents were also less underconfident compared to the younger residency years. Senior residents may perceive that they have higher clinical competency, and that their clinical experience is correctly matched to the cases compared to their juniors. This evidence of clinical experience bridging the gap between their confidence and accuracy of clinical decisions is similar to previous findings that suggest how higher levels of experience leads to an increased awareness of one's capabilities[32], which then improves the alignment of confidence and accuracy. Overconfidence remains consistent among year levels, which exhibits the role of CDSS in all levels of clinical experience.

Consistent with the findings from the previous study where data in this study was derived from, access to online literature search significantly affects clinical decision accuracy (from 45% to 72% accuracy). When stratified based on specialty, this effect was also significant, except among Emergency Medicine residents. In terms of residency year level, the effect was significant except among 4th year residents. These findings stress how providing access to literature search through CDSS affords statistically significant improvements in the accuracy of making evidence-based clinical decisions in a physician's everyday clinical practice[20].

Limitations

The same limitations from the previous study also apply. The sample size limits generalizability, while the use of clinical cases can only approximate actual clinical encounters. The limited number of cases answered per resident may have also reduced the possible variations in the study. These limitations may have contributed to the small effect sizes observed in the data analysis. In future studies, a larger sample size, pool of cases, and use of quantifiable measures for confidence may help address these limitations.

Conclusion

In this study, we discovered how physicians' confidence based on perceptions for needing access to online literature, plays a vital role in clinical cognition and decision-making. These results demonstrated how misalignments in confidence and clinical decisions due to cognitive processes may affect the accuracy of health care delivery to patients. We also showed further that online literature access improves clinical decision-making. The results indicate that clinical information should probably be 'pushed' to physicians who are overconfident and unlikely to seek information. Furthermore, easy and convenient access to information through different strategies such as practice guidelines, smartphones and mHealth at the point of care will likely benefit underconfident and diffident physicians. Previous studies of have shown the importance of convenient access to evidence where it's needed. These findings have diverse implications in design and implementation strategies of CDSS to support evidence-based practice.

Acknowledgement

This research was supported by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM) and Lister Hill National Center for Biomedical Communications (LHNCBC). This research was also supported in part by an appointment to the NLM Research Participation Program, administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the US Department of Energy (DoE) and the NLM.

Disclaimer

The views and opinions of the authors expressed herein do not necessarily state or reflect those of the National Library of Medicine, National Institutes of health or the US Department of Health and Human Services.

Competing Interests

None

References

1. Davies K. The information-seeking behaviour of doctors: a review of the evidence. Health Information and Libraries Journal. 2007;13:78–94. doi: 10.1111/j.1471-1842.2007.00713.x.
2. Osheroff JA, Forsythe DE, Buchanan BG. Physician's information needs: analysis of questions posed during clinical teaching. Ann Intern Med. 1991;114:576–581.
3. Marcelo A, Gavino A, Isip-Tan IT, Apostol-Nicodemus L, Mesa-Gaerlan FJ, Firaza PN, et al. A comparison of the accuracy of clinical decisions based on full-text articles and on journal abstracts alone: a study among residents in a tertiary care hospital. Evid Based Med. 2013 Apr;18(2):48-53. doi: 10.1136/eb-2012-100537.
4. Choudhry NK, Fletcher RH, Soumerai SB. Systematic review: the relationship between clinical experience and quality of health care. Ann Intern Med 2005;142:260–73.
5. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ 1996;312:71-2.
6. Green ML, Ciampi MA, Ellis PJ. Residents' medical information needs in clinic: are they being met. Am J Med. 2000 Aug 15;109(3):218–23.
7. Cogdill KW, Friedman CP, Jenkins CG, Mays BE, and Sharp MC. Information needs and information seeking in community medical education. Acad Med. 2000. May; 75(5):484–6.
8. Buckland MK. Library Services in Theory and Context, 2nd ed. Oxford: Pergamon, 1988.
9. Covell DG, Uman GC, Manning PR. "Information needs in office practice: are they being met?," Ann Intern Med. 1985 Oct;103:596–9.
10. Gorman PN, Helfand M. Information seeking in primary care: how physicians choose which clinical questions to pursue and which to leave unanswered. Med Decis Making. 1995;15(2):113–9.
11. Osheroff J A, Bankowitz R A. Physicians' use of computer software in answering clinical questions. Bull Med Libr Assoc. 1993 Jan;81(1):11–9.
12. Timmermans S, Mauck A. The promises and pitfalls of evidence-based medicine. Health Aff (Millwood) 2005;24:18–28.

13. Weaver RR. Computers and Medical Knowledge: The Diffusion of Decision Support Technology. Boulder, CO: Westview Press, 1991
14. Van der Sijs H, Aarts J, Vulto A, et al. Overriding of drug safety alerts in computerized physician order entry. J Am Med Inform Assoc. 2006;13:5–11.
15. Kahane S. Must we appear to be all-knowing?: patients' and family physicians' perspectives on information seeking during consultations. Can Fam Physician. 2011;57:e228–e236.
16. Barrows HS, Tamblyn RM. Problem-based learning. An approach to medical education. New York: Springer; 1980
17. Charlin B, Tardif J, Boshuizen HP. Scripts and medical diagnostic knowledge: theory and applications for clinical reasoning instruction and research. Acad Med 2000; 75: 18290
18. Friedman C, Gatti G, Elstein A, Franz T, Murphy G, Wolf F. Are clinicians correct when they believe they are correct? Implications for medical decision support. Medinfo. 2001;10(1):454–458.
19. Friedman CP, Gatti GG, Franz TM, Murphy GC, Wolf FM, Heckerling PS, et al. Do physicians know when their diagnoses are correct: implications for decision support and error reduction. J Gen Intern Med. 2005;20(4):334–339.
20. Friedman CP, Elstein AS, Wolf FM, Murphy GC, Franz TM, Heckerling PS, et al. Enhancement of clinicians' diagnostic reasoning by computer-based consultation: a multisite study of 2 systems. JAMA. 1999;13(19):1851–1856. doi: 10.1001/jama.282.19.1851.
21. Studdert DM, Mello MM, Sage WM. et al. Defensive medicine among high-risk specialist physicians in a volatile malpractice environment. JAMA. 2005;14:2609–2617. doi: 10.1001/jama.293.21.2609.
22. Anderson RE. Billions for defense: the pervasive nature of defensive medicine. Arch Intern Med 1999;159:2399–402.

An Integrated Billing Application to Streamline Clinician Workflow

David K. Vawdrey, PhD[1,2]; Colin Walsh, MD[1];
Peter D. Stetson, MD, MA[1,3,4]

[1]Department of Biomedical Informatics, Columbia University, New York, NY
[2]NewYork-Presbyterian Hospital, New York, NY
[3]Department of Medicine, Columbia University, New York, NY
[4]ColumbiaDoctors, New York, NY

Abstract

Between 2008 and 2010, our academic medical center transitioned to electronic provider documentation using a commercial electronic health record system. For attending physicians, one of the most frustrating aspects of this experience was the system's failure to support their existing electronic billing workflow. Because of poor system integration, it was difficult to verify the supporting documentation for each bill and impractical to track whether billable notes had corresponding charges. We developed and deployed in 2011 an integrated billing application called "iCharge" that streamlines clinicians' documentation and billing workflow, and simultaneously populates the inpatient problem list using billing diagnosis codes. Each month, over 550 physicians use iCharge to submit approximately 23,000 professional service charges for over 4,200 patients. On average, about 2.5 new problems are added to each patient's problem list. This paper describes the challenges and benefits of workflow integration across disparate applications and presents an example of innovative software development within a commercial EHR framework.

Introduction

In the United States, medical billing for a hospital encounter typically includes a bill for facility fees, and separate bills for professional services rendered by physicians. Facility fees cover costs for the room, nursing services, time spent in an operating room, supplies and medications, physical and respiratory therapy, and so on. Professional fees are submitted by physicians for the services they provide during the hospital stay. These services include performing evaluation and management tasks (such as conducting physical exams, diagnosing diseases, and formulating treatment plans) as well as performing surgical and other procedures.

Historically, physician bills were submitted using a standardized paper form called the "Universal Billing Form," which required codes for patient diagnoses and services performed. The modern electronic process requires the same coding requirement for diagnoses and procedures. The International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM) terminology is used to represent diagnoses; it is scheduled to be replaced by the ICD-10 code set in October 2014. Physician services and procedures are typically coded using the Current Procedural Terminology (CPT), a proprietary coding scheme maintained by the American Medical Association.

American Medical Association, the "Physicians' Current Procedural Terminology, Fourth Edition" (CPT-4) is a listing of descriptive terms and identifying codes for reporting medical, surgical, and diagnostic services performed by physicians. This five digit numeric coding methodology is not only utilized for billing purposes, but provides a uniform language applicable to patient care education, research, and utilization comparisons. The "International Classification of Diseases, 9th Revision" (ICD-9), is a numeric coding system describing diseases, symptoms, conditions, complications, external causes, as well as drugs and chemicals.

Between 2008 and 2010, our academic medical center transitioned to electronic provider documentation using a commercial electronic health record system (Allscripts Sunrise, Allscripts Corp., Chicago, IL). By December 2010, approximately 30,000 physician notes were electronically authored each month. For attending physicians, one of the most frustrating aspects of the transition to documenting in the EHR was the system's failure to support their existing electronic billing workflow. Physicians were required to author admission, follow-up, consultation, and discharge notes in the inpatient EHR, and log in to a separate system (the EHR used in their ambulatory practices) to submit charges. Users had to manually manage separate patient lists in each application. Furthermore, because of poor system

integration, it was difficult to verify the supporting documentation for each bill and impractical for administrators to track whether billable notes had corresponding charges.

We developed and deployed in 2011 an integrated billing application called "iCharge" with the goal of streamlining clinicians' documentation and billing workflow across two different commercial EHRs, and simultaneously populating the inpatient problem list using billing diagnosis codes in a timely fashion at the point of care. This paper describes the design and implementation of the iCharge application, focusing on the challenges and benefits of workflow integration across disparate applications.

Methods

iCharge Architecture

The integrated charge application was designed to be accessed from the inpatient EHR but provided the same look-and-feel as the physicians' existing billing system, which was a module within a separate ambulatory EHR system (Allscripts Enterprise EHR, Allscripts Corp., Chicago, IL). (In 2010, the inpatient EHR vendor, Eclipsys Corp., was acquired by Allscripts Corp., but the inpatient and outpatient EHR products remain separate products, with different databases and front-end interfaces.) iCharge was developed using Visual C# (Microsoft Corp., Redmond, WA), and communicated with the inpatient EHR using the vendor's included application programming interface (API) known as ObjectsPlus. The API allowed iCharge to leverage core EHR components such as role-based access and security auditing. Charges were submitted through web services (referred to as "Unity") provided by the ambulatory EHR infrastructure and would appear as normal charges in that system, with the substantial improvement of displaying the information (note name and author time) for the documentation that supported the charge. Figures 1–2 show illustrative screenshots of the application for a fake patient.

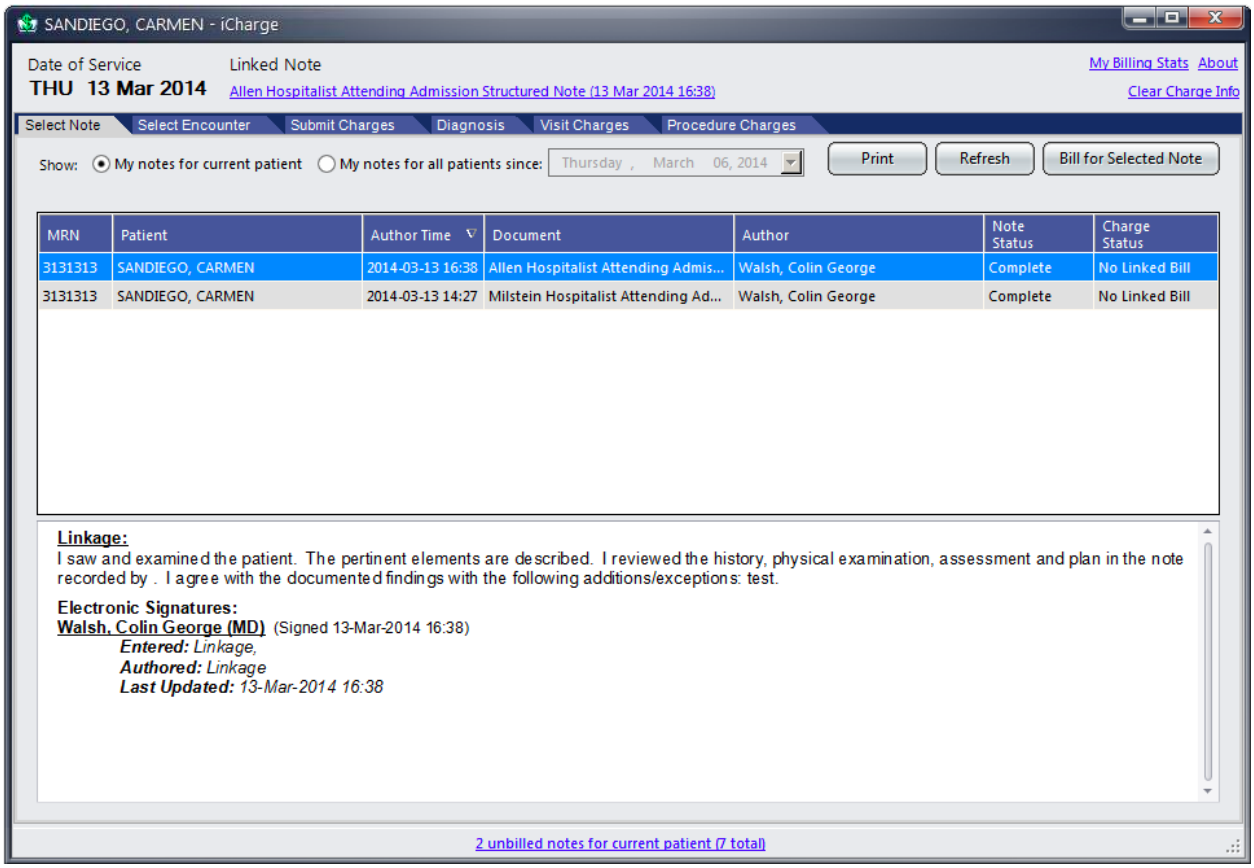


Figure 1. Screenshot of iCharge illustrating the "Note History" feature, where clinicians can review their billing history for each note authored in the inpatient electronic health record.

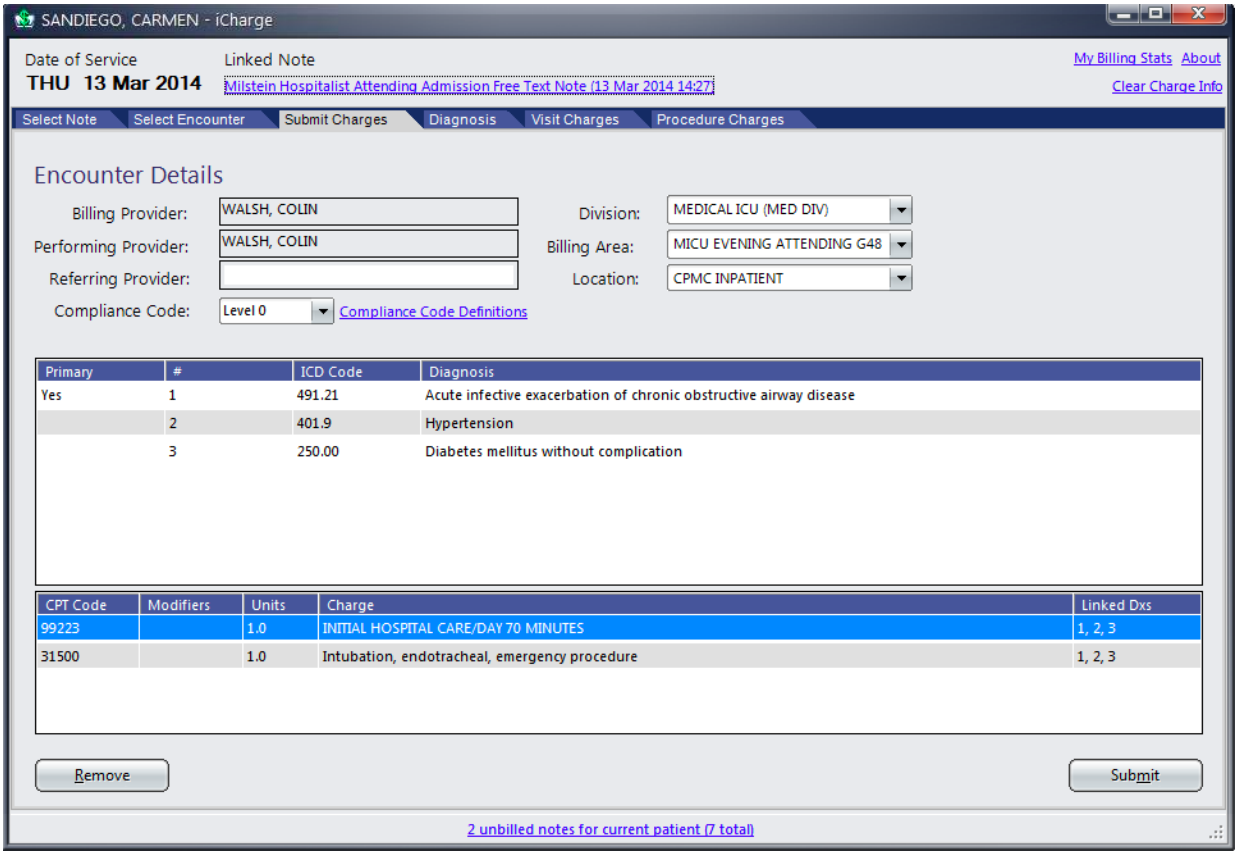


Figure 2. Screenshot of iCharge illustrating the "Note History" feature, where clinicians could review their billing history for each note authored in the inpatient electronic health record.

Workflow & Cognitive Support Characteristics

For health information technology to be most effective, it should provide users with the cognitive support necessary to complete the tasks it is designed to accomplish. (1) Informatics experts have observed that clinical information systems often lack adequate usability testing and demonstrate an apparent lack of clinical input in the design process. (2) The iCharge application was designed to be closely integrated with practitioners' documentation workflow. After a note was electronically signed by an attending physician in the EHR, a pop-up reminder (Figure 3) prompted him or her to generate a bill. Through this workflow, iCharge automatically associated the bill with its supporting documentation. Alternately, physicians could use a "Note History" function in iCharge to review and edit the billing status of all of the notes they authored in a specified time period. Using this function, clinicians could keep track of billable notes they had authored, but for which they had not yet submitted a bill. From within iCharge, clinicians could also view basic statistics about their billing practices compared to other iCharge users (e.g., number of bills submitted, average number of diagnosis codes included in each bill).

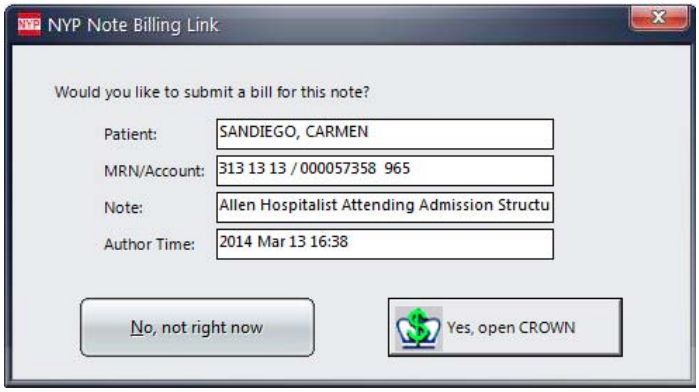


Figure 3. Reminder that appears upon signing an electronic note, prompting clinicians to submit a bill.

The iCharge application was designed to enable fast and efficient searching of diagnoses, which is an important feature as institutions throughout the U.S. prepare for the transition from ICD-9 to ICD-10 billing diagnosis codes. The diagnosis selection module allowed users to select from a patient's previous diagnoses and from lists of 'personal' and 'group' favorites (e.g., the 'Pediatric

Infectious Disease' group contained 16 frequently used billing diagnosis codes). A full catalog search provided suggestions for codes based on partial matches of numerical codes or text descriptions of ICD-9, ICD-10, or IMO (Intelligent Medical Objects) concepts. IMO (Northbrook, IL) provides an extensive terminology of physician-friendly terms that are mapped to ICD codes; the terminology has been incorporated into several commercial EHR systems. The iCharge diagnosis selection module was configured to prompt the clinician to specify a more granular code if an ICD-9 code was selected which mapped to more than one ICD-10 code.

In November 2013, iCharge was modified so that every billing diagnosis was automatically saved to the inpatient EHR's problem list. Prior to that time, physicians could manually save billing diagnoses to the problem list with an additional mouse click. The auto-save function was added as part of an institutional effort to improve the comprehensiveness and accuracy of the problem list.

Measurements

We retrospectively reviewed the system logs generated by iCharge and the inpatient EHR to evaluate system use. With data from the institution's practice management system, we calculated the average "turnaround time" for charges—the number of days between the date of service on a bill and the date the bill arrived in the practice management system.

For a three-month period, we measured monthly use of iCharge in terms of number of users, number of bills submitted, and patients with bills submitted. We queried from the inpatient EHR during the same time period the number of attending physicians authoring notes, number of notes authored, and number of patients with notes authored. We also measured from iCharge the mean number of billing diagnoses included per charge, as well as the average number of new diagnoses added to the inpatient Problem List per patient. Finally, we assessed the delay in days between the time a note was authored in the inpatient EHR and a corresponding charge was submitted via iCharge.

Results

The iCharge application was first made available to beta users in January 2011 and deployed to all attending physicians during the subsequent months. Its use was optional, but over time, the majority of billing clinicians in the institution began using the application (Figure 4). No formal training was conducted to instruct clinicians on the use of the application; however, one of the authors (PDS) created a short instruction guide, referred to locally as a "job aid," and other training materials that were made available to users of the application. By February, 2014, the majority of inpatient services were using iCharge, though some continued to bill using other means. A typical bill submission in iCharge required 30–60 seconds to complete.

Adoption

Use of iCharge was measured by reviewing audit log data from November 1, 2013 to February 1, 2014. Table 1 shows the number of users, charges submitted, average diagnoses per charge, average number of new diagnoses added for each patient.

Table 1. Use of an integrated billing application during three months at Columbia University Medical Center.

Measure	Nov 2013	Dec 2013	Jan 2014
Attending physician EHR note authors	775	793	805
Patients with notes authored by attending physicians	5,040	4,981	5,214
Total attending physician notes authored	32,424	31,413	33,901
iCharge users	427	430	436
Patients with charges submitted via iCharge	4,178	4,050	4,369
Total charges submitted via iCharge	23,225	23,364	24,616
Average number of billing diagnoses per charge in iCharge	2.70	2.81	2.80
Average number of new diagnoses added to inpatient problem list per patient via iCharge	2.54	2.68	2.45

Timeliness of Billing

Figure 5 shows a histogram of charges submitted in iCharge based on day of note authoring. The vast majority of charges were submitted on the same day that the supporting documentation (i.e., the attending physician's follow-up note) was authored. Approximately 9 months after iCharge was implemented, **t**he average turnaround time for charges had decreased from 5.67 days to 4.42 days (95% CI=0.09).

Problem List Impact

For the three months prior to enabling the auto-save of billing diagnoses to the inpatient EHR problem list, the number of manual save events (where physicians clicked to add their billing diagnoses to the problem list) averaged 37.1/day. After the auto-save feature was implemented, billing diagnoses were added to the problem list every time a charge was submitted; on average, 757 times/day. Duplicate diagnoses were not added. The average number of new diagnoses added to the problem list for each patient after the auto-save feature was added was about 2.5/patient (Table 1).

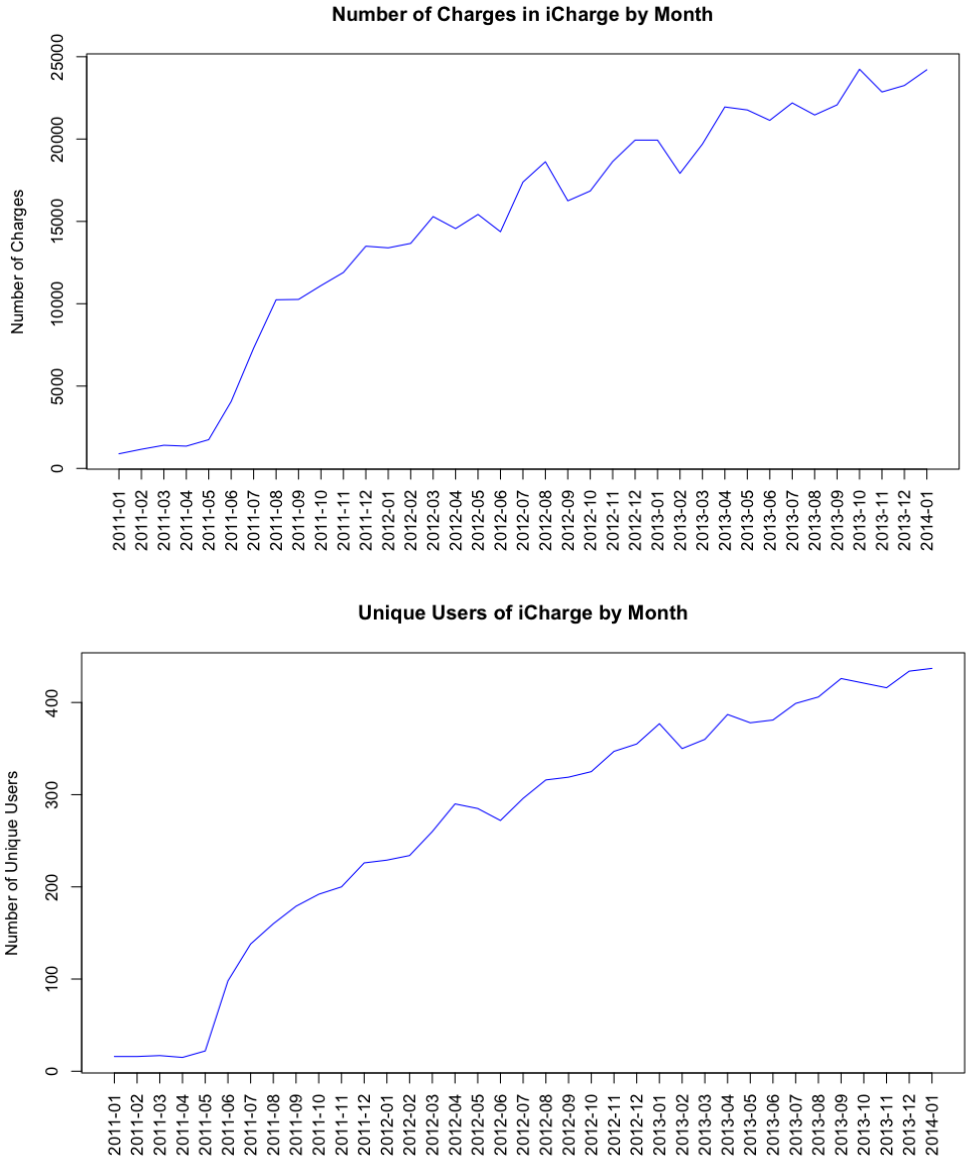


Figure 4. Monthly count of submitted bills and unique users of iCharge.

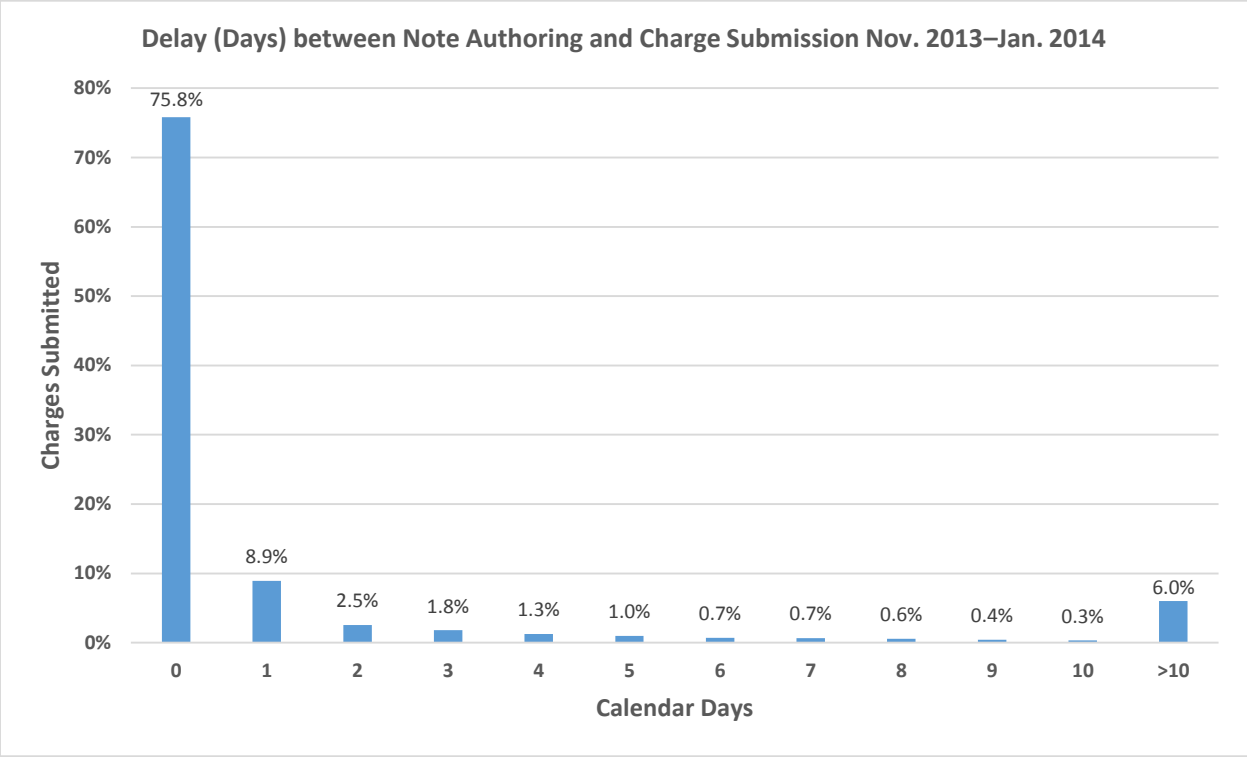


Figure 5. Frequency of charges submitted in iCharge based on day of note authoring.

Discussion

Importance of Workflow Integration

Even the most promising new informatics interventions may have little impact if they do not fit into clinical workflows (3, 4). The iCharge application was voluntarily adopted by over 400 physicians at our medical center primarily because it was closely integrated into their existing documentation workflow. With iCharge, physicians were no longer required to log into multiple systems, maintain multiple patient lists, or rely on their own memory to determine which notes had corresponding bills.

In spite of the fact that EHRs are frequently viewed as a means to promote "improved billing and collection," questions remain regarding the return on investment of these systems. (5-9). Unlike iCharge, the billing functionality in many commercial EHRs is not closely aligned with note-writing, even though there is strong evidence that tying note-writing to inpatient charge capture improves key financial indicators (10). Not only did iCharge offer clinicians the tools to better manage their documentation/billing workflow, it facilitated the creation of administrative reports that uncovered previously undetected trends in charge capture among providers, groups, and departments across our organization.

Benefits and Challenges with Implementing Custom Software in Commercial EHR Environment

Healthcare delivery organizations that use a single EHR system in across all care settings may have less need for solutions like iCharge; however, there are many valid reasons for institutions to develop custom software applications. For example, few institutions are likely to be perfectly happy with any commercial EHR system "out of the box." Most EHR vendors include some capability to develop custom software to address local needs. Moreover, because of the breadth and complexity of healthcare information technology, few organizations can rely on a single software vendor to meet all of their needs. Thus, interfaces and system integration will always be necessary. While many informatics discussions focus on interoperability at the data or system level (11), we believe that "workflow

integration" will become increasingly important. The iCharge project is one successful example of integrating workflows across disparate information systems.

Custom software development is not without its challenges. Healthcare delivery organizations are seldom staffed adequately to engage in large-scale software development efforts, particularly when it comes to testing, documentation, and maintenance. Information technology decision-makers should be judicious and pragmatic in deciding when to "build" and when to "buy" (12). Our organization has experienced a degree of success with custom application development to fill gaps that remained after installing a commercial EHR. We have benefitted from the vendor's open architecture, which enables us to create workflow-sensitive cognitive support tools, including for patient safety activities and regulatory/operational requirements (13).

Improving Problem List Documentation

One of the side benefits of using iCharge was that structured problems, in the form of diagnosis codes, were automatically added to hospital patients' problem lists. On average, patients whose physicians billed using iCharge gained an additional 2.5 problems. The need for a standardized approach to coding a problem list has been emphasized in the literature, but expecting providers to spend additional time coding problems in addition to their usual clinical work places demands on overall productivity (14-16). Prior work has outlined a number of useful approaches to increasing the number of problems on the problem list, including natural language processing, semantic annotation, predictive algorithms (17-22). In contrast to these approaches, the "auto-save" approach in iCharge may be less expensive both financially (i.e., in terms of human effort) and computationally.

Over the years, informatics experts have debated whether ICD-9 codes are sufficiently granular to be beneficial as a coding format for a patient problem list (23-27). We recognize that billing codes—in any format—may not perfectly represent the nuances of clinical care (28, 29), but we also believe that even a basic coded problem list has important clinical implications to ensure alerts are triggering on appropriate patients and that medications are being prescribed to those cohorts who need them (30, 31). The U.S. Meaningful Use financial incentive program requires hospitals to maintain a coded, structured problem list, and the program allows ICD-9 codes as a suitable format for encoding problems. Though it will likely involve a painful transition for many organizations, ICD-10 will provide clinicians with an order of magnitude more codes--and considerable more granularity—than ICD-9 codes afford. Thus, we anticipate that our method for automatically converting provider billing codes to problems on the problem list will continue to yield benefits for our organization, and may also generalize to others.

Conclusion

The integrated billing application that we created within our commercial EHR environment streamlines clinicians' documentation and billing workflow and simultaneously populates the inpatient problem list using billing diagnosis codes. The application has been widely adopted at our institution, demonstrating the importance of health information technology that matches the needs of busy clinicians. The success of the application provides evidence that innovative solutions can be implemented within the framework of a commercial EHR system.

References

1 In: Stead WW, Lin HS, editors. Computational Technology for Effective Health Care: Immediate Steps and Strategic Directions. Washington (DC); 2009.

2 Middleton B, Bloomrosen M, Dente MA, et al. Enhancing patient safety and quality of care by improving the usability of electronic health record systems: recommendations from AMIA. J Am Med Inform Assoc. 2013 Jun;20(e1):e2-8.

3 El-Kareh RE, Gandhi TK, Poon EG, et al. Actionable reminders did not improve performance over passive reminders for overdue tests in the primary care setting. J Am Med Inform Assoc. 2011 Mar-Apr;18(2):160-3.

4 Karsh BT. Beyond usability: designing effective technology implementation systems to promote patient safety. Qual Saf Health Care. 2004 Oct;13(5):388-94.

5 Zandieh SO, Yoon-Flannery K, Kuperman GJ, Langsam DJ, Hyman D, Kaushal R. Challenges to EHR implementation in electronic- versus paper-based office practices. J Gen Int Med. 2008 Jun;23(6):755-61.

6 O'Connell RT, Cho C, Shah N, Brown K, Shiffman RN. Take note(s): differential EHR satisfaction with two implementations under one roof. J Am Med Inform Assoc. 2004 Jan-Feb;11(1):43-9.

7 Menachemi N, Collum TH. Benefits and drawbacks of electronic health record systems. Risk Manag Healthc Policy. 2011;4:47-55.

8 Schmitt KF, Wofford DA. Financial analysis projects clear returns from electronic medical records. Healthc Financ Manage. 2002 Jan;56(1):52-7.

9 Jones SS, Heaton PS, Rudin RS, Schneider EC. Unraveling the IT productivity paradox--lessons for health care. N Engl J Med. 2012 Jun 14;366(24):2243-5.

10 Stetson PD, Keselman A, Rappaport D, et al. Electronic discharge summaries. AMIA Annu Symp Proc. 2005:1121.

11 Berger RG, Baba J. The realities of implementation of Clinical Context Object Workgroup (CCOW) standards for integration of vendor disparate clinical software in a large medical center. Int J Med Inform. 2009 Jun;78(6):386-90.

12 Thompson DI, Classen DC, Haug PJ. EMRs in the fourth stage: the future of electronic medical records based on the experience at Intermountain Health Care. J Healthc Inf Manag. 2007 Summer;21(3):49-60.

13 Vawdrey DK, Stein DM, Fred MR, Bostwick SB, Stetson PD. Implementation of a computerized patient handoff application. AMIA Annu Symp Proc. 2013:1395-400.

14 Kuperman G, Bates DW. Standardized coding of the medical problem list. J Am Med Inform Assoc. : JAMIA. 1994 Sep-Oct;1(5):414-5.

15 McDonald CJ. The barriers to electronic medical record systems and how to overcome them. J Am Med Inform Assoc. 1997;4:213-21.

16 Cusack CM, Hripcsak G, Bloomrosen M, et al. The future state of clinical data capture and documentation: a report from AMIA's 2011 Policy Meeting. J Am Med Inform Assoc. 2013;20:134-40.

17 Meystre S, Haug PJ. Natural language processing to extract medical problems from electronic clinical documents: performance evaluation. J Biomed Inform. 2006;39:589-99.

18 Meystre SM, Haug PJ. Randomized controlled trial of an automated problem list with improved sensitivity. Int J Med Inform. 2008;77:602-12.

19 Wright A, Chen ES, Maloney FL. An automated technique for identifying associations between medications, laboratory results and problems. J Biomed Inform. 2010;43:891-901.

20 Wright A, Pang J, Feblowitz JC, et al. Improving completeness of electronic problem lists through clinical decision support: a randomized, controlled trial. J Am Med Inform Assoc. 2012;19:555-61.

21 Kottke TE, Baechler CJ. An algorithm that identifies coronary and heart failure events in the electronic health record. Prev Chronic Dis. 2013;10:E29.

22 Mowery DL, Jordan P, Wiebe J, Harkema H, Dowling J, Chapman WW. Semantic annotation of clinical events for generating a problem list. AMIA Ann Symp Proc. 2013:1032-41.

23 Chute CG, Elkin PL, Fenton SH, Atkin GE. A clinical terminology in the post modern era: pragmatic problem list development. AMIA Ann Symp Proc. 1998:795-9.

24 Elkin PL, Mohr DN, Tuttle MS, et al. Standardized problem list generation, utilizing the Mayo canonical vocabulary embedded within the Unified Medical Language System. AMIA Ann Symp Proc. 1997:500-4.

25 Fung KW, McDonald C, Srinivasan S. The UMLS-CORE project: a study of the problem list terminologies used in large healthcare institutions. J Am Med Inform Assoc. 2010;17:675-80.

26 Nadkarni PM, Darer JA. Migrating existing clinical content from ICD-9 to SNOMED. J Am Med Inform Assoc. 2010;17:602-7.

27 Perotte A, Pivovarov R, Natarajan K, Weiskopf N, Wood F, Elhadad N. Diagnosis code assignment: models and evaluation metrics. J Am Med Inform Assoc. 2014;21:231-7.

28 Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: a perspective on the tension between structure and flexible documentation. J Am Med Inform Assoc. 2011 Mar-Apr;18(2):181-6.

29 Johnson SB, Bakken S, Dine D, et al. An electronic health record based on structured narrative. J Am Med Inform Assoc. 2008 Jan-Feb;15(1):54-64.

30 Hartung DM, Hunt J, Siemienczuk J, Miller H, Touchette DR. Clinical implications of an accurate problem list on heart failure treatment. J Gen Int Med. 2005 Feb;20(2):143-7.

31 Stockl KM, Le L, Harada AS, Zhang S. Use of controller medications in patients initiated on a long-acting beta2-adrenergic agonist before and after safety alerts. Am J Health Sys Pharm. 2008 Aug 15;65(16):1533-8.

Mining Consumer Health Vocabulary from Community-Generated Text

V.G.Vinod Vydiswaran, PhD[1], Qiaozhu Mei, PhD[1,2], David A. Hanauer, MD, MS[3,1], Kai Zheng, PhD[4,1]
[1]School of Information; [2]Department of Electrical Engineering and Computer Science; [3]Department of Pediatrics; [4]School of Public Health Department of Health Management and Policy. University of Michigan, Ann Arbor, MI

Abstract

Community-generated text corpora can be a valuable resource to extract consumer health vocabulary (CHV) and link them to professional terminologies and alternative variants. In this research, we propose a pattern-based text-mining approach to identify pairs of CHV and professional terms from Wikipedia, a large text corpus created and maintained by the community. A novel measure, leveraging the ratio of frequency of occurrence, was used to differentiate consumer terms from professional terms. We empirically evaluated the applicability of this approach using a large data sample consisting of MedLine abstracts and all posts from an online health forum, MedHelp. The results show that the proposed approach is able to identify synonymous pairs and label the terms as either consumer or professional term with high accuracy. We conclude that the proposed approach provides great potential to produce a high quality CHV to improve the performance of computational applications in processing consumer-generated health text.

Introduction

Over the past decade, there has been a significant increase in the consumption of online health information by the general public. According to a 2013 Pew Research Center survey, 72% of U.S. adult Internet users have looked for health information online; and, among them, 59% have sought information about specific medical conditions.[1] However, it has been long recognized that laypersons and healthcare professionals think about and express health-related concepts very differently,[2] for example, "dry mouth" vs. "xerostomia" and "flu" vs. "influenza". This mismatch in the terminology and style of writing could diminish laypersons' ability to effectively find and comprehend online health information written by professionals or experienced patients.

On the other hand, community-generated data on the Internet have also been increasingly used as a source of information to support professional needs such as public health surveillance and scientific discovery.[3] For example, scientists at Google analyzed search queries submitted by millions of users worldwide to detect the outbreak and spread of influenza-like epidemics,[4] and pharmaceutical companies are routinely monitoring online social conversations for post-market drug research.[5-8] In order to properly extract relevant concepts from community-generated text corpora, a high-quality consumer health vocabulary (CHV) is often needed to specify how a particular health-related concept may be expressed differently in laypersons' terms vs. in a professional language. The accuracy and comprehensiveness of CHVs can be crucial to the performance of computational tools that make use of community-generated health text such as health-related tweets and patient posts in online health forums.

Many resources are available for describing and classifying medical concepts used in the professional settings, such as the Systemized Nomenclature of Medicine Clinical Trials® (SNOMED-CT®), the Logical Observation Identifiers Names and Codes (LOINC), and other biomedical vocabularies and ontologies included in the Unified Medical Library System (UMLS) Metathesaurus. However, vocabularies providing consumer-oriented health terms are relatively less mature. This fact diminishes the performance of named-entity recognition tools for processing community-generated text[9] as well as the potential for building applications that could "translate" professional language into layperson terms to improve readability and facilitate comprehension (e.g. to support the OpenNotes project that shares clinician notes with patients[10]).

Community-generated text corpora could serve as a valuable resource to extract laypersons' expressions of medical concepts (i.e., consumer terms) and their corresponding professional expressions. Wikipedia is one such rich resource that is frequently updated and popularly accessed by the general public. It is estimated that the medical entries on Wikipedia are accessed more than 180 million times a month, and about 1,000–2,000 edits are being made to them each day.[11] Further, it is also estimated that about half of the Wikipedia users who edit the medical entries are healthcare professionals; the remainder are patients, families, and the general public.[11] Thus, it is likely that the

Wikipedia medical entries contain both the professional terminology and laypersons' terms, linked by some semantic relationships (e.g., "influenza, commonly known as the flu"), which provides the basis for this research.

In this paper, we propose a pattern-based text-mining approach to identify pairs of professional terms and their consumer variants from Wikipedia, in addition to their alternative spellings and synonyms. We also describe a computational approach to validate the extractions and to label the terms in a pair as either "professional" or "consumer". This approach is based on the frequencies of a term appearing in MedLine,[12] which indexes scientific papers produced by the professional community, and in MedHelp,[13] a popular online health forum where the content is mainly generated by laypersons. A subsequent manual review of the extracted and labeled pairs of entities was conducted to validate the results generated by the computational approach. The results are very promising.

Background

The use of Wikipedia by laypersons and medical professionals has become a subject of active research in recent years. Several studies have shown that Wikipedia is one of the leading online destinations for health information seekers.[14-16] For example, recent surveys reported that 60% of European doctors use Wikipedia for professional purposes,[14] and nearly 50% of U.S. physicians who go online for information on specific medical conditions use Wikipedia.[15,16] These studies corroborate the findings published in peer-reviewed articles on the use of Wikipedia as a source of information for scientific and medical professionals,[17-19] as well as medical students.[20-25] Studies have also shown a growing use of online medical resources including Wikipedia by patients, caregivers, and healthcare consumers,[26-29] and that Wikipedia articles often appear in the top results provided by Web search engines.[30]

Identifying all medical entities from free text is an active area of research in natural language processing. Lexico-syntactic pattern-based approaches have been well established for over thirty years and have supported numerous information extraction tasks such as hyponym identification,[31] semantic classification,[32,33] meronym identification,[34] and large scale information extraction over the Web.[35] This paper builds on similar ideas to identify and leverage key textual patterns that are frequently used to present synonymous terms in Wikipedia. Automatic term recognition (ATR) techniques, sometimes called named entity recognition (NER) techniques, have also been proposed to identify valid candidate terms in biomedical text corpora,[36-40] using sequential models[41-43] and term scoring approaches.[44,45] In this study, we investigated if the rich formatting styles used by editors in Wikipedia and other wiki-based corpora give sufficient cues to extract consumer vocabulary terms with high accuracy.

Methods

In this section, we present the process used to identify and extract names of medical entities and their alternate synonym variants from Wikipedia. Figure 1 summarizes the system design, and the following paragraphs describe the process in detail.

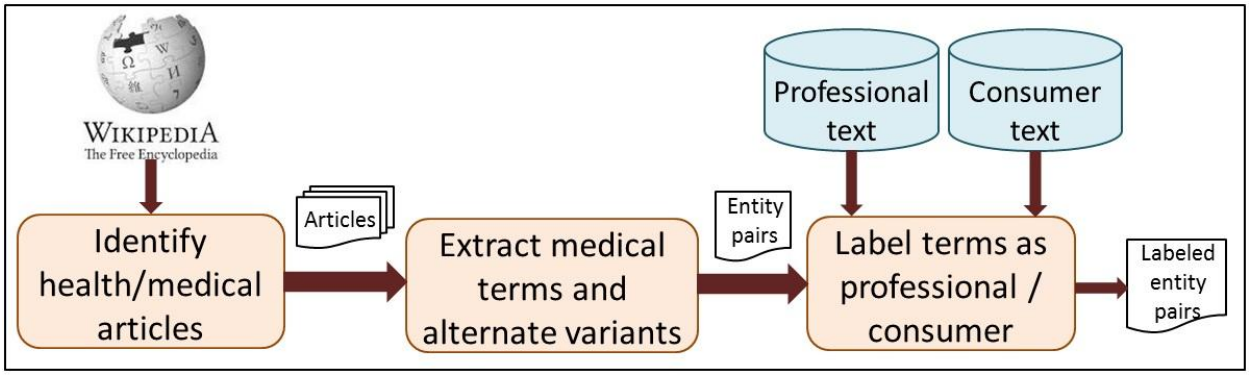


Figure 1. System work flow to extract and label professional and consumer health vocabulary.

Identifying relevant Wikipedia articles

Wikipedia releases periodic snapshots of all the articles published on the website, along with any associated metadata.[46] Each article contains the title, revision information, and the complete text. The text is typically unstructured and the formatting information, such as bold face or italicized fonts, citation information, and hyperlinks to other Wikipedia articles, is added with a marked up schema using special character sequences (such as two consecutive square brackets, three consecutive apostrophes, etc.). For this study, we considered the complete English language Wikipedia, which consists of over four million articles, and nine million additional pages for redirects, stubs, lists, and category pages.

Most Wikipedia articles also contain a list of tags that specify, for each article, a set of relevant categories from Wikipedia's hierarchical topic classification scheme.[47] At the top most level, this scheme consists of twenty five categories. Two of these are *Health* and *Medicine*. These top-level categories are further divided into additional sub-categories. For example, *Clinical medicine* is a sub-category of *Medicine*, and *Medical diagnosis* is a sub-category of *Clinical medicine*. The complete category hierarchy can be obtained by repeatedly traversing the sub-category links.

We collected a list of all sub-categories for *Health* and *Medicine* down to a depth of three, using an external Wikipedia tool called CatScan.[48] CatScan recursively searches an article category to find all articles, sub-categories, images, etc. This resulted in a list of 2,331 candidate categories. Once the relevant categories were identified, all four million Wikipedia articles were programmatically checked to retain only those articles that were tagged with at least one of the candidate categories. This narrowed the number of potential medically relevant articles to about 46,000, which constitutes about 1.2% of all articles in English Wikipedia.

Extracting medical terms and their common alternate names

Wikipedia articles are often written and formatted to serve as an introduction to the topic of the article.[49] Typically, the first sentence is formatted such that the article title appears in bold face as the subject of the definition or introduction. If the title also has alternate forms, such as abbreviations, alternate spellings, or significant alternate titles, these also appear as bold face immediately following or adjacent to the first occurrence of the title. For example, the Wikipedia article on "xerostomia"[50] starts as:

> **Xerostomia** (also termed **dry mouth** or **dry mouth syndrome**) is the medical term for the subjective symptom of dryness in the mouth, …

informing readers that "dry mouth" and "dry mouth syndrome" are alternative names for "xerostomia".

This relatively consistent style lends itself to devising automated text-mining and information extraction techniques. Through an iterative process of reviewing medical articles on Wikipedia, we identified a list of phrases that generally connected a medical concept term to its alternate terminology. Table 1 lists twelve examples of common linking phrases.

Table 1. Common linking phrases.

also called	commonly called	sometimes called	also termed
also known as	commonly known as	sometimes known as	previously known as
also referred to as	commonly referred to as	sometimes referred to as	colloquially known as

In addition to the linking phrases, parentheses also serve as a textual clue to introduce alternate spellings, abbreviations, and synonyms. Multiple alternate terms are mentioned in a comma-separated list, or separated by conjunctions such as "or" and "and". Each alternate form is typeset in bold face or italicized text. Additionally, hyperlinks are used to link phrases to articles about other related concepts.

Wikipedia articles were parsed to identify the title, and the leading text paragraph. All bold face, italicized, and hyperlinked phrases in the lead paragraph were identified as candidates. The common linking patterns were applied to extract pairs of linked candidate entities. For instance, applying this technique to the example above would generate two pairs: ("xerostomia", "dry mouth") and ("xerostomia", "dry mouth syndrome").

Labeling terms as consumer or professional

Although Wikipedia articles mention the alternate variants for a term, they do not specify which terms are likely to be used by consumers and which ones by professionals. Hence, simply extracting these relationships directly from Wikipedia does not provide enough detail to build a CHV, or to map between consumer and professional terms because there is no label assigned to each term.

It can be difficult to categorize terms as belonging to professional or consumer vocabulary. Frequently, health terms used by professionals migrate or evolve into popular vernacular as they become well known.[51] The acceptance or preference of a term can be measured, however, based on how often the term is used by the community. Specifically, if a term is more prominently used in professional text that is generated by and primarily intended for medical professionals, it can be regarded as professional. Conversely, if a term is frequently used by laypersons but not as often by medical professionals, then the term is more likely to be in a consumer-preferred vocabulary.

To quantitatively measure the propensity of a term T to be a consumer-oriented term, we define the following measure:

CHV_propensity (T) = count(T occurs in a consumer text corpus) / (size of the consumer text corpus)

Similarly, we can measure the propensity of a term T to be a professional term, as

PROF_propensity (T) = count(T occurs in a professional text corpus) / (size of the professional text corpus)

We propose that a term is more likely to be a consumer-oriented term if and only if its CHV_propensity is higher than its PROF_propensity. We refer to this intuition as the *propensity argument*.

To label a pair of terms (A, B) as professional or consumer with respect to one another, we can extend the propensity argument to the following. A is said to be the professional term and B is said to be the consumer term, if

$$\frac{\text{PROF\_propensity (A)}}{\text{CHV\_propensity (A)}} > \frac{\text{PROF\_propensity (B)}}{\text{CHV\_propensity (B)}}$$

or alternately,

$$\frac{\text{PROF\_propensity (A)}}{\text{PROF\_propensity (B)}} > \frac{\text{CHV\_propensity (A)}}{\text{CHV\_propensity (B)}}$$

which is the same as saying

$$\frac{\text{count(A occurrences in professional text)}}{\text{count(B occurrences in professional text)}} > \frac{\text{count(A occurrences in consumer text)}}{\text{count(B occurrences in consumer text)}} \quad \text{(Eq. 1)}$$

Conversely, if the condition is not met, then A is said to be the consumer term and B is said to be the professional term. Note that when comparing two terms using Eq. 1, the relative sizes of the corpora do not matter. However, the statistics collected over large corpora are more robust. Even when text corpus sizes are large, some concepts might occur infrequently or not appear at all. Such cases are avoided by smoothing the counts using Laplace smoothing.[52]

We chose online health discussion forums as a representative of consumer language. We crawled all the questions and comments posted by members on community discussion forums on MedHelp.[13] MedHelp is one of the earliest and well-known online forums dedicated to supporting user-driven discussions on health or healthcare related topics. The dataset consists of approximately thirty million messages posted by about a million unique users, and contains approximately 450 million words. This dataset has been a subject of study in other research endeavors.[53] In the following analysis, we refer to this as the **consumer text corpus**.

As a representative of professional text, we chose the abstracts of articles published in scientific journals and included in the 2012 MEDLINE®/PubMed® Baseline distribution.[12] To create a comparable corpus (in terms of word counts) to the consumer text corpus described earlier, we processed two million citations from the Baseline distribution that corresponded to papers published between 2008 and 2012. Titles were excluded from the generated professional text corpus. In the following analysis, we refer to this corpus as the **professional text corpus**.

To evaluate the accuracy of labeling professional and consumer terms in the extracted pairs, a medical expert conducted a manual review. First, the extracted pairs were filtered such that both terms appeared at least five times in both professional and consumer text corpora. A sample of 100 pairs was then randomly selected from this filtered set and manually judged and coded by a medical expert as one of the following classes: (a) valid pairing with correct labeling, (b) valid pairing with incorrect labeling, (c) pairs of equivalent concepts that either have alternative spellings or are synonymous, (d) pairs of related items, but not in a professional-consumer setting, such as an "is-a" relationship, or (e) invalid pairings. The pairs coded as equivalent or related (classes (c) or (d) above) were further coded to check if they were spelling variants, in case of equivalent pairs, or had a hierarchical ("is-a") relationship, if they were initially coded as related. The analysis of the expert judgment is presented in the Results section.

Results

Extracting pairs of medical concepts and their alternate names

Applying these techniques over the filtered set of medical articles from Wikipedia, we obtained 2,721 pairs of concepts and their consumer-preferred alternate names. Table 2 lists the linking patterns used, along with number of concept pairs each pattern generated. We also list a few examples of pairs extracted using that pattern.

Table 2. Patterns used to find pairs of alternate names, along with the count of pairs extracted and a few examples.

Linking pattern	Count	Examples
also known as	1695	(hematocrit, packed cell volume); (hair removal, epilation); (leukopenia, leukocytopenia); (dentures, false teeth)
also called	604	(heat therapy, thermotherapy); (hypersalivation, ptyalism); (nephroptosis, floating kidney); (dark therapy, scototherapy)
commonly known as	157	(nitrous oxide, laughing gas); (calcium oxide, quicklime); (pleurothotonus, Pisa syndrome); (*nepeta cataria*, catnip)
also termed / referred to as	106	(vertebral osteomyelitis, spondylodiskitis); (periapical cyst, radicular cyst); (red blood cells, erythrocyte); (posterior ramus syndrome, Maigne syndrome)
commonly called / referred to as	61	(peripheral vascular disease, peripheral artery disease); (actaea, baneberry); (schizophasia, word salad); (unnecessary health care, overtreatment)
sometimes called / termed	45	(high blood pressure, arterial hypertension); (chalicosis, flint disease); (hemiballismus, ballism); (irritable male syndrome, Del syndrome)
sometimes known as / referred to as	33	(pentazonia, giant pill millipedes); (hypochondria, health phobia); (ocular dominance, eyedness); (sexual addiction, sex addiction)
previously known as / called / referred to as	14	(acute kidney injury, acute renal failure); (erythrovirus, parvovirus B19); (periodic limb movement disorder, nocturnal myoclonus); (ankylosing spondylitis, Bechterew's disease)
colloquially known as / called / referred to as	6	(halitosis, bad breadth); (coal workers pneumoconiosis, black lung disease); (asystole, flatline); (central facial palsy, central seven)
Total pairs	**2721**	

We observe that a majority of extracted pairs come from the pattern "also known as". Further, 90% of the extracted pairs come from the top three patterns. We also observe that for some patterns such as "commonly known as" or "colloquially known as", the consumer-preferred terminology usually occurs as the second part of the extracted pair. However, these patterns contribute only about 8% of the extracted pairs. In pairs extracted using other patterns, the consumer term could appear in either positions.

Labeling the extracted pairs

To evaluate the accuracy of labeling the terms in each extracted pair as "professional" or "consumer" terms, a medical expert reviewed a sample of 100 pairs, as described in the Methods section. Each pair was manually coded by the expert as one of the following classes: (a) valid pairing with correct labeling, (b) valid pairing with incorrect labeling, (c) pairs of equivalent concepts that either have alternative spellings or are synonymous, (d) pairs of related items, but not in a professional-consumer setting, such as an "is-a" relationship, or (e) invalid pairings. Table 3 summarizes the results.

Table 3. Classification of 100 pairs of extractions, with examples. In the examples provided, the professional term as defined by the expert reviewer is shown first and the consumer term is shown second.

Code	Class of instance	Counts	Example pairs
(a.0)	Valid pairing with correct labeling	54	(icterus, jaundice); (pyrosis, heartburn); (oral candidiasis, oral thrush); (somnambulism, sleepwalking); (tinea, ringworm)
(b.0)	Valid pairing with incorrect labeling	4	(chronic renal disease, chronic kidney disease); (ovum, eggs); (brucellosis, Mediterranean fever); (hyperplasia, proliferation)
(c.1)	Alternative spelling variants	8	(leukoplakia, leucoplakia); (fecal incontinence, faecal incontinence); (post-concussion syndrome, postconcussive syndrome)
(c.2)	Synonyms or equivalent	23	(orthostatic hypotension, postural hypotension); (viral load, viral titer); (Lugols iodine, Lugols solution); (mouthwash, mouth rinse); (fecal incontinence, anal incontinence);
(d.1)	Is-a mapping	2	(cannabinoid, endocannabinoid); (radiology, radiation oncology)
(d.2)	Related concepts	9	(hypersensitivity, intolerance); (medical test, diagnostic technique); (medical procedure, technique); (pain management, pain medicine); (keratosis, keratotic); (suffering, aversive); (nerve, innervation); (tracheotomy, tracheostomy); (coccidioidomycosis, cocci)

From Table 3, we first observe that 89% of the pairs are between synonymous or equivalent concepts (Table 3, rows (a.0), (b.0), (c.1), and (c.2)), while the remaining instances were mainly between related items that are not an exact synonym of one another. In the current study, since the focus is on identifying medical terms and their equivalent consumer terms, related concepts (Table 3, rows (d.1) and (d.2)) are less desirable than the valid mappings. None of the 100 extracted pairs were judged invalid or to be between unrelated medical concepts.

Among the 89 pairs identified as valid pairs, 58 pairs (65.2%) were judged to be valid mappings between a professional term and a consumer term. The remaining pairs were either synonyms or equivalent concepts, both of which are valid professional terms. It is instructive to note that even when two terms are profession terms, one is often more widely accepted and used in consumer-generated corpora. For example, the terms "orthostatic hypertension" and "postural hypotension" appeared with similar frequencies in the professional text corpus, but the latter variant was observed 7.5 times more frequently than the former in the consumer text corpus. Hence, our approach labeled "postural hypotension" as the consumer term in the above example pair.

The approach to label terms as consumer or professional was also fairly accurate. For four pairs, the expert labels mismatched those assigned automatically (using the propensity argument, Eq. 1). Comparing against the 54 instances that were judged and labeled the same, this leads to a labeling accuracy of 93.1%. All four instances have been listed as examples in the table (Table 3, row (b.0)).

Identifying new mappings between professional and consumer terms

Finally, we also compared the pairs obtained by the proposed method against the CHV files made available by the open access and collaborative (OAC) CHV initiative.[54] Our approach generated many new pairs that were not included in the existing databases of consumer health vocabulary. Table 4 lists a few such examples of pairs extracted from Wikipedia articles.

Table 4. Examples of new pairs of professional and consumer terms that are not included in available CHV datasets.

Professional term	Equivalent consumer term
ambulatory surgery	outpatient surgery
arterial hypertension	high blood pressure
asphyxiation	lack of oxygen
Biofilms	Plaque
conjunctivitis	pink eye
dermatophytosis	Ringworm
Ethanol	drinking alcohol*
Ethanol	pure alcohol
Fatigue	Lethargy
fertilization	Conception

Professional term	Equivalent consumer term
immunoglobulins	Antibodies
nasopharyngitis	common cold
neuroleptics	Antipsychotics
nutrition	Nourishment
orthostatic hypotension	postural hypotension
periodontitis	gum disease
rofecoxib	Vioxx
social care	home care
stuttering	Stammering
uncompensated care	charity care

* The CHV mentions the pairing between ethanol and drinking alcohol, but lists it as an incorrect mapping instance.

Discussion

One of the benefits of using Wikipedia as a source to identify alternate names is that popular variant names are often nominated by the community and thus mentioned in the Wikipedia article about the primary concept. As language itself evolves, new variants might be introduced,[51] and existing variants might change in popularity. Regularly updating the CHV using the latest snapshot of Wikipedia articles provides one way to keep them current and relevant. Since errors in Wikipedia articles often get corrected by other editors, periodic update of extracted pairs can therefore help eliminate erroneous instances of extracted pairs. For instance, we observed that the latest version of Wikipedia (updated after our analyses were conducted) correctly removed the mention of "radiation oncology" as a variant form for "radiology" (see Table 3, row (d.1)). In the same vein, the professional and consumer text corpora could also be frequently updated to monitor the usage of such terms in the respective communities.

The approach of using free text to compute statistics is subject to the entity disambiguation problems. For example, the term "cocci" has multiple meanings – it could be used either as a shortened term for the disease "coccidioidomycosis", or as the plural form of "coccus". Failure to disambiguate between these two concepts could lead to mislabeling. Although this is a limitation of most statistical and shallow, frequency-based approaches, such occurrences are relatively infrequent.

We have not measured the recall of such pattern-based approaches, to extract *all* instances of consumer terms or alternate variants. Further study is also needed to understand if the words used in the patterns could improve the labeling of terms as professional or consumer. For example, if the pattern is "commonly known as" or "colloquially called", it is more likely that the succeeding concept is a consumer term.

Concerns have been raised about use of Wikipedia as a resource of information in the scientific literature. Bibliometric analysis has shown an increased rate of citing Wikipedia articles in peer-reviewed health science journal publications in recent years,[55,56] and studies have found several limitations with respect to depth of discussion and readability in Wikipedia articles.[57,58] Efforts are underway to encourage medical professionals to actively contribute to Wikipedia,[59] and, in collaboration with other medical and healthcare experts and medical journals, to improve the overall quality of medical articles in Wikipedia.[60,61] Although such concerns are important issues to address in the future, they are beyond the current scope of research presented in this paper.

Conclusion

In this study, we demonstrated the effectiveness of a novel approach that uses community-generated corpora such as Wikipedia to mine pairs of professional terms and their equivalent consumer terms. We measured the propensity of

a term to be a consumer term based on its relative frequencies appearing in the consumer or professional contexts, and demonstrated how this information could be used to properly label the terms. The empirical evaluation results are promising, suggesting that the proposed approach is able to identify and differentiate consumer and professional terms from the Wikipedia corpus with high accuracy. The methods proposed in this paper can therefore be used to augment, update, and maintain existing consumer health vocabularies to enhance the performance of computational applications designed to improve the readability, parsing, and understandability of community-generated health text.

Acknowledgements

The authors would like to acknowledge the contribution of Nirav Mehta, who helped collect data for this study. This study was supported in part by the University of Michigan MCubed Program, the National Center for Advancing Translational Sciences under Award Number UL1TR000433, the National Science Foundation under grant numbers IIS-1054199 and CCF-1048168, and the DARPA under award number W911NF-12-1-0037. The content is solely the responsibility of the authors and does not necessarily represent the official views of funding agencies.

References

1. Fox S, Duggan M. Health online 2013. Pew Research Center's Internet & American Life Project. Published 2013 Jan 15.
2. Zeng QT, Tse T. Exploring and developing consumer health vocabularies. J Am Med Inform Assoc 2006;13:24–9.
3. Okun S, McGraw D, Stang P, et al. Making the case for continuous learning from routinely collected data. Institute of Medicine Discussion Paper, National Academy of Sciences. 2013.
4. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature 2009;457:1012–4.
5. Frost J, Okun S, Vaughan T, Heywood J, Wicks P. Patient-reported outcomes as a source of evidence in off-label prescribing: analysis of data from PatientsLikeMe. J Med Internet Res 2011;13(1):e6.
6. Pearson JF, Brownstein CA, Brownstein JS. Potential for electronic health records and online social networking to redefine medical research. Clin Chem 2011;57(2):196–204.
7. Bian J, Topaloglu U, Yu F. Towards large-scale Twitter mining for drug-related adverse events. Proc Workshop on Smart Health and Wellbeing 2012;25–32.
8. Riding the information technology wave in life sciences: priorities, pitfalls and promise. Retrieved 2014 Mar 12 http://www.imshealth.com/portal/site/imshealth/menuitem.762a961826aad98f53c753c71ad8c22a/?vgnextoid=743a7a4c18394410VgnVCM10000076192ca2RCRD.
9. MacLean DL, Heer J. Identifying medical terms in patient-authored text: a crowdsourcing-based approach. J Am Med Inform Assoc 2013;0:1–8.
10. Delbanco T, Walker J, Darer JD, et al. Open notes: doctors and patients signing on. Ann Intern Med 2010;15(2):121–5.
11. Wi /. kipedia is a massively popular (yet untested) doctor. Retrieved 2014 Feb 22 from http://m.nextgov.com/health/2014/02/wikipedia-massively-popular-yet-untested-doctor/79154/.
12. 2012 MEDLINE®/PubMed® Baseline Database Distribution. Retrieved 2013 Nov 23 from http://www.nlm.nih.gov/bsd/licensee/2012_stats/baseline_med_filecount.html.
13. MedHelp. Retrieved 2013 Mar 20 from http://www.medhelp.org/.
14. Eade D. Dr Wikipedia will see you now… Insight Research Group. 2011 Jun 7. Retrieved 2014 Mar 8 from http://www.pmlive.com/pharma_news/dr_wikipedia_will_see_you_now..._280528.
15. Rosen D. Engaging patients through social media. Report by the IMS Institute for Healthcare Informatics. Published 2014 Jan.
16. Comer B. Docs look to Wikipedia for condition info: Manhattan research. Medical Marketing & Media. 2009 Apr 21. Retrieved 2014 Mar 11 from http://www.mmm-online.com/docs-look-to-wikipedia-for-condition-info-manhattan-research/article/131038/.
17. Hughes B, Joshi I, Lemonde H, Wareham J. Junior physician's use of Web 2.0 for information seeking and medical education: a qualitative study. Int J Med Inform 2009;78:645–55.
18. Brokowski I, Sheehan AH. Evaluation of pharmacist use and perception of Wikipedia as a drug information resource. Ann Pharmacother 2009;43:1912–3.

19. Masters K. For what purpose and reasons do doctors use the Internet: a systematic review. Int J Med Inform 2008;77:4–16.
20. Burgos C, Bot A, Ring D. Evaluating the effectiveness of a wiki Internet site for medical topics. J Hand Microsurg 2012;4:21–4.
21. Varga-Atkins T, Dangerfield P, Brigden D. Developing professionalism through the use of wikis: a study with first-year undergraduate medical students. Med Teach 2010;32:824–9.
22. Haigh CA. Wikipedia as an evidence source for nursing and healthcare students. Nurse Educ Today 2011;31:135–9.
23. Jalali A, Mioduszewski M, Gauthier M, Varpio L. Wiki use and challenges in undergraduate medical education. Med Educ 2009;43:1117.
24. Snodgrass S. Wiki activities in blended learning for health professional students: enhancing critical thinking and clinical reasoning skills. Aus J Educ Technol 2011;27:563–80.
25. Weiner SA, Stephens G, Nour AY. Information-seeking behaviors of first-semester veterinary students: a preliminary report. J Vet Med Educ 2011;38:21–32.
26. Eysenbach G, Powell J, Kuss O, Sa ER. Empirical studies assessing the quality of health information for consumers on the World Wide Web: a systematic review. JAMA 2002;287:2691–700.
27. Deshpande A, Jadad AR. Web 2.0: Could it help move the health system into the 21st century? J Men Health Gender 2006;3:332–6.
28. Thomas GR, Eng L, de Wolff JF, Grover SC. An evaluation of Wikipedia as a resource for patient education in nephrology. Semin Dial 2013;26:159–63.
29. Kinnane NA, Milne DJ. The role of the Internet in supporting and informing carers of people with cancer: a literature review. Support Care Cancer 2010;18:1123–36.
30. Laurent MR, Vickers TJ. Seeking health information online: does Wikipedia matter? J Am Med Inform Assoc. 2009;16:471–9.
31. Hearst MA. Automatic acquisition of hyponyms from large text corpora. Proc Conf Comput Linguist (COLING) Assoc Comput Linguist 1992;539–45.
32. Snow R, Jurafsky D, Ng AY. Learning syntactic patterns for automatic hypernym discovery. Adv Neural Inf Process Syst 2004;17:1297–304.
33. Etzioni O, Cafarella M, Downey D, et al. Unsupervised named-entity extraction from the web: an experimental study. Artif Intell 2005;165(1):91–134.
34. Girju R, Badulescu A, Moldovan D. Automatic discovery of part-whole relations. Comput Linguist Assoc Comput Linguist 2006;32(1):83–135.
35. Yates A, Etzioni O. Unsupervised Methods for Determining Object and Relation Synonyms on the Web. J Artif Intell Res 2009;34:255–96.
36. Aronson AR, Lang FM. An overview of MetaMap: historic perspectives and recent advances. J Am Med Inform Assoc 2010;17(3):229–36.
37. Torii M, Hu Z, Wu CH, Liu H. BioTagger-GM: a gene/protein name recognition system. J Am Med Inform Assoc 2009;16(2):247–55.
38. Harkema H, Gaizauskas R, Hepple M, et al. A large scale terminology resource for biomedical text processing. Proc Workshop Linking Biological Literature Ontologies and Databases Assoc Comput Linguist 2004:53–60.
39. Kageura K, Umino B. Methods of automatic term recognition: a review. Terminology 1996;3(2):259–89.
40. Krauthammer M, Nenadic G. Term identification in the biomedical literatue. J Biomed Inform 2004;37(6):512–26.
41. Collier N, Nobata C, Tsujii J. Extraction of the names of genes and gene products with a hidden Markov model. Proc Conf Comput Linguist (COLING) Assoc Comput Linguist 2000:201–7.
42. de Bruijn B, Cherry C, Kiritchenko S, Martin J, Zhu X. Machine learned solutions for three stages of clinical information extraction: the state-of-the-art at i2b2 2010. J Am Med Inform Assoc 2011;18(5):557–62.
43. Jindal P., Roth D. Using soft constraints in joint inference for clinical concept recognition. Proc Conf Empirical Methods in Natural Language Processing (EMNLP) Assoc Comput Linguist 2013:1808–14.
44. Frantzi KT, Ananiadou S, Mima H. Automatic term recognition of multi-word terms: the C-value/NC-value method. International Journal on Digital Libraries 2003;3(2):115–30.
45. Zeng QT, Tse T, Divita G, et al. Term identification methods for consumer health vocabulary development. J Med Internet Res 2007;9(1):e4.
46. Wikipedia dump of English language articles. Retrieved 2013 Sep 15 from http://dumps.wikimedia.org/enwiki/.
47. Wikipedia: main topic classifications. Retrieved 2013 Sep 15 from http://en.wikipedia.org/wiki/Category:Main_topic_classifications.

48. CatScan. Retrieved 2013 Oct 12 from http://tools.wmflabs.org/catscan2/catscan2.php.
49. Wikipedia: manual of style / lead section. Retrieved 2014 Mar 9 from http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Lead_section.
50. Wikipedia article on Xerostomia. Retrieved 2014 Mar 11 from http://en.wikipedia.org/wiki/Xerostomia.
51. Doing-Harris KM, Zeng-Treitler, Q. Computer-assisted update of a consumer health vocabulary through mining of social network data. J Med Internet Res 2011;13(2):e3.
52. Manning CD, Raghavan P, Schütze M. Introduction to Information Retrieval. Cambridge University Press 2008:240.
53. Vydiswaran VGV, Liu Y, Mei Q, Zheng K, Hanauer D. User-created groups in health forums: What makes them special? Proc Conf Weblogs and Social Media (ICWSM) Assoc Adv Artif Intell 2014:515–24.
54. Open Access and Collabolative Consumer Health Vocabulary Initiative. Retrieved 2013 Nov 10 from http://consumerhealthvocab.org/.
55. Bould MD, Hladkowicz ES, Pigford AA, et al. References that anyone can edit: review of Wikipedia citations in peer reviewed health science literature. BMJ 2014;348:g1585.
56. Noruzi A. Editorial: Wikipedia popularity from a citation analysis point of view. Webology 2009 Jun;6(2):e20.
57. Giles J. Internet encyclopaedias go head to head. Nature 2005;438:900–1.
58. Azer SA. Evaluation of gastroenterology and hepatology articles on Wikipedia: are they suitable as learning resources for medical students? Eur J Gastroenterol Hepatol 2014 Feb;26(2):155–63.
59. Metcalfe D, Powell J. Should doctors spurn Wikipedia? J R Soc Med 2011;104:488–9.
60. Heilman JM, Kemmann E, Bonert M, et al. Wikipedia: a key tool for global public health promotion. J Med Internet Res 2011;13:e14.
61. Mathew M, Joseph A, Heilman J, Tharyan P. Cochrane and Wikipedia: the collaborative potential for a quantum leap in the dissemination and uptake of trusted evidence. Cochrane Database Syst Rev 2013 Oct 22;10:ED000069.

Adverse Drug Event-based Stratification of Tumor Mutations: A Case Study of Breast Cancer Patients Receiving Aromatase Inhibitors

Chen Wang, PhD, Michael T. Zimmermann, PhD, Naresh Prodduturi, Christopher G. Chute, MD, Dr.PH, Guoqian Jiang, MD, PhD
Department of Health Sciences Research, Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN

Abstract

Adverse drug events (ADEs) are a critical factor for selecting cancer therapy options. The underlying molecular mechanisms of ADEs associated with cancer therapy drugs may overlap with their antineoplastic mechanisms; an aspect of toxicity. In the present study, we develop a novel knowledge-driven approach that provides an ADE-based stratification (ADEStrata) of tumor mutations. We demonstrate clinical utility of the ADEStrata approach through performing a case study of breast invasive carcinoma (BRCA) patients receiving aromatase inhibitors (AI) from The Cancer Genome Atlas (TCGA) (n=212), focusing on the musculoskeletal adverse events (MS-AEs). We prioritized somatic variants in a manner that is guided by MS-AEs codified as 6 Human Phenotype Ontology (HPO) terms. Pathway enrichment and hierarchical clustering of prioritized variants reveals clusters associated with overall survival. We demonstrated that the prediction of per-patient ADE propensity simultaneously identifies high-risk patients experiencing poor outcomes. In conclusion, the ADEStrata approach could produce clinically and biologically meaningful tumor subtypes that are potentially predictive of the drug response to the cancer therapy drugs.

1 Introduction

Adverse drug events (ADEs) have been well recognized as a cause of patient morbidity and increased health care costs in the United States. With rapid developments in genomics technology, the contribution of genetic factors to ADEs is being considered and has already influenced clinical recommendations for drug dosage and toxicity (1, 2), thus representing a major component of the movement to pharmacogenomics and individualized medicine (3, 4). Genetic susceptibility is an important feature of severe ADEs and there is considerable interest in developing genetic tests to identify at-risk patients prior to prescription (5). Preliminary studies also suggested that drug therapies based on an individual's genetic makeup may result in a significant reduction in adverse outcomes (6).

To conduct a pharmacogenomics study of an ADE, ideally, multiple sources of evidence should be integrated to fully characterize the potential pharmacogenomics mechanism relevant to the ADE. For instance, a project known as PharmGKB (7, 8), initiated by the National Institute of Health (NIH), has a mission of collecting and disseminating human-curated information about the impact of human genetic variation on drug responses. In our previous studies, we proposed a knowledge-driven framework that aims to support pharmacogenomics-target prediction of ADEs (9). In the framework, we integrated a semantically annotated literature corpus, Semantic MEDLINE, with a semantically coded ADE knowledge base known as ADEpedia (10) using a Semantic Web-based framework. We developed a knowledge-discovery approach leveraging a network-based analysis of a protein-protein interaction (PPI) network to mine the knowledge of drug-ADE-gene interactions.

The recent advances in sequencing technology have underpinned the progress in several large-scale projects to systematically compile genomic informatics related to human cancer (11, 12). A notable example is The Cancer Genome Atlas (TCGA) (13) and projects that have focused on identifying links between cancer and genomic variation. More promisingly, TCGA Pan-Cancer Project (14) has been initiated to assemble coherent datasets across tumor types, analyze the data in a consistent fashion, and finally provide comprehensive interpretation. Tumor stratification has been regarded as one of the fundamental goals of cancer informatics, enabling Pan-Cancer studies in which the molecular profiles of tumors are used to determine subtypes (15), regardless of the organ in which it is manifest. In particular, the somatic mutation profile is emerging as a rich new source of data for uncovering tumor subtypes with different causes and clinical outcomes. A network-based stratification using the knowledge of molecular signaling could produce robust tumor subtypes that are biologically informative and have a strong association to clinical outcomes and emergence of drug resistance (15).

Preliminary studies have demonstrated that the underlying molecular mechanism of common ADEs known to cancer therapy drugs may overlap with that of the efficacy of the therapeutic drugs themselves. For example, breast cancer patients receiving aromatase inhibitors (AI) have a high incidence of musculoskeletal adverse events (MS-

AEs); about half of patients treated with AIs have joint-related complaints (16, 17). Musculoskeletal complaints have been the most frequent reason given by patients on a clinical trial comparing the non-steroidal AI anastrozole with the steroidal AI exemestane as adjuvant therapy for early breast cancer (18). A case-control genome-wide association study (GWAS) from a Mayo Clinic group identified SNPs associated with MS-AEs in women treated with AIs, one of which created an estrogen response element (18). Another study in the same group at Mayo Clinic confirmed that single nucleotide polymorphisms (SNPs) in the aromatase CYP19 gene contribute to response to neoadjuvant AI therapy (19), two of which are significantly associated with both a greater change in aromatase activity after AI treatment and higher plasma estradiol levels pre- and post-AI treatment.

The objective of the present study is to develop a novel knowledge-driven approach that provides an ADE-based stratification of tumor mutations (ADEStrata). Our assumption here is that the ADE-based tumor stratification would potentially produce clinically and biologically meaningful tumor subtypes that are predictive of the drug response to the cancer therapy drugs. To test the assumption, we performed a case study of breast cancer patients receiving the AIs from TCGA. We utilized a variant prioritization tool upon the somatic mutation profiles of TCGA breast invasive carcinoma (BRCA) patients treated with three AI drugs. The phenotype input of the variant prioritization tools contains a set of MS-AEs represented by standard Human Phentoype Ontology (HPO) terms. We utilized the prioritized variants to cluster the target patients into subgroups and investigated their associations with clinical outcomes.

2 Materials and Methods

2.1 Materials

2.1.1 SIDER: A Side Effect Resource

The SIDER (SIDe Effect Resource) is a public, computer-readable side effect resource that contains reported adverse drug reactions(20). The information is extracted from public documents and package inserts; in particular, from the United States Food and Drug Administration (FDA) Structured Product Labels (SPLs). The standardized Coding Symbols for a Thesaurus of Adverse Reaction Terms (COSTART), which are part of the Unified Medical Language System (UMLS) Metathesaurus, were used as the basic lexicon of side effects. In the present study, we utilized the latest version SIDER 2 that was released on October 17, 2012.

2.1.2 HPO: Human Phenotype Ontology

The HPO project aims to provide a standardized vocabulary of phenotypic abnormalities encountered in human diseases (21). The HPO is being developed in collaboration with the OBO Foundry using information from Online Mendelian Inheritance in Man (OMIM) and medical literatures. The ontology contains more than 10,000 terms and equivalence mappings to other standard vocabularies such as MedDRA and UMLS. In the present study, we used the latest version of HPO-MedDRA mapping file that is publicly available from the HPO website (22).

2.1.3 eXtasy: A Variant Prioritization Tool

eXtasy is a pipeline developed at the University of Leuven, for ranking the likelihood that a given nonsynonymous single nucleotide variants (nSNVs) is related to given phenotype (23, 24). The pipeline utilizes a genomic data fusion methodology (25) that takes into account multiple strategies to detect the deleteriousness of mutations and prioritizes them in a phenotype-specific manner. The ultimate goal of the tool is to discriminate between putatively mildly deleterious rare variants and actual disease-causing variants. The eXtasy tool is open-source and publicly downloadable from its github site (26). The eXtasy pipeline takes a Variant Call File (VCF) and one or more gene prioritization files. Each prioritization file is pre-computed for a specific phenotype (HPO term). In the present study, we downloaded and installed the tool on a local Ubuntu server.

2.1.4 TCGA Data Portal

TCGA Data Portal provides a platform for researchers to search, download, and analyze data sets generated by TCGA (27). It contains clinical information, genomic characterization data, and high level sequence analysis of the tumor genomes. There are two data access tiers: Open Access data tier and Controlled Access. The data of more 30 tumor types are available from the data portal. As of February 2014, there are 1043 cases of breast invasive carcinoma (BRCA) with data. In the present study, we utilized the BRCA clinical data (including clinical drug data and follow-up data) and somatic mutation data through the Open Access data tier.

2.2 Methods

2.2.1 Identifying HPO ADE Terms Relevant to Aromatase Inhibitors

In the present study, we aim to conduct variant prioritization in a manner that can be guided by the ADEs relevant to AI drugs. In other words, the ADEs induced by AI drugs are treated as the phenotypes required by the eXtasy tool. In order to enable the use of the eXtasy tool for variant prioritization in a phenotype-specific manner, we need to identify the phenotypes that are represented in HPO terms.

We first mapped the ADE terms represented in MedDRA UMLS concept unique identifiers (CUIs) from the SIDER 2 database file to the HPO terms using an HPO-MedDRA mapping file produced by HPO development team. Second, we annotated those HPO terms with a flag using the eXtasy HPO term list to indicate whether a HPO-based ADE term can be processed by eXtasy or not. Third, we retrieved those entries (with drug-ADE pairs) using the drug names “anastrozole”, “exemestane” and “letrozole” which are the third-generation AI drugs. We reviewed all the ADE terms and identified those ADE terms belonging to musculoskeletal adverse events (MS-AEs) and their HPO term annotations.

2.2.2 Identifying Patient Cohorts by AI Drugs and Somatic Mutations from TCGA

We utilized the clinical drug file of the BRCA patients from TCGA data portal through its Open-Access HTTP Directory. The spelling corrections were taken for all variants of the three drugs to maximize the sample size of the patient cases. We then identified a set of patient cases (represented by patient barcodes) that were prescribed for the three AI drugs (AI cases).

We also downloaded the somatic mutation file of the BRCA patients from TCGA data portal in a Mutation Annotation Format (MAF). The format is a tab-delimited file containing somatic mutations for each patient. As eXtasy requires a VCF file as input, we converted the MAF file into a collection of VCF files. Each VCF file contains somatic mutations for a single patient tumor sample. We combined all VCF files for all AI cases into a single VCF file using the patient barcodes identified in the step above.

2.2.3 Variant Prioritization Using HPO ADE Terms

As mentioned above, we installed an instance of the eXtasy tool in a local server and ran the tool with a custom Ruby script. The input consists of a VCF file (produced in the Section 2.2.2) and a set of pre-computed gene prioritization files for those phenotypes represented by the HPO ADE terms of interest (identified in the Section 2.2.1). The output is a file with scores for the individual variants’ likelihood of impacting an individual HPO term. Order statistics (25) and aggregate scores are generated and range from 0 to 1, where 0 is likely to be disease-causing and 1 unlikely (in contrast with the normal eXtasy scores). This is a pseudo p-value that represents the probability that a variant is high-ranking in all different phenotypes given the null-hypothesis of random rankings. To understand how the variants affect function, we first classified the input variants into three functional impact categories, calling a variant “high” if it is a frameshift, nonsense, nonstop, or splice-site; and “medium” if it is a missense; and “silent” if it is a mutation not causing protein coding changes. And then we analyzed the function of those variants scored by eXtasy for AI-related HPO terms.

2.2.4 Tumor Mutation Stratification and Clinical Outcome Association Studies

We first selected statistically significant variants based on the eXtasy order statistics (pseudo p-value <0.05). Second, we aggregated genes affected by these prioritized variants across 1,320 canonical pathways collected from the Molecular Signature Database (MSigDB) (28, 29). In order to reduce false discoveries, multiple criteria were applied to further filter out less relevant pathways (binomial distribution p-value >0.05) or pathways containing too few genes (<10 genes). We excluded pathways with less than 10 genes, based on the consideration that small pathways are often subcomponents of larger pathways, and inclusion of them tends to introduce unnecessary redundancy. Third, we performed hierarchical clustering to highlight pathway-level patterns among AI-treated patients.

We used overall survival (OS) time (months) as a clinical endpoint to measure the outcome of TCGA patients in the identified cohort. We performed both univariate analysis and multivariate cox-regression to assess the association of clusters (produced by hierarchical clustering) with survival. In multivariate analysis, patient age and tumor stage were adjusted for to evaluate the independent contribution of each cluster. We also analyzed the distribution of patient age and tumor stage in the clusters identified.

3 Results

Out of 4,492 unique MedDRA terms represented by the UMLS CUIs from the SIDER database file, 2,827 (62.9%) MedDRA terms had mappings to 1,491 unique HPO terms. Out of the 1,491 HPO terms, 844 (56.6%) HPO terms are included in the eXtasy phenotype list. We identified 6 unique HPO terms representing the MS-AEs relevant to three AI drugs. The 6 HPO terms are *HP:0003418|Back pain, HP:0002653|Bone pain, HP:0003011|Abnormality of musculature, HP:0001369|Arthritis, HP:0009763|Limb pain, and HP:0002758|Osteoarthritis*. Table 1 shows the SIDER database entries with HPO terms identified for the musculoskeletal adverse events (MS-AEs) relevant to three AI drugs.

Table 1. Entries with HPO terms identified for the musculoskeletal adverse events (MS-AEs) relevant to three AI drugs

Drug label	Meddra umls cui	Meddra label	HPO id	HPO label	HPO Term in eXtasy
anastrozole	C0004604	Back pain	HP:0003418	Back pain	YES
anastrozole	C0151825	Bone pain	HP:0002653	Bone pain	YES
anastrozole	C0026857	Musculoskeletal disorder	HP:0003011	Abnormality of musculature	YES
anastrozole	C0003864	Arthritis	HP:0001369	Arthritis	YES
exemestane	C0004604	Back pain	HP:0003418	Back pain	YES
exemestane	C0151825	Bone pain	HP:0002653	Bone pain	YES
exemestane	C0030196	Pain in extremity	HP:0009763	Limb pain	YES
exemestane	C0026857	Musculoskeletal disorder	HP:0003011	Abnormality of musculature	YES
exemestane	C0029408	Osteoarthritis	HP:0002758	Osteoarthritis	YES
letrozole	C0004604	Back pain	HP:0003418	Back pain	YES
letrozole	C0026857	Musculoskeletal disorder	HP:0003011	Abnormality of musculature	YES
letrozole	C0151825	Bone pain	HP:0002653	Bone pain	YES
letrozole	C0030196	Pain in extremity	HP:0009763	Limb pain	YES
letrozole	C0003864	Arthritis	HP:0001369	Arthritis	YES

Using the clinical drug file of TCGA BRCA patients, we identified a cohort of 212 patients who were prescribed with one of the three AI drugs (AI cases).

The algorithm eXtasy ranks coding variants according to their probability of being related to a given phenotype. We found that 23.8% of the input variants are silent and are ignored by eXtasy, while 11.6% are of high impact (see section 2.2.3) and almost assuredly affect the normal physiologic function of the affected gene. Of the variants scored by eXtasy for AI-related HPO terms, 43% are highly conserved among placental mammals. Variants were prioritized for each patient across the MS-AE phenotypes represented by 6 HPO terms (listed in Table 1), producing aggregate prioritization scores (max and order statistics). Table 2 lists the top 20 prioritized variants.

Table 2. Top 20 variants prioritized for the MS-AE phenotypes using the eXtasy in the AI cases

Chromosome	Ref base	Alt base	Position	Gene region	eXtasy combined max	eXtasy combined order statistics
X	G	C	77289124	ATP7A	0.952	2.57E-13
10	C	G	89692883	PTEN	0.866	2.76E-13

12	G	A	115118782	TBX3	0.88	3.47E-13
17	G	A	7577094	TP53	0.852	1.28E-12
3	A	T	49455277	AMT	0.936	1.34E-12
9	G	T	132576329	TOR1A	0.958	1.91E-12
19	C	G	41838160	TGFB1	0.874	2.26E-12
10	A	T	8115746	GATA3	0.906	2.35E-12
5	C	T	174156285	MSX2	0.98	2.97E-12
12	A	G	115120669	TBX3	0.846	3.16E-12
17	C	T	7577547	TP53	0.892	3.70E-12
12	G	A	121431482	HNF1A	0.942	5.98E-12
17	A	C	7577144	TP53	0.834	6.21E-12
17	G	A	7577105	TP53	0.83	7.33E-12
7	C	G	5567503	ACTB	0.876	9.62E-12
17	A	G	7577129	TP53	0.832	1.07E-11
12	C	T	110781179	ATP2A2	0.908	1.92E-11
12	C	A	121426687	HNF1A	0.928	2.23E-11
3	G	A	30729932	TGFBR2	0.792	4.07E-11

From the eXtasy output for the AI cases, 2,164 statistically significant variants were selected for pathway enrichment and clustering analysis. Among 1,320 canonical pathways, 63 of them passed the filtering criteria defined in section 2.2.4. By hierarchical clustering, 3 distinct patient clusters, organized by pathways (affected by prioritized variants), were identified and are displayed in Figure 1 containing 91, 60, and 22 patients each. Patients in Cluster 1 exhibit relatively silent pathway aberrations, while Cluster 2 and Cluster 3 have much stronger pathway activities. A summary of the 63 selected pathways, enriched in MS-AE relevant variants, can be found in Supplemental Table 1 posted at http://catargets.org.

Table 3 shows the results of the univariate and multivariate cox-regression analysis for the three clusters. We found that although Cluster 3 has a relatively small number of patients, the cluster is significantly association with poorer survival time in both univariate and multivariate analysis. Table 4 shows the distribution of age and stage in the 3 clusters identified. There is no significant association between the 3 clusters and age/stage, although we noticed that Cluster 3 is enriched with more Stage 2 patient cases.

Figure 2 shows a Kaplan-Meier plot of survival time for the 3 clusters, derived from our pathway-level analysis. Interestingly, Cluster 2 does not have a significantly altered survival time, despite its similarity to Cluster 3. The exception is for a few activated pathways responsible for DNA damaging repair and apoptosis. In addition, we observed that many patients in Cluster 1 have somatic variants associated with ATM/thyroid pathways, while Cluster 2 has many other pathway features but lacks the ATM/thyroid pathway enrichment. Cluster 3, however, has the pathway features of both Cluster 1 and Cluster 2. This “two hits” pattern may account for the worse survival outcome associated with Cluster 3.

4 Discussion

In this study, we demonstrated that the ADE-based tumor stratification could produce clinically and biologically meaningful tumor subtypes that are potentially predictive of the drug response to the cancer therapy drugs. The preliminary results from the case study of TCGA breast cancer patients receiving AI drugs are very promising. We consider that our study approach and results have several important implications in terms of how to further understanding of disease, drug action or to improve treatment outcome. First, it would be possible that new patients can be assigned to different groups to inform clinical decision-making. Second, our approach could be potentially used in prediction of treatment outcomes or probability of ADEs based on tumor genome. Third, it would be possible that our study results could be used to gain a greater understanding of mechanism of action of targeted drugs or underlying causes of ADEs.

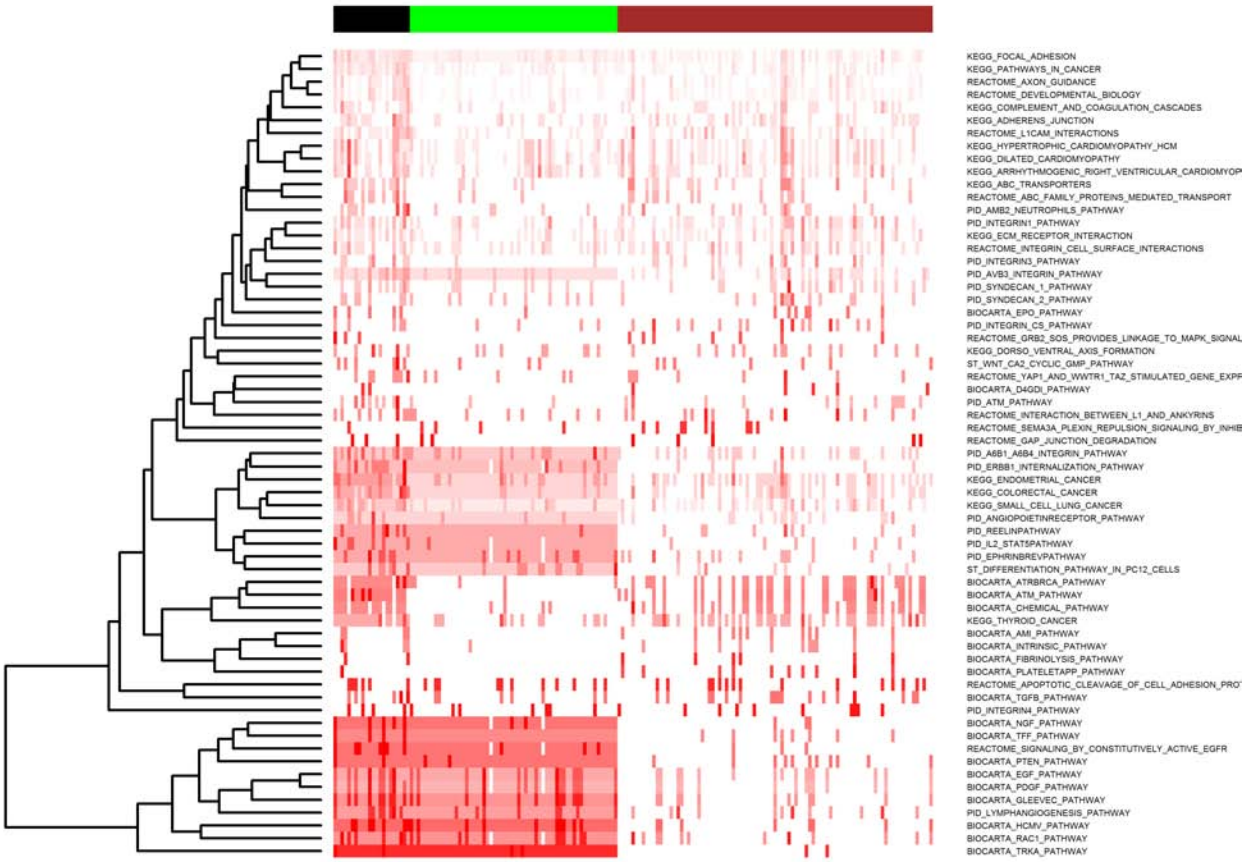


Figure 1. This ordered heatmap shows pathway-level clustering of 173 patients treated with AI across 63 pathways-enriched MS-AE relevant variants. The color of heatmap from white to red indicates low to high percentages (0% to 100%) of genes affected by MS-AE relevant variants. Column color-bar on top of the heatmap indicates three clusters of samples: Cluster 1 (brown), Cluster 2 (green) and Cluster 3 (black). Note that the number of the patients (n=173) with pathway enrichment is less than total number of the identified cohort (n=212) is because not all patients have prioritized variants listed.

Table 3. The univariate and multivariate cox-regression analysis results of cluster labels. In multivariate analysis, patient age and tumor stage were adjusted for to determine the independent contribution of cluster membership. * denotes p<0.05.

	Univariate analysis		***Multivariate analysis***	
	p-value	**HR [95% CI]**	**p-value**	**HR [95% CI]**
Cluster 2 (n=60) (Cluster 1 as ref)	0.63	0.57 [0.06, 5.55]	0.64	0.57 [0.06, 5.76]
Cluster 3 (n=22) (Cluster 1 as ref)	0.03*	5.03 [1.13, 22.55]	0.04*	4.86 [1.07, 22.161]
Overall model (log-rank test)	0.02*	NA	0.07	NA

Table 4. The disribution of age and stage in the 3 clusters identified. * p-value for age vs. cluster association was computed using ANOVA test; p-value for stage vs. cluster association was computed using Fisher's exact test.

	Cluster 1 (n=91)	Cluster 2 (n=60)	Cluster 3 (n=22)	p-value*
Age **mean [Q1, median, Q3]**	61 [55.2, 62.9, 66.1]	63.4 [54.9, 62.8, 71.9]	58.4 [54.1, 58.7, 62.8]	0.16
Stage **number (% per cluster)**				0.22
s1	32 (35.2%)	23 (38.3%)	5 (22.7%)	
s2	48 (52.7%)	27 (45%)	15 (68.2%)	
s3	11 (12.1%)	8 (13.3%)	1 (4.5%)	
s4	0 (0%)	2 (3.3%)	1 (4.5%)	

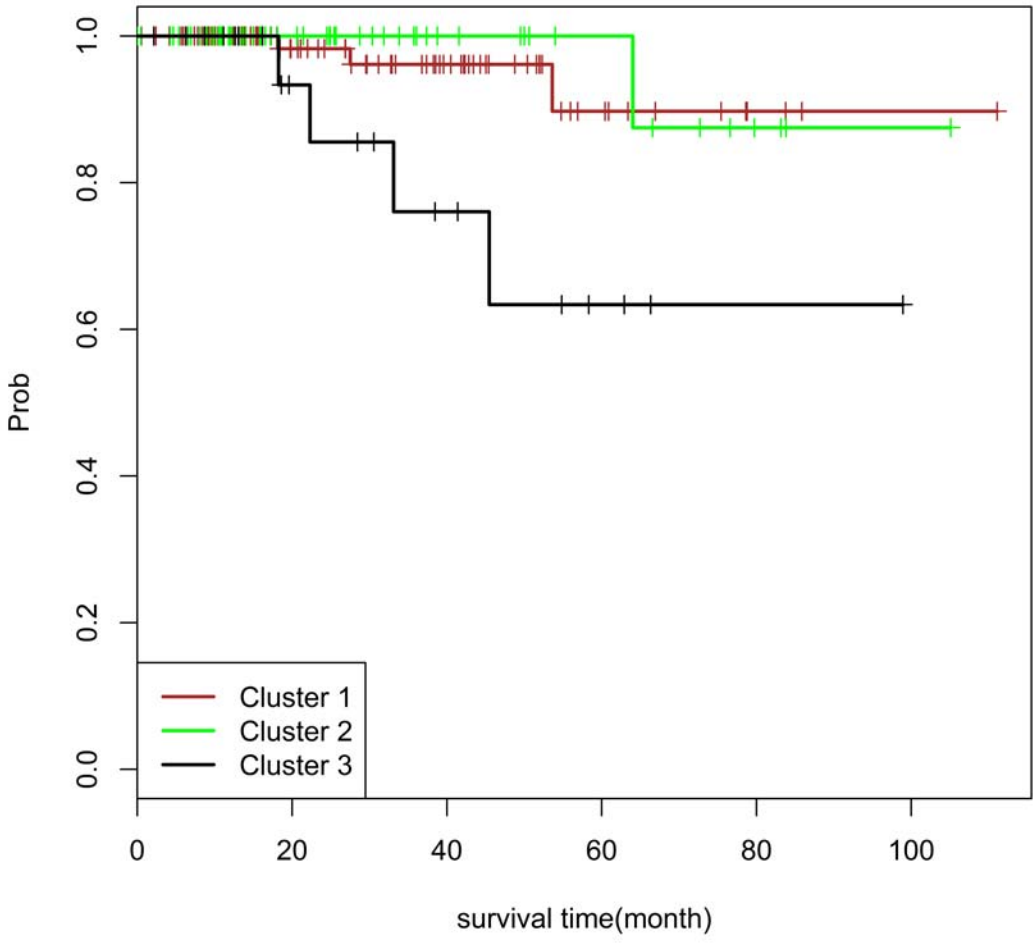


Figure 2. Kaplan-Meier plot of survival time for patients within the 3 pathway-pattern clusters .

One of the key components of our approach is to leverage known ADE knowledge for tumor stratification. We found that a semantically coded ADE knowledge base is extremely useful for extracting known ADEs relevant to target drugs. We utilized the SIDER ADE dataset, in which the ADEs are annotated with the MedDRA codes

represented by UMLS CUIs. In addition, the HPO team has produced a mapping file between MedDRA UMLS CUIs and HPO terms. These standard codes facilitated our ability to integrate and identify ADE terms across different datasets. For example, in the present study, eXtasy requires HPO terms as phenotype input and we were able to identify those MS-AE terms from the SIDER dataset represented in UMLS CUIs.

Use of standard phenotype/diagnosis codes enables the potential of an automatic mechanism to retrieve a set of ADE codes for a specific patient or disease domain. Although we manually identified and retrieved 6 ADE codes related to MS-AEs in this study, it would be valuable to build a valueset definition mechanism based on the semantic hierarchical relationship (eg, parent-child relation) asserted in standard terminologies, ideally through standard terminology services such as the recent development of the OMG/HL7 standard - common terminology services 2 (CTS2) (30, 31). Drug data normalization is another typical use case for utilizing standard terminologies and an important prerequisite for accurate ADE association. We noticed that the scope of drug name normalization must move beyond synonym mapping (itself a difficult task) as commonly used public resources also contain spelling errors and other types of irregularities. For example, from TCGA clinical drug data, the names "CYOTXAN", "CYTOXAN", "CYTOXEN", "CYCLOPHASPHAMIDE", "CYCLOPHOSPHAMID" are used to record the drug "Cyclophosphamide". A more detailed investigation on drug data normalization is beyond the scope of the present study and is addressed in a separate paper (32).

We also found that the severity and frequency of the ADEs are important factors for enabling our tumor stratification approach. As we mentioned in the Introduction section, breast cancer patients receiving AIs have a high incidence of MS-AEs; about half of patients treated with AIs have joint-related complaints (16, 17). Musculoskeletal complaints have been the most frequent reason given by patients on a clinical trial comparing the non-steroidal AI anastrozole with the steroidal AI exemestane as adjuvant therapy for early breast cancer (18). We believe that an ADE knowledge base with such severity and frequency information would be greatly useful in selecting the ADE phenotypes for tumor stratification, which is one of the goals of our ongoing ADEpedia project. In a previous study, we have developed an approach to build a severe ADE knowledge base based on the FDA Adverse Event Reporting System (AERS) reporting data (33). In this work, we propose a computational approach to screen genomic variants from individual patients and relate them to the probability of that patient experiencing an ADE while on a particular treatment.

In this study, we utilized the variant prioritization tool eXtasy for two primary reasons. First, eXtasy uses summary statistics of multiple criteria to evaluate the importance and contextual relevance of nonsynonymous SNVs according to conservation, interaction networks derived from experimental and knowledge databases, gene ontology, etc. This provides a succinct computable mechanism for aggregating knowledge associated with the nonsynonymous SNVs. Second, variant prioritization is guided by specific phenotypes, which in particular are organized using standard HPO terms. This provides a standard interface, allowing us to use the ADEs in HPO terms as input to the tool. In the future, we plan to evaluate the eXtasy tool in comparison with other variant prioritization algorithms. Pre-filtering the variants to only those we believe likely to affect function may be worth exploring, but may suffer from eXtasy having less power to make an appropriate association. The eXtasy algorithm only utilizes non-synonymous variants; all silent mutations (about 23.8%) are filtered out. It is intuitive that those affecting splicing, frameshift, or truncating would be more impactful than missense variants, but they are also more rare.

The rich cancer genomic data produced by TCGA Research Network provides a major opportunity to develop an integrated picture of commonalities, differences and emergent themes across tumor lineages. TCGA Pan-Cancer project (14) envisions that there are six types of omics characterization that create a data stack for maximizing the potential of integrative analysis. The six types comprise mutation, copy number, gene expression, DNA methylation, MicroRNA, reverse-phase proteomic arrays (RPPA) and clinical data. Hofree, *et al*, introduced a network-based stratification (NBS) method to integrate somatic tumor genomes with gene networks, which could identify subtypes in ovarian, uterine and lung cancer cohorts from TCGA (15). They demonstrated that the subtypes are predictive of clinical outcomes such as patient survival, response to therapy or tumor histology. By integrating mRNA, microRNA (miRNA), and DNA methylation next-generation sequencing data from TCGA, Volinia, *et al*. performed survival analysis on a cohort of 466 patients with primary invasive ductal carcinoma (IDC), and produced an integrated RNA signature that has been demonstrated prognostic to the IDC patients (34). The novelty of the present study is to build a tumor stratification method that utilizes the ADE-based variant prioritization, with the assumption that the underlying molecular mechanism of common ADEs known to cancer therapy drugs may overlap with that of the efficacy of the therapeutic drugs themselves, or have common indications. Although not presently utilizing the full data stack, our approach did identify subtypes predictive of patient survival time. We believe that an integrative

analysis with more omics data types would provide greater insights into the underlying biological mechanisms of the identified subtypes.

Since the impact of an individual variant is often difficult to interpret, in particular those only mutated in one (or a few) patient(s), directly comparing the landscape of genomic differences across patients is of great difficulty. To address the challenge, we performed a pathway enrichment analysis with multiple criteria and identified 63 canonical pathways that are highly related to the prioritized variants selected using the MS-AE phenotypes. In general, the definition of a pathway is a convenient abstraction for underlying molecular regulations, but cross-talk between pathways is often observed. Capitalizing upon cross-talk, we were able to perform a hierarchical clustering analysis to highlight the pathway-level patterns for the somatic variants among AI treated patients. Three clusters were identified, in which we found that the pathway pattern in Cluster 1 demonstrated its own characteristics in terms of pathway aberrations in comparison with the pathway patterns in Cluster 2 and 3. More promisingly, clinical outcome association analysis demonstrated that the survival time among the three clusters is significantly different, with Cluster 3 having the worst survival time (see Figure 2). We find that Clusters 2 and 3 have a similar pathway pattern in general, except for a few activated pathways responsible for DNA damaging repair and apoptosis. This perhaps represents a "two hits" pattern for Cluster 3 that may be responsible for the poorer survival outcome. We consider that our pathway-based clustering approach would make the findings from clinical outcome association studies more interpretable.

In summary, we developed a novel knowledge-driven approach that provides an ADE-based stratification of tumor mutations. We demonstrated that the prediction of per-patient ADE propensity simultaneously identifies high-risk patients experiencing poor outcomes. We plan to evaluate and validate our approach by incorporating more other data types (eg, germline variants) and other tumor types, and explore its potential in enabling pan-cancer studies in the future. The datasets and supplementary results produced by the study are publicly available at http://catargets.org.

References

1 Relling MV, Klein TE. CPIC: Clinical Pharmacogenetics Implementation Consortium of the Pharmacogenomics Research Network. Clinical pharmacology and therapeutics. 2011 Mar;**89**(3):464-7.

2 Green RC, Berg JS, Grody WW, et al. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. Genetics in medicine : official journal of the American College of Medical Genetics. 2013 Jul;**15**(7):565-74.

3 Wang L. Pharmacogenomics: a systems approach. Wiley interdisciplinary reviews Systems biology and medicine. 2010 Jan-Feb;**2**(1):3-22.

4 Karczewski KJ, Daneshjou R, Altman RB. Chapter 7: Pharmacogenomics. PLoS computational biology. 2012;**8**(12):e1002817.

5 Daly AK. Pharmacogenomics of adverse drug reactions. Genome medicine. 2013 Jan 29;**5**(1):5.

6 Phillips KA, Veenstra DL, Oren E, Lee JK, Sadee W. Potential role of pharmacogenomics in reducing adverse drug reactions: a systematic review. JAMA : the journal of the American Medical Association. 2001 Nov 14;**286**(18):2270-9.

7 Thorn CF, Klein TE, Altman RB. Pharmacogenomics and bioinformatics: PharmGKB. Pharmacogenomics. 2010 Apr;**11**(4):501-5.

8 Whirl-Carrillo M, McDonagh EM, Hebert JM, et al. Pharmacogenomics knowledge for personalized medicine. Clinical pharmacology and therapeutics. 2012 Oct;**92**(4):414-7.

9 Jiang G, Wang C, Zhu Q, Chute CG. A Framework of Knowledge Integration and Discovery for Supporting Pharmacogenomics Target Predication of Adverse Drug Events: A Case Study of Drug-Induced Long QT Syndrome. AMIA Summits on Translational Science proceedings AMIA Summit on Translational Science. 2013;**2013**:88-92.

10 Jiang G, Solbrig HR, Chute CG. ADEpedia: a scalable and standardized knowledge base of Adverse Drug Events using semantic web technology. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2011;**2011**:607-16.

11 Vazquez M, de la Torre V, Valencia A. Chapter 14: Cancer genome analysis. PLoS computational biology. 2012;**8**(12):e1002824.
12 Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Jr., Kinzler KW. Cancer genome landscapes. Science. 2013 Mar 29;**339**(6127):1546-58.
13 The Cancer Genome Atlas. [cited February 17, 2014]; Available from: http://cancergenome.nih.gov/
14 Weinstein JN, Collisson EA, Mills GB, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nature genetics. 2013 Oct;**45**(10):1113-20.
15 Hofree M, Shen JP, Carter H, Gross A, Ideker T. Network-based stratification of tumor mutations. Nature methods. 2013 Nov;**10**(11):1108-15.
16 Crew KD, Greenlee H, Capodice J, et al. Prevalence of joint symptoms in postmenopausal women taking aromatase inhibitors for early-stage breast cancer. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2007 Sep 1;**25**(25):3877-83.
17 Henry NL, Giles JT, Ang D, et al. Prospective characterization of musculoskeletal symptoms in early stage breast cancer patients treated with aromatase inhibitors. Breast cancer research and treatment. 2008 Sep;**111**(2):365-72.
18 Ingle JN, Schaid DJ, Goss PE, et al. Genome-wide associations and functional genomic studies of musculoskeletal adverse events in women receiving aromatase inhibitors. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2010 Nov 1;**28**(31):4674-82.
19 Wang L, Ellsworth KA, Moon I, et al. Functional genetic polymorphisms in the aromatase gene CYP19 vary the response of breast cancer patients to neoadjuvant therapy with aromatase inhibitors. Cancer research. 2010 Jan 1;**70**(1):319-28.
20 Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P. A side effect resource to capture phenotypic effects of drugs. Molecular systems biology. 2010;**6**:343.
21 Kohler S, Doelken SC, Mungall CJ, et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic acids research. 2014 Jan 1;**42**(1):D966-74.
22 The Human Phenotype Ontology URL. [cited February 17, 2014]; Available from: http://www.human-phenotype-ontology.org/
23 Sifrim A, Popovic D, Tranchevent LC, et al. eXtasy: variant prioritization by genomic data fusion. Nature methods. 2013 Nov;**10**(11):1083-4.
24 eXtasy URL. [cited February 15, 2014]; Available from: http://homes.esat.kuleuven.be/~bioiuser/eXtasy/
25 Aerts S, Lambrechts D, Maity S, et al. Gene prioritization through genomic data fusion. Nature biotechnology. 2006 May;**24**(5):537-44.
26 eXtasy GitHub URL. [cited February 15, 2014]; Available from: https://github.com/asifrim/eXtasy
27 TCGA Data Portal. [cited February 17, 2014]; Available from: https://tcga-data.nci.nih.gov/tcga/
28 Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011 Jun 15;**27**(12):1739-40.
29 Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America. 2005 Oct 25;**102**(43):15545-50.
30 Pathak J, Bailey KR, Beebe CE, et al. Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium. Journal of the American Medical Informatics Association : JAMIA. 2013 Dec;**20**(e2):e341-8.
31 CTS2 Wiki. [cited March 12, 2014]; Available from: http://informatics.mayo.edu/cts2/index.php/Main_Page
32 Jiang G, Sohn S, Zimmermann MT, Liu H, Chute CG. Drug Normalization for Cancer Therapeutic and Druggable Genome Target Discovery. Proceedings of ICBO 2014 - International VDOS Workshop (in submission). Houston, TX; 2014.
33 Jiang G, Wang L, Liu H, Solbrig HR, Chute CG. Building a knowledge base of severe adverse drug events based on AERS reporting data using semantic web technologies. Studies in health technology and informatics. 2013;**192**:496-500.
34 Volinia S, Croce CM. Prognostic microRNA/mRNA signature from the integrated analysis of patients with invasive breast cancer. Proceedings of the National Academy of Sciences of the United States of America. 2013 Apr 30;**110**(18):7413-7.

Clinical Risk Prediction by Exploring High-Order Feature Correlations

Fei Wang, Ping Zhang, Xiang Wang, Jianying Hu
IBM T. J. Watson Research Center, Yorktown Heights, NY
{fwang,ping,wangxi,jyhu}@us.ibm.com

Abstract

Clinical risk prediction is one important problem in medical informatics, and logistic regression is one of the most widely used approaches for clinical risk prediction. In many cases, the number of potential risk factors is fairly large and the actual set of factors that contribute to the risk is small. Therefore sparse logistic regression is proposed, which can not only predict the clinical risk but also identify the set of relevant risk factors. The inputs of logistic regression and sparse logistic regression are required to be in vector form. This limits the applicability of these models in the problems when the data cannot be naturally represented vectors (e.g., medical images are two-dimensional matrices). To handle the cases when the data are in the form of multi-dimensional arrays, we propose *HOSLR*: High-Order Sparse Logistic Regression, which can be viewed as a high order extension of sparse logistic regression. Instead of solving one classification vector as in conventional logistic regression, we solve for K classification vectors in *HOSLR* (K is the number of modes in the data). A block proximal descent approach is proposed to solve the problem and its convergence is guaranteed. Finally we validate the effectiveness of *HOSLR* on predicting the onset risk of patients with Alzheimer's disease and heart failure.

1 Introduction

Predictive modeling of clinical risk, such as disease onset [1] or hospitalization [2], is an important problem in medical informatics. Effective risk prediction can be very helpful for the physician to make proper decision and provide the right service at point-of-care.

Typically we need three steps to perform patient clinical risk prediction:

1. Collecting all potential risk factors from patient historical data and utilizing them to properly represent each patient (e.g., as a vector [1][3]).
2. Identifying important risk factors from the risk factor pool collected in the first step, such that the value change of the selected risk factors could generate big impact on the predicted risk.
3. Training a proper predictive model based on the patients represented with the selected risk factors from the second step. Such model will be used to score the clinical risk of new testing patients.

One representative clinical risk prediction work that follows those three steps is the work by Sun *et al.* [1], where the goal is to predict the onset risk for potential heart failure patients. The authors first collect all potential risk factors from the two year patient electronic health records, and then designed an scalable orthogonal regression method to identify important risk factors, which will be used to train a logistic regression model for risk prediction at last. The authors showed that they can achieve the state-of-the-art performance as well as identify clinically meaningful risk factors for heart failure.

Note that in practice, depending on the concrete risk factor identification method and predictive model, step 2 and 3 could be combined into one step, i.e., a unified model can be constructed for both prediction and risk factor identification (e.g., LASSO [4]). This will make the constructed model more integrative and interpretable. Sparse Logistic Regression [5] is one such model. As is known to all that logistic regression is a popular model for clinical risk prediction [6] [1][3]. However, the pool of potential risk factors is usually very large and noisy, which would affect the efficiency and performance of predictive modeling. The main difference between sparse and convectional logistic regression is it adds an one norm regularizer on the model coefficients to encourage the model sparsity, so that only

those *important* risk factors will contribute to the final predictions. In recent years people have also been doing research on constructing different regularization terms to enforce different sparsity structures on the model coefficients, such as the ℓ_p norm [7], group sparsity [8] and elastic net regularization [9].

One limitation of the existing sparse logistic regression type of approaches is that they assume vector based inputs, which means that we need to have a vector based representation for each patient before we can use those methods to evaluate the patient's clinical risk. However, we are in the era of *big data* with *variety* as one representative characteristic, so does medical data, i.e., there are many medical data are not naturally in vector form. For example, typical medical images (e.g., X-Ray and MRI) are two dimensional matrices, with some more advanced medical imaging technologies can even generate three-dimensional image sequences (e.g., functional Magnetic Resonance Imaging (fMRI)). In a recent paper, Ho *et al*. [10] proposed a *tensor* (which can be viewed as high order generalization of matrix) based representation of patient Electronic Health Records (EHRs) to capture the interactions between different *modes* in patient EHRs. For example, medication order information for every patient could be captured by a 2nd order tensor with 2 modes, where each mode is an aspect of a tensor: a) medication and b) diagnosis. With such a representation we can take into consideration the correlation between diagnosis and drugs when predicting the patient risk. If there are more inter-correlated modes in the data then we will need to represent the patient in higher order tensors. In these cases, if we still want to apply logistic regression one straightforward way is to stretch those matrices and tensors into vectors as people did in image processing, but this will lose the correlation information among different dimensions. Moreover, after stretching the dimensionality of the data objects will become very high, which will make traditional sparse logistic regression inefficient.

In recent years, there has been a lot of research on extending traditional vector based approaches to 2nd (matrix based) or higher order (tensor based) settings. Two representative examples are two-dimensional Principal Component Analysis (PCA) [11] and Linear Discriminant Analysis [12], which have been found to be more effective on computer vision tasks compared to traditional vector based PCA and LDA. Recently, Huang and Wang [13] developed a matrix variate logistic regression model and applied it in electroencephalography data analysis. Tan *et al.* [14] further extended logistic regression to tensor inputs and achieved good performance in a video classification task.

In this paper, we propose *HOSLR*, a *H*igh-*O*rder *S*parse *L*ogistic *R*egression method that can perform prediction based on matrix or tensor inputs. Our model learns a linear decision vector on every mode of the input, and we added an ℓ_1 regularization term on each decision vector to encourage sparsity. We developed a *Block Proximal Gradient* (BPG) [15] method to solve the problem iteratively. The convergence of the proposed algorithm can be guaranteed by the Kurdyka–Lojasiewicz inequality [16] (for proof details please see a more technical version of this paper [17]). Finally we validate the effectiveness of our algorithm on two real world medical scenarios on the risk prediction of patients with Alzheimer's Disease and Heart Failure.

The rest of this paper is organized as follows. Section 2 reviews some related works. The details along with the convergence analysis of *HOSLR* is introduced in Section 3. Section 4 presents the experimental results, followed by the conclusions in Section 5.

2 Related Work

Logistic regression [18] is a statistical prediction method that has widely been used in medical informatics [1][6][19]. Suppose we have a training data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$, where $\mathbf{x}_i \in \mathbb{R}^d$ $(1 \leqslant i \leqslant n)$ is the i-th training data vector with dimensionality d, and associated with each $\mathbf{x}_i$ we also have its corresponding label $y_i \in \{+1, -1\}$. The goal of logistic regression is to train a linear decision function $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$ to discriminate the data in class +1 from the data in class -1 by minimizing the following logistic loss

$$\ell_{org}(\mathbf{w}, b) = \frac{1}{n}\sum_{i=1}^{n} \log\left[1 + \exp(-y_i(\mathbf{w}^\top \mathbf{x}_i + b))\right] \tag{1}$$

where $\mathbf{w} \in \mathbb{R}^d$ is the decision vector and b is the bias. They can be learned with gradient descent type of approaches.

In many medical informatics applications, the data vectors $\{\mathbf{x}_i\}_{i=1}^n$ are sparse and high-dimensional (e.g., each patient could be a tens of thousands dimensional vector with bag-of-feature representation [1]). To enhance the interpretability

of the model in these scenarios, we can minimize the following ℓ_1-regularized logistic loss

$$\ell_{sp}(\mathbf{w}, b) = \frac{1}{n}\sum_{i=1}^{n} \log\left[1 + \exp(-y_i(\mathbf{w}^\top \mathbf{x}_i + b))\right] + \lambda\|\mathbf{w}\|_1 \tag{2}$$

where $\|\cdot\|_1$ is the vector $\ell 1$ norm and $\lambda > 0$ is a factor trading off the prediction accuracy and model sparsity. The resultant model is usually referred to as sparse logistic regression model [5][20]. Compared with the conventional logistic regression model obtained by minimizing $\mathcal{J}_{org}$, the $\mathbf{w}$ obtained by minimizing $\mathcal{J}_{sp}$ is sparse thanks to the ℓ_1 norm regularization. In this way, we can not only get a predictor, but also know what are the feature dimensions that are important to the prediction (which are the features with nonzero classification coefficients).

Sparse logistic regression has widely been used in health informatics because it can achieve a good balance between model accuracy and model interpretability. For example, sparse logistic regression has been used in the prediction of Leukemia [21], Alzheimer's disease [22] and cancers [23]. In recent years people also designed different regularization terms [7][8][9] to enforce more complex sparsity patterns on the learned model. However, all these works require a vector based data representations. Under this framework, if the data naturally come as tensors (like medical imaging), we need to first stretch them into vectors before we can apply sparse logistic regression. This may lose the correlation structure among different modes in the original data, while for the *HOSLR* method proposed in this paper, we directly work with data in tensor representations. Fig.1 provides a graphical illustration on the difference of traditional vector based logistic regression and high order logistic regression when working on multi-dimensional data.

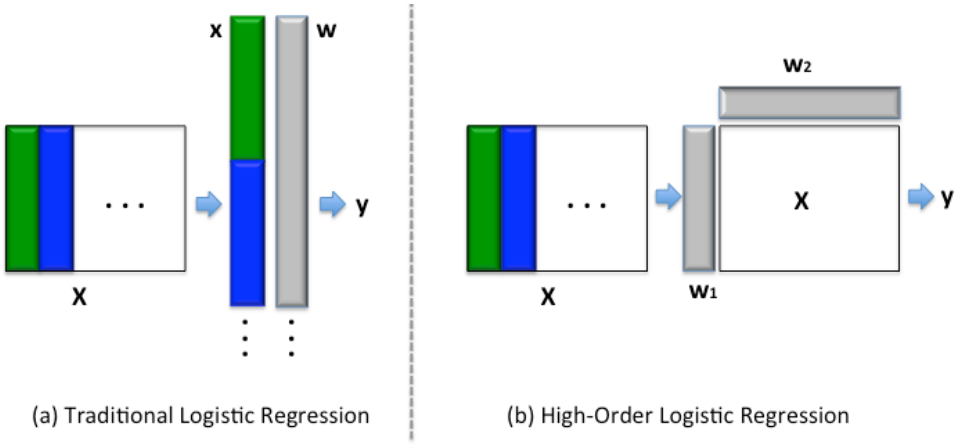


Figure 1: Traditional vector based logistic regression and high-order logistic regression work on multi-dimensional data.

3 Methodology

We introduce the details of *HOSLR* in this section. First we will formally define the problem.

3.1 Problem Statement

Without the loss of generality, we assume each observation is a tensor $\mathcal{X}^i \in \mathbb{R}^{d_1 \times d_2 \times \cdots d_K}$, suppose its corresponding response is $y^i \in \{0, 1\}$, then *HOSLR* assumes

$$y^i \leftarrow \mathcal{X}^i \times_1 \mathbf{w}^1 \times_2 \mathbf{w}^2 \cdots \times_K \mathbf{w}^K + b \tag{3}$$

where $\times_k$ is the mode-k product, and $\mathbf{w}^k \in \mathbb{R}^{d_k \times 1}$ is the prediction coefficients on the k-th dimension. Then

$$\mathcal{X}^i \times_1 \mathbf{w}^1 \times_2 \mathbf{w}^2 \cdots \times_K \mathbf{w}^K = \sum_{i_1=1}^{d_1}\sum_{i_2=1}^{d_2}\cdots\sum_{i_K=1}^{d_K} w^1_{i_1} w^2_{i_2} \cdots w^K_{i_K} X^i_{i_1 i_2 \cdots i_K} \tag{4}$$

Let $\mathcal{W} = \{\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_K\}$ be the set of prediction coefficient vectors. The loss we want to minimize is

$$\begin{aligned} \ell(\mathcal{W}, b) &= \frac{1}{n}\sum_{i=1}^{n} \ell(\mathcal{X}_i, y_i, \mathcal{W}) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \mathcal{X}^i \times_1 \mathbf{w}^1 \times_2 \mathbf{w}^2 \cdots \times_K \mathbf{w}^K + b) \\ &= \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f_{(\mathcal{W},b)}(\mathcal{X}^i)) \end{aligned}$$

where for notational convince, we denote

$$f_{(\mathcal{W},b)}(\mathcal{X}^i) = \mathcal{X}^i \times_1 \mathbf{w}^1 \times_2 \mathbf{w}^2 \cdots \times_K \mathbf{w}^K + b \tag{5}$$

The loss function we considered in this paper is *Logistic Loss*:

$$\ell_l(\mathcal{W}, b) = \log[1 + \exp(-y_i f_{(\mathcal{W},b)}(\mathcal{X}^i))] \tag{6}$$

We also introduce the regularization term

$$\mathcal{R}(\mathcal{W}) = \mathcal{R}_1(\mathcal{W}) + \mathcal{R}_2(\mathcal{W}) = \sum_{k=1}^{K} \lambda_k \|\mathbf{w}^k\|_1 + \frac{1}{2}\sum_{k=1}^{K} \mu_k \|\mathbf{w}^k\|_2^2 \tag{7}$$

which is usually referred to as *elastic net* regularization [24]. This regularizer is a combination of ℓ_1 and ℓ_2 norm regularizations, thus it can achieve better numerical stability and reliability [24]. Then the optimization problem we want to solve is

$$\min_{\mathcal{W}} \mathcal{J}(\mathcal{W}, b) = \ell(\mathcal{W}, b) + \mathcal{R}(\mathcal{W}) \tag{8}$$

We adopt a *Block Coordinate Descent* (BCD) procedure to solve the problem. Starting from some initialization $(\mathcal{W}_{(0)}, b_{(0)})$, at the i-th step of the t-th round of updates, we update $(\mathbf{w}^k_{(t)}, b_{(t)})$ by

$$(\mathbf{w}^k_{(t)}, b_{(t)}) = \arg\min_{(\mathbf{w},b)} \left[\ell(\mathcal{W}^{1\sim(k-1)}_{(t)}, \mathbf{w}, \mathcal{W}^{(k+1)\sim K}_{(t-1)}, b) + \lambda_k \|\mathbf{w}\|_1 + \frac{\mu_k}{2}\|\mathbf{w}\|_2^2\right]$$

where $\mathcal{W}^{1\sim(k-1)}_{(t)} = \left\{\mathbf{w}^1_{(t)}, \mathbf{w}^2_{(t)}, \cdots, \mathbf{w}^{k-1}_{(t)}\right\}$ and $\mathcal{W}^{(k+1)\sim K}_{(t-1)} = \left\{\mathbf{w}^{k+1}_{(t-1)}, \mathbf{w}^{k+2}_{(t-1)}, \cdots, \mathbf{w}^{K}_{(t-1)}\right\}$.

Algorithm 1 Block Coordinate Descent Procedure

Require: Data set $\{\mathcal{X}_i, y_i\}_{i=1}^n$, Regularization parameters $\{\lambda_k, \mu_k\}_{k=1}^K$
1: **Initialization**: $(\mathcal{W}_{(0)}, b_{(0)})$, $t = 0$
2: **while** Not Converge **do**
3: **for** $k = 1 : K$ **do**
4: Update $(\mathbf{w}^k_{(t)}, b_{(t,k)})$ by solving problem (9)
5: $t = t + 1$
6: **end for**
7: **end while**

3.2 Proximal Gradient Descent

Algorithm 2 summarized the whole algorithmic flow of our algorithm, where $\alpha^k_{(t)} = \lambda_k/(\tau^k_{(t)} + \mu_k)$ and $\mathcal{S}_{\alpha^k_{(t)}}$ is the component-wise shrinkage operator defined as

$$\left(\mathcal{S}_{\alpha^k_{(t)}}(\mathbf{v})\right)_i = \begin{cases} v_i - \alpha^k_{(t)}, & \text{if } v_i > \alpha^k_{(t)} \\ v_i + \alpha^k_{(t)}, & \text{if } v_i < -\alpha^k_{(t)} \\ 0, & \text{if } |v_i| \leqslant |\alpha^k_{(t)}| \end{cases} \tag{9}$$

At each iteration the most time consuming part is evaluating the gradient, which takes $O(n\prod_{i=1}^K d_i)$ time, that is linear with respect to data set size and data dimension. The detailed algorithm derivation can be referred to [17].

Algorithm 2 Block Proximal Gradient Descent for Multilinear Sparse Logistic Regression

Require: Data set $\{\mathcal{X}_i, y_i\}_{i=1}^n$, Regularization parameters $\{\lambda_k, \mu_k\}_{k=1}^K$, $r_0 = 1$, $\delta_\omega < 1$

Initialization: $(\mathcal{W}_{(0)}, b_{(0)})$, $t = 1$
while Not Converge **do**
  for $k = 1 : K$ **do**
    Compute $\tau_{(t)}^k$ with $\tau_{(t)}^k = \frac{\sqrt{2}}{n} \sum_{i=1}^n \left(\left\| \nabla_{\mathbf{w}^k}^{(t,k)} f_{(\mathcal{W},b)}(\mathcal{X}^i) \right\|_2 + 1 \right)^2$
    Compute $\omega_{(t)}^k$ with $\omega_{(t)}^k = \min\left(\omega_{(t)}, \delta_\omega \sqrt{\frac{\tau_{(t-1)}^k}{\tau_{(t)}^k}}\right)$
    Compute $\widetilde{\mathbf{w}}_{(t)}^k$ with $\widetilde{\mathbf{w}}_{(t)}^k = \mathbf{w}_{(t-1)}^k + \omega_{(t)}^k(\mathbf{w}_{(t-1)}^k - \mathbf{w}_{(t-2)}^k)$
    Update $\mathbf{w}_{(t)}^k$ by $\mathbf{w}_{(t)}^k = \mathcal{S}_{\alpha_{(t)}^k}\left(\frac{\tau_{(t)}^k \widetilde{\mathbf{w}}_{(t)}^k - \nabla_{\mathbf{w}^k} \ell_{(t)}^k(\widetilde{\mathbf{w}}_{(t)}^k, b_{(t,k-1)})}{\tau_{(t)}^k + \mu_k}\right)$
    Compute $\widetilde{b}_{(t,k)}$ with $\widetilde{b}_{(t,k)} = b_{(t,k-1)} + \omega_{(t)}^k(b_{(t,k-1)} - b_{(t,k-2)})$
    Update $b_{(t,k)}$ by $b_{(t,k)} = \widetilde{b}_{t,k} - \frac{1}{\tau_{(t)}^k} \nabla_b \ell_{(t)}^k(\mathbf{w}_{(t)}^k, \widetilde{b}_{(t,k)})$
  end for
  if $\ell(\mathcal{W}_{(t-1)}, b_{(t-1,K)}) \leqslant \ell(\mathcal{W}_{(t)}, b_{(t,K)})$ **then**
    Reupdate $\mathbf{w}_{(t)}^k$ and $b_{(t,k)}$ using $\mathbf{w}_{(t)}^k = \mathcal{S}_{\alpha_{(t)}^k}\left(\frac{\tau_{(t)}^k \widetilde{\mathbf{w}}_{(t)}^k - \nabla_{\mathbf{w}^k} \ell_{(t)}^k(\widetilde{\mathbf{w}}_{(t)}^k, b_{(t,k-1)})}{\tau_{(t)}^k + \mu_k}\right)$ and $b_{(t,k)} = \widetilde{b}_{t,k} - \frac{1}{\tau_{(t)}^k} \nabla_b \ell_{(t)}^k(\mathbf{w}_{(t)}^k, \widetilde{b}_{(t,k)})$, with $\widetilde{\mathbf{w}}_{(t)}^k = \mathbf{w}_{(t-1)}^k$ and $\widetilde{b}_{(t,k)} = b_{(t,k-1)}$
  end if
  $t = t + 1$
end while

4 Experiments

In this section we will present the experimental results on applying *HOSLR* to predict the onset risk of potential Alzheimer's Disease patients from their fMRI images, and the onset risk of potential heart failure patients from their EHR data.

4.1 Experiments on Predicting the Onset Risk of Alzheimer's Disease

Alzheimer's disease (AD) is the most common form of dementia. It worsens as it progresses and eventually leads to death. There is no cure for the disease. AD is usually diagnosed in elder people (typically over 65 years of age), although the less-prevalent early-onset Alzheimer's can occur much earlier. There are currently more than 5 million Americans living with Alzheimer's disease and that number is poised to grow to as many as 16 million by 2050. The care for has been the country's most expensive condition, which costs the nation $203 billion annually with projections to reach $1.2 trillion by 2050 [25].

Early detection of AD is of key importance for its effective intervention and treatment, where *functional magnetic resonance imaging or functional MRI* (fMRI) [26] is an effective approach to investigate alterations in brain function related to the earliest symptoms of Alzheimer's disease, possibly before development of significant irreversible structural damage.

In this set of experiment, we adopted a set of fMRI scans collected from real clinic cases of 1,005 patients [27], whose cognitive function scores (semantic, episodic, executive and spatial - ranges between -2.8258 and 2.5123) were also acquired at the same time using a cognitive function test. There are three types of MRI scans that were collected from the subjects: (1) FA, the fractional anisotropy MRI gives information about the shape of the diffusion tensor at each voxel, which reflects the differences between an isotropic diffusion and a linear diffusion; (2) FLAIR, Fluid attenuated inversion recovery is a pulse sequence used in MRI, which uncovers the white matter hyperintensity of the brain; (3)

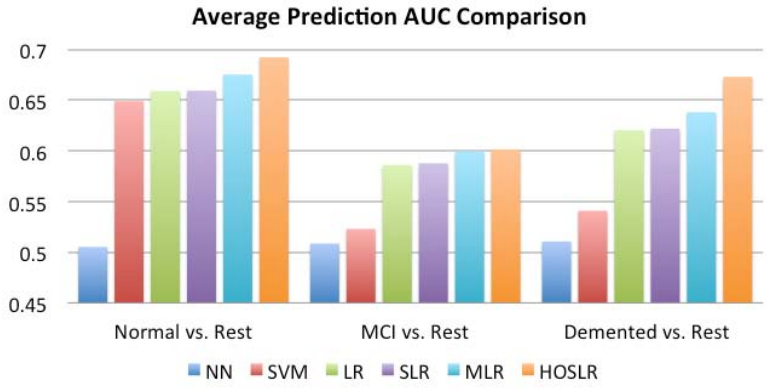


Figure 2: Average prediction AUC over 5-fold cross validation comparison for different methods.

GRAY, gray MRI images revealing the gray matter of the brain. In the raw scans, each voxel has a value from 0 to 1, where 1 indicates that the structural integrity of the axon tracts at that location is perfect, while 0 implies either there are no axon tracts or they are shot (not working). The raw scans are preprocessed (including normalization, denoising and alignment) and then restructured to 3D tensors with a size of $134 \times 102 \times 134$. Associated with each sample we have a label, which could be either *normal*, *Mild Cognitive Impairment* (MCI) or *demented*.

We constructed three binary classification problems to test the effectiveness of our *HOSLR* method, i.e., *Normal* vs. *Rest* (MCI and Demented), *MCI* vs. *Rest* (Normal and Demented), *Demented* vs. *Rest* (Normal and MCI). For *HOSLR*, because the input fMRI images are three dimensional tensors, we set the ℓ_1 term regularization parameters on all three dimensions equal, i.e., $\lambda_1 = \lambda_2 = \lambda_3$ and tune it from the grid $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$ with five fold cross validation. The ℓ_2 term regularization parameters are set to $\mu_1 = \mu_2 = \mu_3 = 10^{-4}$. For comparison purpose, we also implemented the following baseline algorithms:

- **Nearest Neighbor** (NN). This is the one nearest neighbor classifier with standard Euclidean distance.
- **Support Vector Machine** (SVM). This is the regular vector based SVM method.
- **Logistic Regression** (LR). This is the traditional vector based logistic regression method.
- **Sparse Logistic Regression** (SLR). This is the vector based sparse logistic regression.
- **Multilinear Logistic Regression** (MLR). This is equivalent to *HOSLR* with all $\ell 1$ regularization parameters setting to 0.

We use *LIBLINEAR* [28] for the implementation of LR and SLR, and *LIBSVM* [29] for the implementation of SVM. Note that in order to test those vector based approaches, we need to stretch those fMRI tensors into very long vectors (with dimensionality 1,831,512). Fig.2 summarized the average performance over 5-fold cross validation in terms of Areas Under the receiver operating characteristics Curve (AUC) values. The data we used are the FLAIR images. From the figure we can observe that *HOSLR* beats all other competitors in all three tasks. This is because *HOSLR* can not only take into consideration the spatial correlation between three different dimensions in those fMRI images, but also exploring their joint sparsity structures (the FLAIR images are sparse in nature).

4.2 Experiments on Predicting the Onset Risk of Congestive Heart Failure Patients

Congestive heart failure (CHF), occurs when the heart is unable to pump sufficiently to maintain blood flow to meet the needs of the body, is a major chronic illness in the U.S. affecting more than five million patients. It is estimated CHF costs the nation an estimated $32 billion each year [30]. Effective prediction of the onset risk of potential CHF

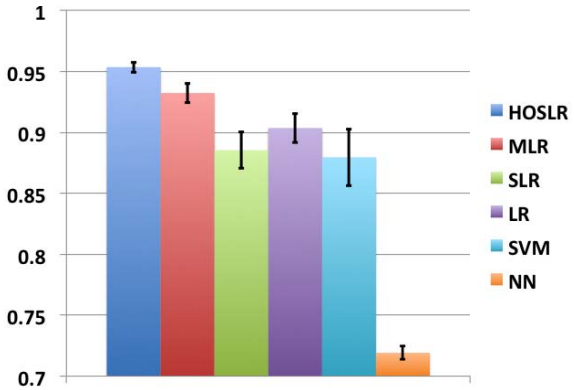


Figure 3: Prediction performance for different methods on the CHF onset prediction task in terms of averaged AUC value with 5-fold cross validation along with their standard deviations.

patients would help identify the patient at risk in time, and thus the decision makers can provide the proper treatment. This can also help save huge amount of unnecessary costs.

The data set we use in this set of experiments is from a real world Electronic Health Record (EHR) data warehouse including the longitudinal EHR of 319,650 patients over 4 years. On this data set, we identified 1,000 CHF case patients according to the diagnostic criteria in [3]. Then we obtained 2,000 group matched controls according to patient demographics, comorbidities and primary care physicians similar as in [3]. We use the medication orders of those patients within two years from their operational criteria date (for case patients, their operational criteria dates are just their CHF confirmation date; for control patients that date is just the date of their last records in the database). On each medication order we use the corresponding pharmacy class according to the United States Pharmacopeial (USP) convention[1] and the primary diagnosis in terms of Hierarchical Condition Category (HCC) codes [31] for the medication prescription. In total there are 92 unique pharmacy classes and 195 distinct HCC codes appeared in those medication orders. Therefore each patient can be represented as a 92×195 matrix, where the (i, j)-th entry indicates the frequency that the i-th drug was prescribed during the two years with the j-th diagnosis code as primary diagnosis.

The parameters for *HOSLR* are set in the same manner as the experiments in last subsection. For comparison purpose, we also implemented NN, SVM, LR, SLR, MLR and reported the averaged AUC value over 5-fold cross validation along with their standard deviations on Fig.3. From the figure we can get similar observations as we saw in Fig.2.

Another interesting thing to check is which medications and diagnosis play key roles during the decision. Because in this set of experiments we have two feature modes: medications and diagnosis, we will get two decision vectors $\mathbf{w}_{\text{med}}$ and $\mathbf{w}_{\text{diag}}$, one on each mode. The bilinear decision function in this case can be written as

$$f(\mathbf{X}) = \mathbf{w}_{\text{med}}^\top \mathbf{X} \mathbf{w}_{\text{diag}} = \mathbf{1}^\top \left((\mathbf{w}_{\text{med}} \mathbf{w}_{\text{diag}}^\top) \odot \mathbf{X} \right) \mathbf{1} \tag{10}$$

where $\mathbf{1}$ denotes all-one vector of appropriate dimension, $\odot$ is element-wise matrix product. The importance of the (i, j)-th feature X_{ij} to the decision can be evaluated as $w_{\text{med}}(i) w_{\text{diag}}(j)$. Therefore if both the magnitudes of $w_{\text{med}}(i)$ and $w_{\text{diag}}(j)$ are large, then the feature pair (medication i, diagnosis j) will definitely be important. We list in Table 1 the top diagnoses and medications according to their coefficient magnitudes in $\mathbf{w}_{\text{diag}}$ and $\mathbf{w}_{\text{med}}$. From the table we can see that the diagnoses are mainly hypertension, heart disease and some common comorbidities of heart failure including chronic lung disease (e.g., *Chronic Obstructive Pulmonary Disease* (COPD) [32]) and chronic kidney disease [33]. The top medications include drugs for treating heart disease such as Beta blockers and calcium blockers, and medicine for treating lung disease such as *Corticosteroids*. There are also medicine for treating heart failure related symptom, such as Gout, which is a well-known Framingham symptom [34]. Vaccine is also an important treatment for reducing the stress on heart [35].

[1] http://www.usp.org/

Diagnosis	
Heart Disease	Congestive Heart Failure
	Acute Myocardial Infarction
	Specified Heart Arrhythmias
	Ischemic or Unspecified Stroke
Hypertension	Hypertension
	Hypertensive Heart Disease
Lung Disease	Fibrosis of Lung and Other Chronic Lung Disorders
	Asthma
	Chronic Obstructive Pulmonary Disease (COPD)
Kidney Disease	Chronic Kidney Disease, Very Severe (Stage 5)
	Chronic Kidney Disease, Mild or Unspecified (Stage 1-2 or Unspecified)

Medication
Antihyperlipidemic
Antihypertensive
Beta Blockers
Calcium Blockers
Cardiotonics
Cardiovascular
Corticosteriods
Diuretics
General Anesthetics
Gout
Vaccines

Table 1: Top diagnosis and medications according to the magnitude of their corresponding decision coefficient

5 Conclusions

We propose a high order sparse logistic regression method called *HOSLR* in this paper, which can directly take data matrices or tensors as inputs and do prediction on that. *HOSLR* is formulated as an optimization problem and we propose an effective BCD strategy to solve it. We validate the effectiveness of *HOSLR* on two real world medical scenarios on predicting the onset risk of Alzheimer's disease and heart failure. We demonstrate that *HOSLR* can not only achieve good performance, but also discover interesting predictive patterns.

References

[1] Jimeng Sun, Jianying Hu, Dijun Luo, Marianthi Markatou, Fei Wang, Shahram Edabollahi, Steven E Steinhubl, Zahra Daar, and Walter F Stewart. Combining knowledge and data driven insights for identifying risk factors using electronic health records. In *AMIA Annual Symposium Proceedings*, volume 2012, page 901. American Medical Informatics Association, 2012.

[2] Edward F Philbin and Thomas G DiSalvo. Prediction of hospital readmission for heart failure: development of a simple risk score based on administrative data. *Journal of the American College of Cardiology*, 33(6):1560–1566, 1999.

[3] Jionglin Wu, Jason Roy, and Walter F Stewart. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. *Medical care*, 48(6):S106–S113, 2010.

[4] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[5] Shirish Krishnaj Shevade and S Sathiya Keerthi. A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, 19(17):2246–2253, 2003.

[6] Marc Miravitlles, Tina Guerrero, Cristina Mayordomo, Leopoldo Sánchez-Agudo, Felip Nicolau, and José Luis Segú. Factors associated with increased risk of exacerbation and hospital admission in a cohort of ambulatory COPD patients: a multiple logistic regression analysis. *Respiration*, 67(5):495–501, 2000.

[7] Zhenqiu Liu, Feng Jiang, Guoliang Tian, Suna Wang, Fumiaki Sato, Stephen J Meltzer, and Ming Tan. Sparse logistic regression with Lp penalty for biomarker identification. *Statistical Applications in Genetics and Molecular Biology*, 6(1), 2007.

[8] Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group LASSO for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

[9] Srikanth Ryali, Kaustubh Supekar, Daniel A Abrams, and Vinod Menon. Sparse logistic regression for whole-brain classification of fMRI data. *NeuroImage*, 51(2):752–764, 2010.

[10] Joyce C Ho, Joydeep Ghosh, Steve Steinhubl, Walter Stewart, Joshua C Denny, Bradley A Malin, and Jimeng Sun. Limestone: High-throughput candidate phenotype generation via tensor factorization. *Journal of Biomedical Informatics*, 2014.

[11] Jian Yang, David Zhang, Alejandro F Frangi, and Jing-yu Yang. Two-dimensional PCA: a new approach to appearance-based face representation and recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(1):131–137, 2004.

[12] Jieping Ye, Ravi Janardan, Qi Li, et al. Two-dimensional linear discriminant analysis. In *Advances in Neural Information Processing Systems*, volume 17, 2004.

[13] Hung Hung and Chen-Chien Wang. Matrix variate logistic regression model with application to EEG data. *Biostatistics*, 14(1):189–202, 2013.

[14] Xu Tan, Yin Zhang, Siliang Tang, Jian Shao, Fei Wu, and Yueting Zhuang. Logistic tensor regression for classification. In *Intelligent Science and Intelligent Data Engineering*, pages 573–581. Springer, 2013.

[15] Yangyang Xu. Alternating proximal gradient method for sparse nonnegative Tucker decomposition. *Mathematical Programming Computation*, pages 1–32, 2013.

[16] Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. The Lojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007.

[17] Fei Wang, Ping Zhang, Buyue Qian, Xiang Wang, and Ian Davidson. Clinical risk prediction with multilinear sparse logistic regression. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014.

[18] David W Hosmer Jr and Stanley Lemeshow. *Applied logistic regression*. John Wiley & Sons, 2004.

[19] Shuo Xiang, Lei Yuan, Wei Fan, Yalin Wang, Paul M Thompson, and Jieping Ye. Multi-source learning with block-wise missing data for Alzheimer's disease prediction. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 185–193. ACM, 2013.

[20] Jun Liu, Jianhui Chen, and Jieping Ye. Large-scale sparse logistic regression. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 547–556. ACM, 2009.

[21] Tapio Manninen, Heikki Huttunen, Pekka Ruusuvuori, and Matti Nykter. Leukemia prediction using sparse logistic regression. *PloS one*, 8(8), 2013.

[22] Anil Rao, Ying Lee, Achim Gass, and Andreas Monsch. Classification of Alzheimer's disease from structural MRI using sparse logistic regression with optional spatial regularization. In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, pages 4499–4502. IEEE, 2011.

[23] Yongdai Kim, Sunghoon Kwon, and Seuck Heun Song. Multiclass sparse logistic regression for classification of multiple cancer types using gene expression data. *Computational Statistics & Data Analysis*, 51(3):1643–1655, 2006.

[24] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[25] William Thies and Laura Bleiler. 2013 Alzheimer's disease facts and figures. *Alzheimer's & dementia: the journal of the Alzheimer's Association*, 9(2):208–245, 2013.

[26] Scott A Huettel, Allen W Song, and Gregory McCarthy. *Functional magnetic resonance imaging*, volume 1. Sinauer Associates Sunderland, MA, 2004.

[27] Buyue Qian, Xiang Wang, Fei Wang, Hongfei Li, Jieping Ye, and Ian Davidson. Active learning from relative queries. In *Proceedings of International Joint Conference on Artificial Intelligence*, pages 1614–1620, 2013.

[28] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[29] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[30] Paul A Heidenreich, Justin G Trogdon, Olga A Khavjou, Javed Butler, Kathleen Dracup, Michael D Ezekowitz, Eric Andrew Finkelstein, Yuling Hong, S Claiborne Johnston, Amit Khera, et al. Forecasting the future of cardiovascular disease in the United States a policy statement from the American heart association. *Circulation*, 123(8):933–944, 2011.

[31] Gregory C Pope, Randall P Ellis, Arlene S Ash, JZ Ayanian, DW Bates, H Burstin, LI Iezzoni, E Marcantonio, and B Wu. Diagnostic cost group hierarchical condition category models for Medicare risk adjustment. *Health Economics Research, Inc. Waltham, MA*, 2000.

[32] Frans H Rutten, Maarten-Jan M Cramer, Jan-Willem J Lammers, Diederick E Grobbee, and Arno W Hoes. Heart failure and chronic obstructive pulmonary disease: an ignored combination? *European journal of heart failure*, 8(7):706–711, 2006.

[33] Ali Ahmed, Michael W Rich, Paul W Sanders, Gilbert J Perry, George L Bakris, Michael R Zile, Thomas E Love, Inmaculada B Aban, and Michael G Shlipak. Chronic kidney disease associated mortality in diastolic versus systolic heart failure: a propensity matched study. *The American journal of cardiology*, 99(3):393–398, 2007.

[34] Patrick A McKee, William P Castelli, Patricia M McNamara, and William B Kannel. The natural history of congestive heart failure: the Framingham study. *New England Journal of Medicine*, 285(26):1441–1446, 1971.

[35] Matthew M Davis, Kathryn Taubert, Andrea L Benin, David W Brown, George A Mensah, Larry M Baddour, Sandra Dunbar, and Harlan M Krumholz. Influenza vaccination as secondary prevention for cardiovascular disease: a science advisory from the american heart association/american college of cardiology. *Journal of the American College of Cardiology*, 48(7):1498–1502, 2006.

Exploring Joint Disease Risk Prediction

Xiang Wang, PhD, Fei Wang, PhD, Jianying Hu, PhD, Robert Sorrentino, MD
IBM T. J. Watson Research Center, Yorktown Heights, NY
{wangxi,fwang,jyhu,sorrentino}@us.ibm.com

Abstract

Disease risk prediction has been a central topic of medical informatics. Although various risk prediction models have been studied in the literature, the vast majority were designed to be single-task, i.e. they only consider one target disease at a time. This becomes a limitation when in practice we are dealing with two or more diseases that are related to each other in terms of sharing common comorbidities, symptoms, risk factors, etc., because single-task prediction models are not equipped to identify these associations across different tasks. In this paper we address this limitation by exploring the application of multi-task learning framework to joint disease risk prediction. Specifically, we characterize the disease relatedness by assuming that the risk predictors underlying these diseases have overlap. We develop an optimization-based formulation that can simultaneously predict the risk for all diseases and learn the shared predictors. Our model is applied to a real Electronic Health Record (EHR) database with 7,839 patients, among which 1,127 developed Congestive Heart Failure (CHF) and 477 developed Chronic Obstructive Pulmonary Disease (COPD). We demonstrate that a properly designed multi-task learning algorithm is viable for joint disease risk prediction and it can discover clinical insights that single-task models would overlook.

1 Introduction

Disease risk prediction [1] has been extensively studied in the literature. The latest trends include building risk prediction models based on a large amount of features from Electronic Health Record (EHR) databases and adopting state-of-the-art machine learning algorithms, such as Generalized Linear Regression, Support Vector Machine, Bayesian Networks, etc. [2, 3, 4]. What many existing risk prediction models have in common is that they are designed to be single-task, i.e. they only predict the risk of a single disease at a time. With a given target disease, a single-task risk model would select features that are most informative for this particular target, and then train the model using the training data set.

However, in practice we often have training data for multiple different yet related target diseases at the same time, e.g. hypertension and heart diseases, diabetes and cataract, depression and obesity, etc. In these application scenarios, single-task risk models have two significant limitations. First, by treating different target diseases separately, they fail to identify the underlying correlation between these diseases such as their common causes, similar symptoms, comorbid conditions, and distinguishing factors. These information are sometimes more important to clinical practitioners than risk prediction itself because they lead to insights on the underlying mechanisms of diseases. Second, applying a single-task model limits us to the training data that have been labeled for that particular disease, even though the training data from other related diseases can also be helpful. For instance, a single-task model will not be able to augment a small training data set labeled for heart failure with another large training data set labeled for hypertension, even though it is common knowledge that hypertension is closely related to heart failure and share many important risk predicators.

In the machine learning literature multi-task learning has been extensively studied [5, 6, 7]. However, existing multi-task learning techniques cannot be directly applied to the problem of EHR-based risk prediction because the validity of each algorithm relies on the specific assumption it makes about task relatedness and these assumptions often fail to hold for many clinical applications. For example, some models assume different tasks are close to each other as if they are derived from the same underlying distribution [7, 8], or alternatively, assume the tasks have group structure and are similar within each group [9, 10, 11]. Other models formalize task relatedness by assuming all tasks share a latent feature space [6, 5, 12]. Our proposed framework falls into this category.

In this paper we explore the viability of designing a multi-task framework for EHR-based risk prediction. In order to do this, we first need to define the relatedness between multiple diseases. In particular, we make the assumption that

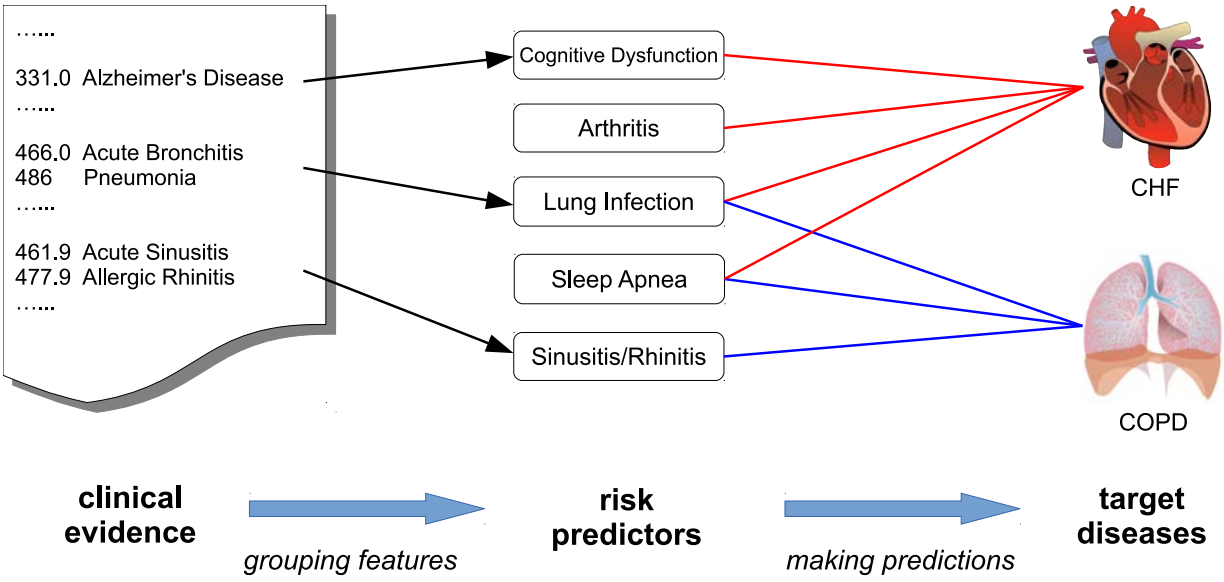


Figure 1: An intuitive example to demonstrate the problem setting for multi-task risk prediction. The lines between target diseases and risk predictors indicate strong connection. The arrows show the mapping from the raw clinical evidence (diagnosis code in this case) to the high-level risk predictors. We want to learn both the lines and the arrows.

if two (or more) diseases are related then they must share some common risk predictors and these risk predictors can be characterized by groups of the clinical evidence from the EHR database.

To intuitively understand what our assumption implies in practice, consider a group of individuals who are at risk of two diseases: Congestive Heart Failure (CHF) and Chronic Obstructive Pulmonary Disease (COPD). Traditional risk models attribute the risks directly to the raw medical features from the EHR database, such as individual diagnosis codes, lab results, vitals, etc., which are often noisy and sparse [13]. Under our framework, we attribute the risks to some higher-level latent risk predictors, which are modeled as groups of the raw medical features. As show in Figure 1, our two target diseases, CHF and COPD, share some common risk predictors such as sleep apnea, hypertension, respiratory system infection [14, 15, 16, 17]. Therefore, studying them jointly will help us more accurately pinpoint these underlying predictors and consequently facilitate the risk prediction. In addition, these two diseases also have their own risk predictors, such as Rhinitis for COPD and Arthritis for CHF. It will be beneficial to identify such predictors because they can help us better distinguish these two diseases with very similar symptoms and comorbidities [18].

Given our mild assumption, which will hold for a wide range of diseases and EHR data, our goal becomes how to jointly identify these common predictors across different diseases. To do so, we develop an optimization based formulation that simultaneously learns the feature groups and predicts the risk for all diseases based on the identified predictors. We show how to solve our objective function efficiently using an alternating minimization algorithm. To validate our proposed framework, we apply it to a real EHR database with 7,839 patients, among which 1,127 developed Congestive Heart Failure and 477 developed Chronic Obstructive Pulmonary Disease. By using diagnosis codes as underlying features, we demonstrate that our model is able to identify a meaningful set of shared risk predictors for CHF and COPD and good prediction accuracy ensues.

2 Study Design

We chose two tasks for joint prediction: CHF onset and COPD onset. CHF and COPD are well known to have significant overlap in terms of common comorbidities, risk factors, and symptoms [14, 15, 16]. In fact they are so

similar that in practice they are often misdiagnosed for each other [18]. Thus it is highly desirable if we could risk stratify them jointly and identify not only the common predictors that they share but also the unique predictors that distinguish them.

Our study is on a real-world EHR data warehouse including the records of 319,650 patients over 4 years. We identified 1,127 CHF case patients with 3,850 control match, and 477 COPD case patients with 2,385 control match. We extracted the same features for both cohorts, namely ICD-9 diagnosis codes.

2.1 CHF Cohort Construction

We defined CHF diagnosis using the following criteria (which are similar to the criteria used in [3]): (1) ICD-9 diagnosis of heart failure appeared in the EHR for two outpatient encounters, indicating consistency in clinical assessment; (2) at least one medication were prescribed with an associated ICD-9 diagnosis of heart failure. The diagnosis date was defined as the first appearance in the EHR of the heart failure diagnosis. These criteria have also been previously validated as part of Geisinger Clinical involvement in a Centers for Medicare & Medicaid Services (CMS) pay-for-performance pilot [19]. With this criteria, we extracted from the database 1,127 CHF case patients. Following the case-control match strategy in [3], a primary care patient was eligible as a control patient if they are not in the case list, and had the same PCP as the case patient. Approximately 10 eligible clinic-, sex-, and age-matched (in five-year age intervals) controls were selected for each heart failure case. In situations where 10 matches were not available, all available matches were selected. Following this strategy, we got CHF 3,850 control patients, so on average each case patient was matched with approximately 3 controls.

2.2 COPD Cohort Construction

We defined COPD diagnosis also using two criteria: (1) the occurrence of at least one COPD-related ICD-9 diagnosis code; (2) the prescription of at least one COPD-related medication. The diagnosis date was defined as the date when both criteria were met. In the end we identified 477 COPD case patients. We matched these case patients by identifying control patients were similar in age, sex, and PCP. An eligible control patient should also have a valid ICD-9 outpatient diagnosis which is not (a) a symptom code (b) a screening diagnosis, or (c) a COPD-related diagnosis. In the end we identified 2,385 control patients, which gave us a 1:5 case-control match.

2.3 Feature Extraction

For all patients we extracted their ICD-9 codes from the EHR database. We only considered the medical records that occurred from 540 days prior to the diagnosis date till 180 days prior to the diagnosis date. In other words, we used about a year worth of data to make prediction at least half a year before the disease onset. Patients who had insufficient amount of records were not included. For control patients, we set the last day of their available records as the diagnosis date and followed the same rule. In total there were 4,784 unique ICD-9 codes from all patients. After removing infrequent features, i.e. ICD-9 codes that occurred to fewer than 100 different patients (case and control combined), we had 267 distinct features left. In total 80,472 medical records were considered, which indicates our input data was extremely sparse. After features were extracted, we discarded the temporal order within the observation window and used binary encoding to record wether or not a certain feature was assigned to a certain patient during that time.

3 Model Derivation

Objective: Suppose we have T target diseases (tasks), D features from the EHR database, and N_t patients for the t-th task. For each task we have an observation matrix $X_t \in \mathbb{R}^{D\times N_t}$. The (j,i)-th entry of X_t denotes the occurrence of feature j to patient i. $\mathbf{x}_{t_i}$ is the i-th column of X_t. $\mathbf{y}_t \in \{-1,1\}^{N_t}$ is the response vector for task t: $\mathbf{y}_{t_i} = 1$ means patient i is diagnosed disease t, -1 otherwise. $U \in \{0,1\}^{D\times K}$ is a mapping from the D medical features to K groups. The rows of U sum up to one, which means each feature belongs to one group. $\mathbf{w}_t \in \mathbb{R}^K$ is the regression coefficients

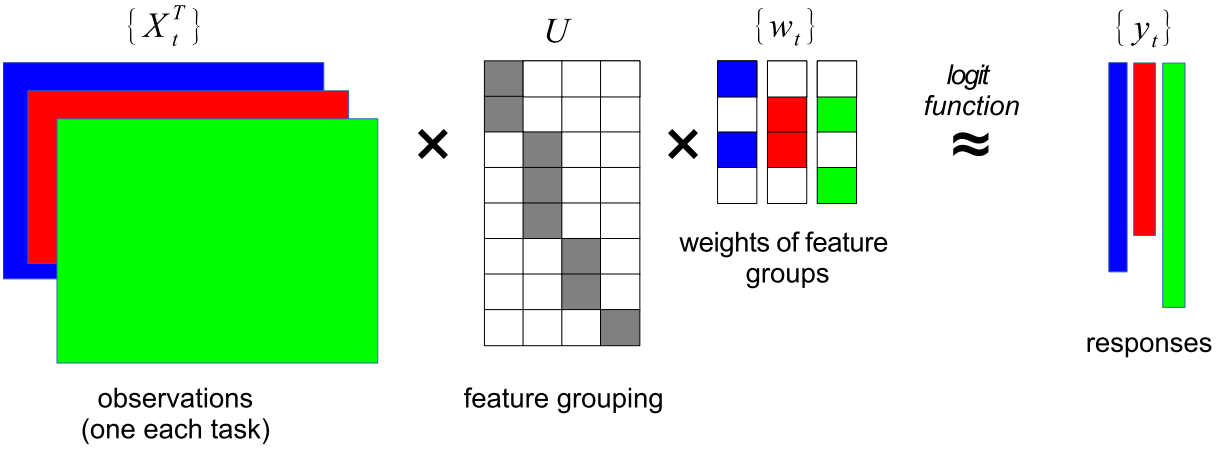


Figure 2: An illustration of our multi-task learning framework. The clinical features from the EHR database are assigned into feature groups, as defined by the assignment matrix U that is shared across all diseases. $\{X_t, \mathbf{y}_t\}$ are input and we want to learn U and $\{\mathbf{w}_t\}$. In this illustration, $D = 8, K = 4, T = 3$.

over the K feature groups for the t-th disease. A positive entry in $\mathbf{w}_t$ means that feature group contributes positively to the risk of disease t and vice versa. Figure 2 is an illustration of our framework.

Our objective is to learn the feature grouping U and the regression coefficients $\{\mathbf{w}_t\}$ simultaneously and jointly over T diseases. Formally it can be written as:

$$\underset{\{\mathbf{w}_t \in \mathbb{R}^K\}, U \in \{0,1\}^{D \times K}}{\operatorname{argmin}} \sum_{t=1}^{T} \left(\frac{1}{N_t} \sum_{i=1}^{N_t} \log(1 + \mathbf{e}^{-\mathbf{y}_{t_i} \mathbf{x}_{t_i}^T U \mathbf{w}_t}) + \lambda \|\mathbf{w}_t\|_1 \right) \quad \text{s.t.} \sum_{k=1}^{K} U_{dk} = 1, \forall d = 1, \ldots, D \tag{1}$$

where $\|\cdot\|_1$ is element-wise $\ell 1$ norm $\|A\|_1 = \sum_{i,j} |A_{ij}|$, $\lambda > 0$ is a user-specified parameter.

Interpretation: The first term inside the summation of Eq.(1) is the empirical loss. Here we choose the logistic loss, one that is mostly commonly used for clinical risk models. The second term is a regularizer that enforces sparsity on the regression coefficients $\mathbf{w}_t$. Intuitively this term wants each disease to be explained by a smaller number of groups (thus a simpler explanation). Additional regularizers can be added according to the practical needs. The constraint term in Eq.(1) says the rows of U should sum up to 1, which implies the K feature groups are a disjoint partition of the D medical features. This is to make the feature groups semantically distinct.

Eq.(1) is intractable due to the combinatorial nature of U. To overcome this, we relax the constraint on U by allowing the entries in U to take real values. After the relaxation, our objective becomes:

$$\underset{\{\mathbf{w}_t \in \mathbb{R}^K\}, U \in \{0,1\}^{D \times K}}{\operatorname{argmin}} \sum_{t=1}^{T} \left(\frac{1}{N_t} \sum_{i=1}^{N_t} \log(1 + \mathbf{e}^{-\mathbf{y}_{t_i} \mathbf{x}_{t_i}^T U \mathbf{w}_t}) + \lambda \|\mathbf{w}_t\|_1 \right) \quad \text{s.t. } U^T U = I_K \tag{2}$$

Note that the orthogonality constraint now replaces the original constraint in Eq.(1) to enforce distinction among different groups. Eq.(2) now allows an efficient solution.

Optimization: To solve the objective function in Eq.(2), we alternate between U and $\{\mathbf{w}_t\}$ by fixing one and updating the other to minimize Eq.(2) until a local optimum is reached. When U is fixed, Eq.(2) becomes:

$$\underset{\{\mathbf{w}_t \in \mathbb{R}^K\}}{\operatorname{argmin}} \sum_{t=1}^{T} \left(\frac{1}{N_t} \sum_{i=1}^{N_t} \log(1 + \mathbf{e}^{-\mathbf{y}_{t_i} \tilde{x}_{t_i}^T \mathbf{w}_t}) + \lambda \|\mathbf{w}_t\|_1 \right) \tag{3}$$

where $\tilde{x}_{t_i}^T = \mathbf{x}_{t_i}^T U$. This is a set of T standard $\ell 1$-regularized logistic regression problems [20] and can be solved independently using a variety of ready-to-use solvers.

Table 1: Feature groups identified by our model with top-5 features with highest weights ($K = 3$).

Weights	ICD-9	Description
Feature Group 1: Predictors shared by CHF and COPD ($\mathbf{w}_{CHF} = 2.176, \mathbf{w}_{COPD} = 1.390$)		
0.329	715	Osteoarthrosis and Allied Disorders
0.257	729	Disorders of Soft Tissues
0.246	724	Disorders of Back
0.229	V72	Special Investigations and Examinations
0.227	719	Cardiac Dysrhythmias
Feature Group 2: Predictors mainly associated with CHF ($\mathbf{w}_{CHF} = 1.739, \mathbf{w}_{COPD} = 0.767$)		
0.248	427	Cardiac Dysrhythmias
0.224	250	Diabetes Mellitus
0.196	414	Chronic Ischemic Heart Disease
0.120	429	Ill-Defined Descriptions of Heart Disease
0.109	411	Acute Ischemic Heart Disease
Feature Group 3: Predictors mainly associated with COPD ($\mathbf{w}_{CHF} = 0.082, \mathbf{w}_{COPD} = 5.084$)		
0.318	493	Asthma
0.171	592	Kidney Stones
0.162	388	Disorders of Ear
0.148	461	Acute Sinusitis
0.144	305	Tobacco Use Disorder

When $\{\mathbf{w}_t\}$ is fixed, Eq.(2) becomes

$$\underset{U \in \mathbb{R}^{D \times K}}{\operatorname{argmin}} \sum_{t=1}^{T} \frac{1}{N_t} \sum_{i=1}^{N_t} \log(1 + \mathbf{e}^{-\mathbf{y}_{t_i} \mathbf{x}_{t_i}^T U \mathbf{w}_t}), \text{ s.t. } U^T U = I_K \tag{4}$$

We solve this subproblem by using the Augmented Lagrange Multipliers method, which minimizes the following Lagrangian of Eq.(4):

$$F(U, \Lambda) = \sum_{t=1}^{T} \frac{1}{N_t} \sum_{i=1}^{N_t} \log(1 + \mathbf{e}^{-\mathbf{y}_{t_i} \mathbf{x}_{t_i}^T U \mathbf{w}_t}) + \text{tr}(\Lambda(U^T U - I_K)) + \frac{\rho}{2} \|U^T U - I_K\|_F^2 \tag{5}$$

where $\Lambda \in \mathbb{R}^{K \times K}$ is the Lagrange multipliers and $\rho > 0$ is a given constant.

4 Results

4.1 Identified Feature Groups

The key difference between our multi-task model and the single-task models is that our model uses feature groups instead of individual raw features as predictors. Therefore we first examine the feature groups identified from the real patient cohort for their clinical validity. An important parameter of our model is K, the number of the feature groups we want to project the raw features into. Generally speaking, K should increase with the number of tasks T to ensure enough descriptive power. Here since we only have two tasks, we set $K = 3$ for the clarity of presentation. Also for the sake of clarity, we pre-grouped ICD-9 codes into ICD-9 group codes, i.e. the first 3 digits. For each feature group we display in Table 1 the top ICD-9 group codes according to their weights, U_{ki}. A larger weight indicates this ICD-9 group code is associated more strongly with the feature group. For each feature group we also display its regression coefficients $\mathbf{w}_t$ for both tasks, from which we can tell how much this feature group as a predictor contributes to the risk of respective diseases.

The first group in Table 1 consists of ICD-9 codes that are associated with both CHF and COPD. We can see that they are mainly musculoskeletal disorders. These are common problems that can be caused either by CHF or COPD. The

second group consists of diagnosis that are mainly associated with CHF, such as heart arrhythmia, diabetes, ischemic heart diseases, and so on. These are all known risk factors for CHF. As a contrast, the third feature group shows diagnosis that are mainly associated with COPD, where we can find tobacco use, a leading risk factor for COPD, as well as asthma, a major comorbid condition for COPD.

Table 1 suggests that our model is indeed able to identify significant risk predictors across different tasks and group them based on whether they are shared by different tasks or only belongs to a specific task. However, we observe that there are still overlapping between the feature groups (e.g. 427 Cardiac Dysrhythmias in Group 1 and 2). This calls for more detailed investigation by physicians because the same ICD-9 (group) code can have different implications in different medical contexts. Therefore there is no simple "ground truth" on whether a specific feature should be associated with a specific target disease.

Note that after our relaxation from exclusive grouping to orthogonality constraint, every feature will appear not only in one feature group but in all feature groups with different weights (and the weights can be either positive or negative). Therefore, for each feature we need to consider the sum of its weights across all groups before we can interpret its overall contribution to the risk of task t.

4.2 Prediction Accuracy

Next we evaluate the performance of our approach in terms of prediction accuracy. The measurement we used was Area Under Receiver Operating Characteristic Curve (AUC) [21], which is a commonly used evaluation metric for risk prediction models. An AUC score of 1 means the prediction perfectly matches the ground truth whereas 0.5 means the prediction is no better than a random guess. We used 10-fold cross validation and report the median, the 25th, and the 75th percentiles as boxplot in Figure 3. For each fold we sampled 60% of the entire dataset for training and used the rest for validation.

We compare our approach (Multi-Risk) to two baseline methods. The first one is denoted Single-PCA. Instead of learning U jointly from all diseases, Single-PCA used a fixed U derived from the top-K principal components of all observed. Single-PCA represents the result we get in the single-task setting where the feature groupings are learned without supervision. The second baseline is denoted Single-Raw, which means U is set to be an identity matrix I_D. This is an extreme case of our framework where $K = D$ and all the raw features are used directly for logistic regression (without grouping). For all methods we used the same parameter settings for logistic regression. For Single-PCA and our model (Multi-Risk) we set $K = 3$.

From Figure 3 we observe that our model significantly outperformed Single-PCA after joint feature grouping. This is expected because the groupings are learned with supervision to maximize the discriminative power. On the other hand, the AUCs of our model and Single-Raw were comparable (no significant difference). However, we would like to point out that Single-Raw is designed to achieve the highest prediction accuracy possible without considering the potential association between different diseases, whereas our model can capture common predictors across different diseases without sacrificing prediction accuracy.

5 Discussion

This work is a first step towards multi-task clinical risk prediction. Our main purpose is to demonstrate that a properly designed multi-task learning objective can indeed capture the latent relatedness between different target diseases. It provides a principled way of identifying the common predictors and the discriminative predictors, both of which are clinically interesting. We also show that the identified feature groups can achieve the state-of-the-art prediction accuracy.

In our experiment, the two tasks shared a same training set. Namely, we have only one patient cohort and each patient was labeled for both tasks. This setting does not fully demonstrate the advantage of a multi-task risk model, which can combine disjoint patient cohorts from different tasks. For example, the CHF task can have one patient cohort labeled only for CHF from one database and the COPD task can have a different patient cohort (with different patients) labeled only for COPD. The only requirement is that these two cohorts must share the same feature space. Such augmentation

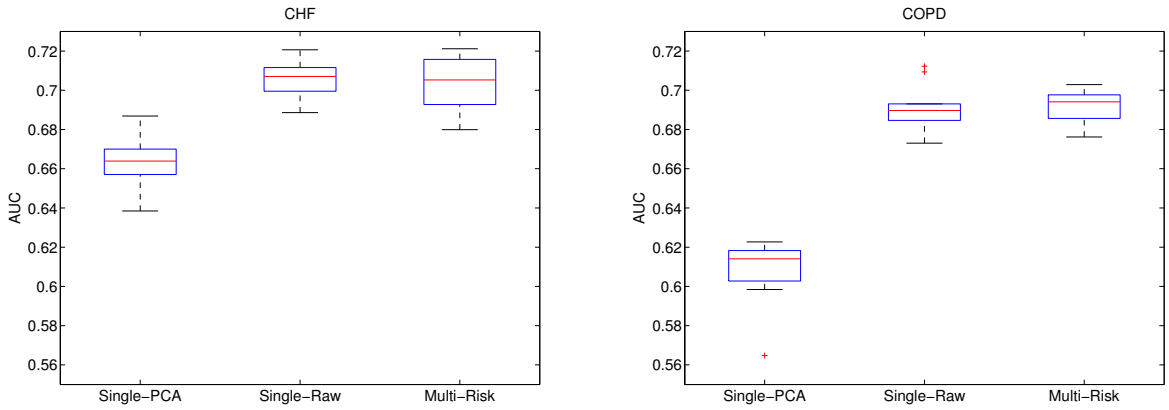


Figure 3: Accuracy of our model (Multi-Risk) for both tasks with comparison to two baseline methods. Showing the median, 25th, and 75th percentiles over 10-fold cross validation.

of the training data can potential further improve the risk prediction accuracy of our multi-task model and gives it advantage over the state-of-the-art single-task models. We leave this to future investigation.

Our multi-task learning framework can be extended in many aspects. For instance, it will be very interesting to jointly study multiple tasks from different problem domains. In our experiments, we used two similar tasks (disease onset prediction) that share the same feature set (ICD-9 codes). However, our framework can be extended to incorporate tasks from different domains, e.g. disease onset vs. hospitalization vs. need for social assistance, by grouping heterogeneous features together (diagnosis codes, medications, social behaviors, and so on). This will be a big step forward because we will be able introduce predictors from different domains for more comprehensive risk analysis.

A second direction to extend our model is to incorporate prior knowledge from domain experts. Currently our model only groups features based on how they contribute to the respective outcomes. It is not aware the clinical meaning of each feature, therefore the groups are not coherent in terms of clinical interpretation. This can be overcome if we incorporate prior knowledge into the model, which specifies the semantics of each group. For instance, domain experts can assign a few ICD-9 codes related to cardiovascular diseases to a certain group and let our model automatically identify the remaining ICD-9 codes that should also belong to the same group. Such integration can be achieved by carefully initializing U and/or introducing an additional term to regularize key entries in U.

In order to make our objective function tractable, we used the orthogonality constraint as a surrogate for exclusive group partition. The side effect of this relaxation was that U will have negative weights, which are not clinically interpretable. To overcome this, we can replace the orthogonality constraint with non-negativity constraint, i.e. $U \geq 0$. This change calls for a different optimization algorithm but produces more interpretable results.

6 Conclusion

In this work we explore a multi-task framework for joint disease risk prediction. Our framework exploits the assumption that related diseases share some common risk predictors that can be represented by groups of clinical evidence. We use the proposed model to simultaneously predict the onset risk of a CHF cohort and a COPD cohort. Preliminary results suggest that our model can identify both common and discriminative risk predictors for both diseases while pertaining good prediction accuracy. We discussed the potential of using our model to integrate patient cohorts from multiple sources and problem domains. We also discussed future improvements that can be made to enhance the interpretability of our model.

References

[1] Miller CC, Reardon MJ, Safi HJ. Risk stratification: a practical guide for clinicians. Cambridge University Press; 2001.

[2] Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet. 2012 Jun;13(6):395–405.

[3] Wu J, Roy J, Stewart WF. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. Med Care. 2010 Jun;48(6 Suppl):S106–113.

[4] Kennedy EH, Wiitala WL, Hayward RA, Sussman JB. Improved cardiovascular risk prediction using nonparametric regression and electronic health record data. Med Care. 2013 Mar;51(3):251–258.

[5] Argyriou A, Evgeniou T, Pontil M. Convex multi-task feature learning. Machine Learning. 2008;73(3):243–272.

[6] Ando RK, Zhang T. A Framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data. Journal of Machine Learning Research. 2005;6:1817–1853.

[7] Evgeniou T, Pontil M. Regularized multi–task learning. In: KDD; 2004. p. 109–117.

[8] Yu K, Tresp V, Schwaighofer A. Learning Gaussian processes from multiple tasks. In: ICML; 2005. p. 1012–1019.

[9] Kumar A, Daumé III H. Learning Task Grouping and Overlap in Multi-task Learning. In: ICML; 2012. .

[10] Kang Z, Grauman K, Sha F. Learning with Whom to Share in Multi-task Feature Learning. In: ICML; 2011. p. 521–528.

[11] Zhou J, Chen J, Ye J. Clustered Multi-Task Learning Via Alternating Structure Optimization. In: NIPS; 2011. p. 702–710.

[12] Zhou J, Yuan L, Liu J, Ye J. A multi-task learning formulation for predicting disease progression. In: KDD; 2011. p. 814–822.

[13] Lasko TA, Denny JC, Levy MA. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. PLoS ONE. 2013;8(6):e66341.

[14] Angermann C. Comorbidities in heart failure: a key issue. European Journal of Heart Failure Supplements. 2009;8(suppl 1):i5–i10.

[15] Lang CC, Mancini DM. Non-cardiac comorbidities in chronic heart failure. Heart. 2007 Jun;93(6):665–671.

[16] Decramer M, Janssens W, Miravitlles M. Chronic obstructive pulmonary disease. Lancet. 2012 Apr;379(9823):1341–1351.

[17] Baty F, Putora PM, Isenring B, Blum T, Brutsche M. Comorbidities and burden of COPD: a population based case-control study. PLoS ONE. 2013;8(5):e63285.

[18] Hawkins NM, Petrie MC, Jhund PS, Chalmers GW, Dunn FG, McMurray JJ. Heart failure and chronic obstructive pulmonary disease: diagnostic pitfalls and epidemiology. Eur J Heart Fail. 2009 Feb;11(2):130–139.

[19] Pfisterer M, Buser P, Rickli H, Gutmann M, Erne P, Rickenbacher P, et al. BNP-guided vs symptom-guided heart failure therapy. JAMA: the journal of the American Medical Association. 2009;301(4):383–392.

[20] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.; 2001.

[21] Zou KH, O'Malley AJ, Mauri L. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. Circulation. 2007 Feb;115(5):654–657.

Clinical Decision Support for Whole Genome Sequence Information Leveraging a Service-Oriented Architecture: a Prototype

Brandon M. Welch, MS, PhD[1,2],* Salvador Rodriguez-Loya[3],* Karen Eilbeck, MS, PhD[2], Kensaku Kawamoto, MD, PhD[2]
Medical University of South Carolina, Charleston, SC; University of Utah, Salt Lake City, UT; University of Sussex, East Sussex, United Kingdom
***co-first authors**

Abstract

Whole genome sequence (WGS) information could soon be routinely available to clinicians to support the personalized care of their patients. At such time, clinical decision support (CDS) integrated into the clinical workflow will likely be necessary to support genome-guided clinical care. Nevertheless, developing CDS capabilities for WGS information presents many unique challenges that need to be overcome for such approaches to be effective. In this manuscript, we describe the development of a prototype CDS system that is capable of providing genome-guided CDS at the point of care and within the clinical workflow. To demonstrate the functionality of this prototype, we implemented a clinical scenario of a hypothetical patient at high risk for Lynch Syndrome based on his genomic information. We demonstrate that this system can effectively use service-oriented architecture principles and standards-based components to deliver point of care CDS for WGS information in real-time.

Introduction

Having a patient's whole genome sequence (WGS) information available to guide clinical decision-making has been an important goal of genome research, as WGS information could be used to improve clinical diagnosis, guide preventative efforts, and inform therapeutic decisions in the clinic[1]. Indeed, several clinical examples illustrate how WGS information has been used for the clinical diagnosis and treatment of patients with rare or previously undiagnosed diseases[2–4]. A patient's WGS information available at the point of care can increase a clinician's capacity to practice personalized medicine in a routine clinical care setting[5]. In fact, as a result of the increasing availability of sequencing technology, as well as exponentially declining costs of obtaining one's genomic information, the application of WGS information to routine clinical care is becoming increasingly possible[6].

The need for clinical decision support for WGS

Although WGS information has the potential to be a valuable resource for clinical decision-making, its effective application to routine clinical care will likely be hindered by a number of challenges. Examples of such challenges include (1) static laboratory reports intended for human consumption, (2) the complexity of genetic analysis, (3) limited physician proficiency in genetics, and (4) the lack of genetics professionals in the clinical workforce[7]. Nevertheless, these challenges could be overcome by means of clinical decision support (CDS). CDS entails providing clinicians, patients, and other healthcare stakeholders with pertinent knowledge and/or person-specific information, intelligently filtered or presented at appropriate times, to enhance health and health care[8]. CDS provided within the clinical workflow, and at the point and time of decision-making within the electronic health record (EHR), has been shown to be the most effective way to deliver CDS to clinicians[9]. To realize the potential of genome-guided clinical care, CDS leveraging a patient's WGS information must be provided in real-time within the clinical workflow and the EHR[7].

Desiderata for integrating genomic information with the EHR and CDS

To address the complexity and challenges of integrating genomic information with EHRs, Masys *et al.* developed a set of desired technical requirements for EHRs to support the integration of genomic information[10]. See Table 1 for the list of these requirements.

Table 1. Desiderata for the integration of genomic data into EHRs described by Masys *et al*.

1.	Maintain separation of primary molecular observations from the clinical interpretations of those data
2.	Support lossless data compression from primary molecular observations to clinically manageable subsets
3.	Maintain linkage of molecular observations to the laboratory methods used to generate them
4.	Support compact representation of clinically actionable subsets for optimal performance
5.	Simultaneously support human-viewable formats and machine-readable formats in order to facilitate implementation of decision support rules
6.	Anticipate fundamental changes in the understanding of human molecular variation
7.	Support both individual clinical care and discovery science

Welch *et al.* developed an additional set of requirements, extending the Masys *et al.* desiderata, specifically addressing the technical requirements related to CDS for WGS information[11]. See Table 2 for a list of these requirements.

Table 2. Additional desiderata for the technical integration of WGS with CDS described by Welch *et al.*

8.	CDS knowledge must have the potential to incorporate multiple genes and clinical information.
9.	Keep CDS knowledge separate from variant classification.
10.	CDS knowledge must have the capacity to support multiple EHR platforms with various data representations with minimal modification.
11.	Support a large number of gene variants while simplifying the CDS knowledge to the extent possible.
12.	Leverage current and developing CDS and genomics infrastructure and standards.
13.	Support a CDS knowledge base deployed at and developed by multiple independent organizations.
14.	Access and transmit only the genomic information necessary for CDS.

Both of these efforts are aimed at providing guidance to system developers who are developing health IT capabilities for WGS information.

Proposed CDS architecture for WGS information

To satisfy the desiderata requirements described above, we previously proposed the use of a service-oriented architecture (SOA) to provide automatic CDS for WGS information at the point of care[12]. SOA is a software design approach which uses separate, independent software components known as services, which are self-contained components that have well-defined, understood capabilities[13]. SOA facilitates the reusability and standardization of processes, allowing for independent evolution and modifications to a particular service, reducing the burden of change on the overall system. Indeed, SOA offers many advantages to health information technology and CDS; as a result, its use is growing in health care[14]. Figure 1 shows a diagram of the SOA architecture proposed and described in further detail by Welch *et al*.[12]

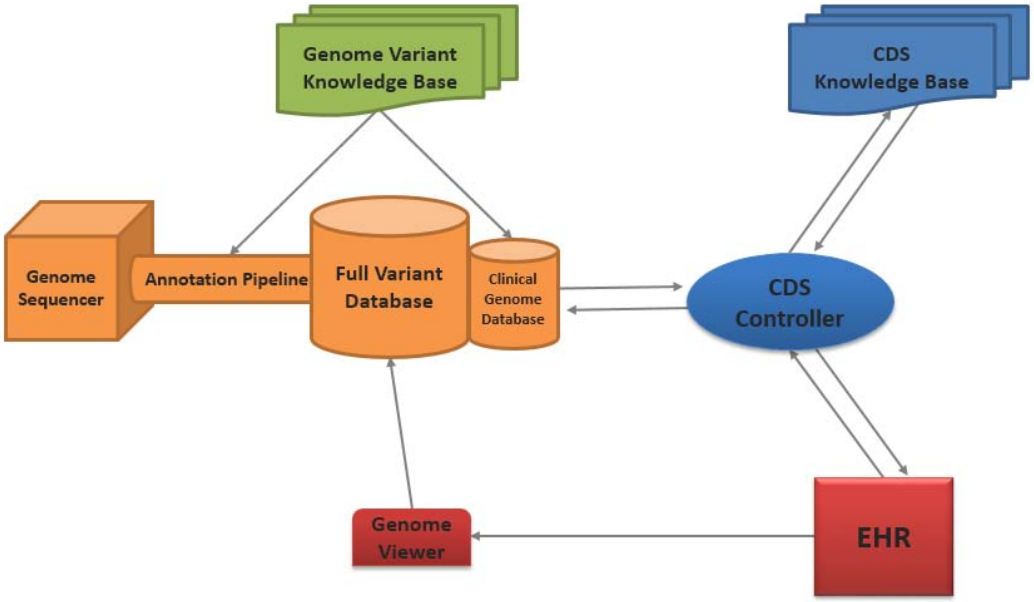


Figure 1: Overview of proposed CDS architecture for WGS information implemented as a prototype in this paper

A brief summary of each component is provided here:

- *Genome sequencing and annotation*- This component of the architecture is responsible for making the patient's genome information available and providing the genome with relevant annotations (e.g. location of variants relative to genes, impact of variants on genes, etc.). The annotation process leverages variant knowledge contained in the genome variant knowledge base.
- *Genome variant knowledge base*- This is a knowledge repository of known genome variants and assigned clinical interpretations. The primary responsibility of these knowledge repositories is to maintain the most up-to-date and accurate variant interpretations, which is important as variant interpretations are known to change over time.[15]
- *Genome database*- In the proposed architecture, a patient's genome is stored in a genome database, separate from clinical information (e.g. EHR). This database securely stores patients' genomes along with the most up-to-date variant interpretations from the genome variant knowledge base. This database also provides a standardized interface for access to a patient's genomic information.
- *CDS knowledge base*- The role of the CDS knowledge base is to process the patient-specific clinical and genomic information provided to it. It subsequently returns patient-specific, knowledge-based recommendations and/or information to support clinical care.
- *Electronic health record*- The EHR provides the patient's clinical information used by the CDS knowledge base. The EHR is also responsible for actions which trigger CDS requests and the presentation of CDS results within the clinical workflow.
- *CDS controller*- The CDS controller is responsible for processing a CDS request and assuring that all required data for CDS knowledge is available. If additional patient data is needed for CDS, the CDS controller makes a request to other data repositories (e.g. genome database) for the required information.

This manuscript describes our effort to build a functional prototype of the previously proposed CDS architecture for WGS information[12]. In this paper, we describe the open-source components and health IT standards used to develop this prototype. Furthermore, we describe the methods used to evaluate the prototype using a clinical use case involving a patient at increased risk for Lynch Syndrome. Finally, we have identified areas that will require additional research and development in the future. While others have built CDS capabilities for genomics[16–18], to our knowledge, this work is the first to describe a system that adheres to the technical desiderata[10,11] and uses a SOA approach to provide WGS-guided CDS within the clinical workflow and within the EHR.

Materials and Methods

Components used and configuration

Genome data acquisition and annotation

To obtain real WGS data used in this prototype, we used the 10Gen data set, which represents the first ten publicly available human genomes in a standardized genome variation format[19]. These genomes represent three different ethnicities (African, Asian, and Caucasian) and several sequencing platforms including SOLiD, Illumina, Sanger, Roche 454, CGenomics, and Helicos[20]. To prepare these genomes for clinical use, they were annotated using the Web-based Omicia Opal genome annotation platform.

Genome database

To persistently store a patient's annotated genomic information and make it available for CDS data requests, we implemented a relational database using MySQL Community Server (version 5.6.15) and the HeidiSQL open-source database manager[21]. In this database, we created a table named 'patient_genome' consisting of seven columns (with column names in quotes): (1) 'MRN' which is the patient's unique medical record number (MRN) that matches the patient's MRN in the Tolven open-source EHR; (2) 'gene' is the gene where the variant resides, represented using HUGO (Human Genome Organization) Gene Nomenclature Committee (HGNC) standardized nomenclature; (3) 'refSNP' is the reference SNP ID number; (4)'nuc_var' is the nucleotide variant represented in Human Genome Variation Society (HGVS) nomenclature; (5) 'pro_var' is the protein variant also represented in HGVS nomenclature; (6) 'interpretation' is the clinical impact of the variant provided by the ClinVar genome variant knowledge base; and (7) 'id' is an auto-incremented value for the primary key of the table. The 10Gen genomes were exported from Omicia as CSV files and then imported into the database through a database import process available through HeidiSQL.

To support external access to a patient's genomic data stored in the database, we developed a Web service interface deployed as a Java application within the JBoss Enterprise Application Platform version 6.1. This Web service

interface provides access to the database content using the HL7 Retrieve, Locate, and Update Service (RLUS) standard[22]. The RLUS specification can be adapted to different semantic content formats. For the present prototype, we used the HL7 Virtual Medical Record (vMR) as the semantic content format within the RLUS-based Web service.

ClinVar genome variant knowledge base

We used ClinVar as the genome variant knowledge base in this prototype. ClinVar is a publically available repository of human genome sequence variations and associated phenotypes supported by the National Center for Biotechnology Information of the U.S. National Library of Medicine[23]. ClinVar currently does not support service calls to its knowledge base, so we replicated this functionality by developing a second table in the genome database called 'genomekb' and imported a subset of ClinVar data into this table. A ClinVar full data release was downloaded directly from the ClinVar FTP site to a Linux server and transformed into a format suitable for database import using XSLT[24]. We created a SQL statement to update the interpretation field in the 'patient_genome' table from the interpretation field in the 'genomekb' table, based upon matches in the gene and variant between the two tables.

Tolven electronic health record

To store the patient's clinical information and provide a clinical workflow interface for CDS, we used the open-source Tolven electronic Clinician Health Record (eCHR™), which is provided as part of the Tolven Platform.[25] Tolven eCHR™ is an Office of the National Coordinator (ONC) certified clinical information system which supports basic clinical processes and information exchange. The Tolven Platform supports several additional components including ePrescribing, scheduling, and analytics. Tolven was selected because of (1) its open-source code; (2) its ability to be configured and customized; (3) its data model, which is based on the HL7 Reference Information Model; and (4) its use of several terminology standards such as Current Procedural Terminology (CPT), Logical Observation Identifiers Names and Codes (LOINC), and RxNorm. Furthermore, third-party plugins can be developed to extend the functionality of the Tolven Platform. We developed a plugin for Tolven using Java Enterprise Edition (version 1.6). The plugin is designed to serve three important purposes for the prototype: (1) to access the patient's clinical data stored in Tolven; (2) to create a vMR document containing the patient's data; and (3) to communicate with an external CDS Web service using the HL7 Decision Support Service (DSS) standard.

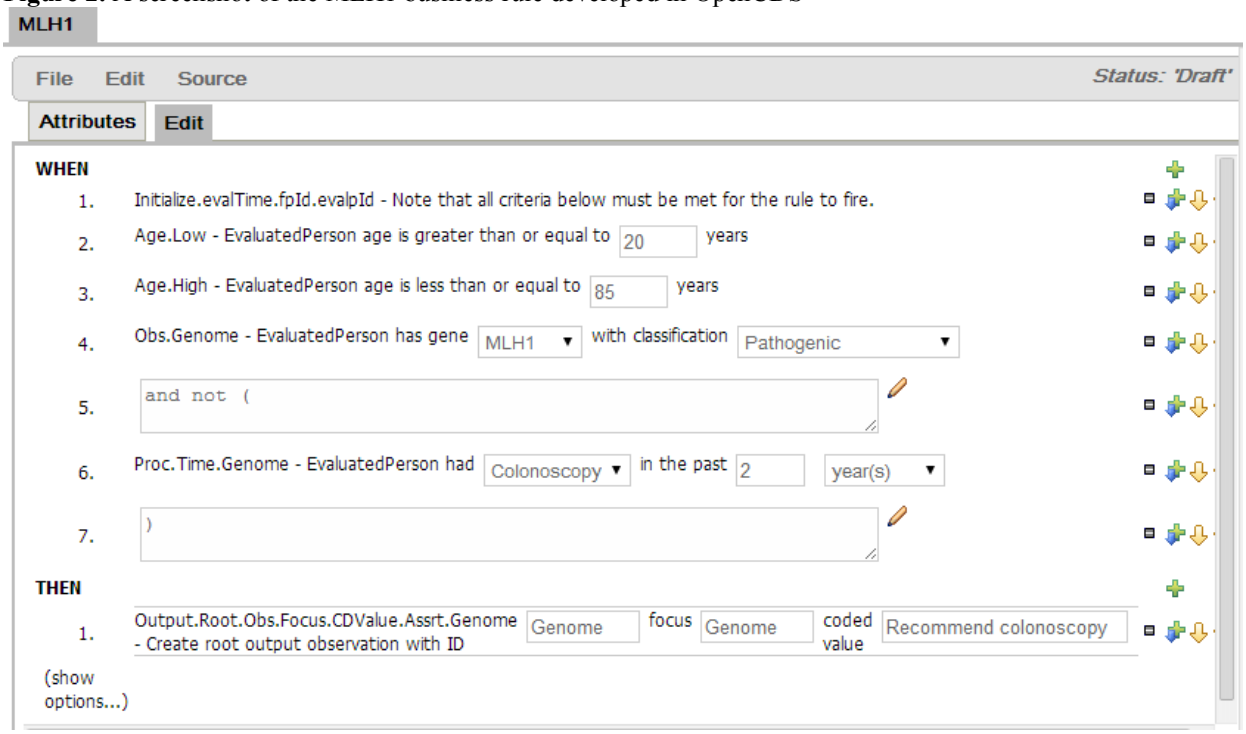
SwitchYard CDS controller

The CDS controller plays a central coordinating role in this architecture. It first processes a CDS request from Tolven and checks that all required data for CDS inferencing is available. In our prototype, it identifies that the genomic data is not available, so it makes a data request to the genome database for additional information. We used the open source solution SwitchYard (JBoss) to provide these functions. SwitchYard is a component-based SOA development framework which combines several useful components and functions for SOA into one application, including Apache Camel, Java Enterprise Edition, business process management (BPM), orchestration, routing, validation, and transformation. Within SwitchYard, we developed a composite service that includes five components: CamelServiceRoute, ProcessComponent, VerifyGenomeData, RequestGenomeData and IntegrateGenomeData. We also configured two Web service adaptors for SwitchYard to request and receive data from other components of the CDS architecture. These include (1) an adaptor for the HL7 RLUS standard to interface with the genome database Web service and (2) an adaptor for the HL7 DSS standard to interface with the Tolven EHR and OpenCDS.

OpenCDS

For the final component of the architecture, we used OpenCDS (http://www.OpenCDS.org) to serve as the CDS knowledge base. OpenCDS is an open-source, Java-based CDS framework designed to enable the delivery of CDS as a Web service compliant with the HL7 DSS standard. OpenCDS provides a knowledge authoring environment and CDS rules engine. Internally, OpenCDS uses the vMR as its data model and JBoss Drools as its inferencing engine. OpenCDS also manages and uses its own terminology within the Apelon Distributed Terminology System (DTS), with mappings created to standard terminologies such as SNOMED CT and LOINC. For this prototype, we deployed a new instance of OpenCDS and created the CDS rule logic in the Web-based knowledge authoring environment for Drools known as Guvnor. Figure 2 shows a CDS business rule authored in Guvnor and used in this prototype. After the CDS rule was created and tested in Guvnor, it was deployed within the OpenCDS run-time environment. Several vocabulary terms required by the rules were added to the OpenCDS terminology in the Apelon DTS instance used by OpenCDS.

Figure 2. A screenshot of the MLH1 business rule developed in OpenCDS



Standards Used

HL7 Decision Support Service standard

The DSS standard specifies a standard interface for providing CDS as a service and is adopted by both HL7 and the Object Management Group (OMG) standards development organizations[26]. The DSS standard includes three major interfaces including (1) *evaluation*, used to evaluate patient data and generate patient-specific results; (2) *metadata discovery*, to identify metadata and knowledge modules of a service; and (3) *query*, which is used to query for knowledge modules of interest. For this prototype, we used the evaluateAtSpecifiedTime request interface, which includes a payload section in which a base64 encoded version of the vMR is placed.

HL7 Virtual Medical Record standard

The Virtual Medical Record (vMR) is a standard HL7 clinical data model designed for CDS[27]. The vMR data model can represent patient-specific classes such as demographics, encounters, procedures, problems, medications, laboratory results, and observations. A patient's clinical data is modeled using these various vMR classes and elements. To represent the patient's genome information, we used the *ObservationResult* class. Within this class, *observationFocus* was used to represent the gene name in HGNC format, *observationValue* was used to represent the nucleotide variant in HGVS format, and *interpretation* was used to represent the variant clinical interpretation in LOINC.

HL7 Retrieve, Locate, and Update Service

The HL7 Retrieve, Locate, and Update Service (RLUS) defines the service interface to locate, retrieve, and update resources among and within healthcare organizations[22]. This specification was designed to support SOA in healthcare. The HL7 RLUS specification defines several methods including *describe*, *discard*, *get*, *initialize*, *list*, *locate*, and *put*. In our prototype, we used RLUS to request a patient's genomic information from the genome database using the *describe* and *get* methods. *Describe* returns a detailed schema definition and *get* retrieves patient data based on parameters supplied in the RLUS retrieval request.

Gene and variant nomenclature

Several genome standards were used to represent genomic information in this prototype. To represent genes, we used names approved by HGNC[28]. We used three types of standardized genome variant representations in the prototype:

(1) the refSNP number assigned to a sequence variant by dbSNP[29]; (2) the nucleotide variant in HGVS nomenclature format for coding sequence[30]; and (3) the protein variant in HGVS nomenclature for protein variation. Finally, we used LOINC to represent the possible interpretations of gene variants, which include pathogenic, presumed pathogenic, unknown significance, benign, and presumed benign[31].

The CDS process

The previous sections describe the components used in the prototype and their configurations. This section describes the overall process for providing patient-specific CDS using WGS information within the EHR. With all components of the prototype in place and functioning properly, the sequential steps of the application are as follows:

1. A patient's clinical data is recorded in Tolven by the clinician. When the patient's chart is modified and saved, a CDS request is triggered by the OpenCDS plugin in Tolven. The plugin proceeds to gather the patient's clinical data from Tolven and transform the data into the vMR format. When the patient's clinical data is all in vMR format, it is base64-encoded, placed into the DSS payload, and sent to the SwitchYard CDS controller.
2. When SwitchYard receives the DSS request with clinical data in vMR format, it validates the data against a vMR schema template defining the required data for the requested CDS knowledge module. In this case, SwitchYard identifies that the genomic data required by the CDS module is missing. It then creates a RLUS request to obtain that required information from the genome database.
3. The genome database Web service interface receives the RLUS request from SwitchYard. This interface retrieves the requested information from the genome database, which includes the gene and clinical interpretation (the return of specific variants are also available upon request). The interface creates a response RLUS message and inserts a vMR payload of the genomic information.
4. SwitchYard receives the RLUS response from the genome Web service interface and merges the genome vMR with the clinical data vMR from Tolven. If all the required data is present, a DSS request with the newly merged vMR is sent to OpenCDS.
5. OpenCDS receives the DSS request and processes the data contained in the vMR against rules in the requested knowledge module (MLH1). A CDS result is produced and sent back to SwitchYard in a DSS response.
6. SwitchYard receives the response from OpenCDS and forwards the message to Tolven.
7. When the Tolven plugin receives the response, it renders the response as an alert within the Tolven user interface, such that the CDS is provided within the clinical workflow and at the time of decision-making.

Clinical use case

To assess whether the prototype is functional, we implemented the following clinical scenario and evaluated the performance of the prototype.

Clinical use case: Lynch Syndrome risk assessment

Lynch syndrome (or hereditary nonpolyposis colorectal cancer) is an autosomal dominant genetic condition caused by pathogenic mutations in mismatch repair genes such as MSH2, MLH1, MSH6, and PMS2[32]. These mutations impact the ability of a cell to repair DNA replication errors that occur during cell division. As the cell continues to divide, these mutations continue to accumulate throughout the genome. These mutations may ultimately lead to uncontrolled cell proliferation and thus cancer. Patients who have a pathogenic variant in any of these genes possess the greatest risk for colorectal cancer, and they are also at increased risk for cancer in the stomach, intestines, liver, brain, skin and other body sites. Every year in the U.S., approximately 4,000-7,000 new cases of colorectal cancer are caused by Lynch Syndrome[32]. It is recommended that patients at increased risk for Lynch syndrome should receive a colonoscopy every one to two years starting around the age of 20 years[33]. While family history is a strong indicator for risk, genetic testing is the definitive test for a patient's Lynch Syndrome risk. Genetic screening for the disease, although currently expensive, will likely become easier when WGS information is clinically available for routine care[34]. Even then, some clinicians may not be familiar with Lynch Syndrome nor the best practices for risk mitigation and management. CDS can play an important role in the awareness and management of Lynch Syndrome.

Implementation of the clinical use case

To assess the ability of this prototype architecture to provide point-of-care CDS for Lynch Syndrome, we created and implemented the example clinical scenario found in Box 1. To implement this clinical scenario in the prototype, we created a new 32 year old patient in Tolven. For simplicity and demonstration purposes, we tested the architecture

using a single gene (MLH1) with a pathogenic variant. We assigned the publically available Venter genome to the test patient, which did not have a pathogenic mutation in the MLH1 gene[35]. Therefore, we created a known pathogenic MLH1 mutation (NG_007109.2:g.32058C>T). This genome change was created previous to a genome knowledge base update of the 'interpretation' database field in the patient_genome table. A CDS business rule representing the MLH1 recommendation for colonoscopy was created in Guvnor and deployed on the OpenCDS run time environment (see Figure 2). For simplicity, the CDS query is triggered automatically each time the patient's record is modified in Tolven.

Box 1: High-risk disease risk assessment clinical scenario

> The patient is a 32 years old male, with no personal history of colon cancer and no prior colonoscopies. The patient does not know his family health history. This patient previously had his genome sequenced, and this WGS information is available for assessment by CDS. This patient has a pathogenic mutation in the MLH1 gene, a gene associated with a high risk for Lynch Syndrome. The clinical recommendation established by his healthcare organization is to recommend that patients over 20 years old with pathogenic mutations in the MLH1 gene receive a recommendation if one has not been performed in the past 1-2 years.

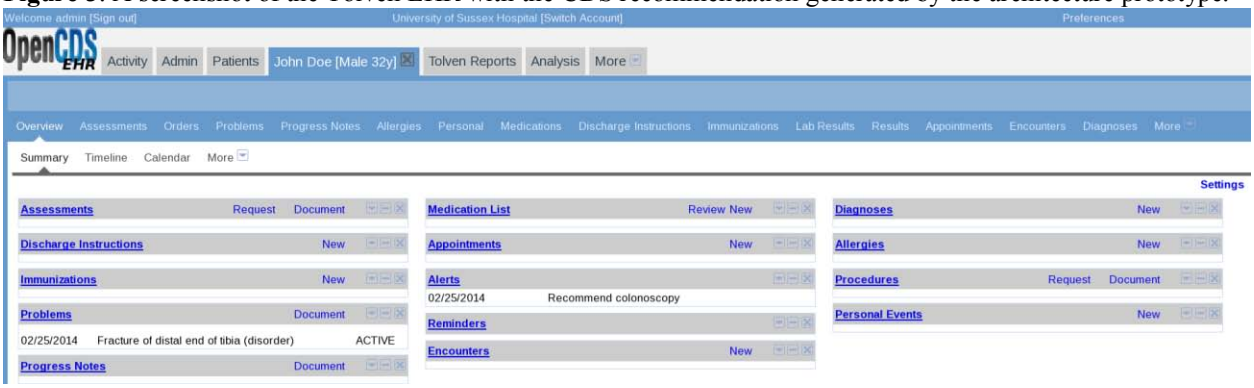
Performance evaluation

To test the performance of the architecture we used the free, open-source load testing software LoadUI (www.loadui.org) running on a Windows 7 (64-bit) machine with two processors and four gigabytes of RAM. In order to identify sources of latency, we tested each service component of the architecture separately, which included the genome application service running on a Linux Ubuntu 12.04 (64-bit) server with eight processors and 16 gigabytes of RAM, as well as the OpenCDS service running on a Linux CentOS v5.8 (64-bit) with four processors and four gigabytes of RAM. We also tested the performance of the SwitchYard CDS Controller (also running on the Linux Ubuntu server previously described) which included service calls to both the genome application and OpenCDS within its evaluation. As such, the evaluation of the CDS controller most closely represents the overall performance of the architecture. Each component was tested with a random load balance of 100 simultaneous users every 10 seconds using the same data requirements described in the clinical scenario. The overall performance evaluation was limited to 20 simultaneous users as a result of performance issues related to running Switchyard on a machine with limited processor capabilities.

Results

The objective of this study was to develop a functional prototype of our proposed architecture[12]. To that end, we implemented a prototype and verified that a patient with a clinical scenario described in Box 1 was evaluated properly by the prototype. Figure 3 shows a screenshot of the CDS recommendation successfully generated through the use of this prototype architecture within the Tolven EHR.

Figure 3. A screenshot of the Tolven EHR with the CDS recommendation generated by the architecture prototype.



Architecture performance

The genome application (which includes the service interface and database request) handled 3,109 requests over a five minute period. This genome application's fastest request took 25 milliseconds (ms), and the slowest took 697 ms, with an average of 40 ms (SD 47.44 ms). OpenCDS handled 3,015 requests over a five minute period. The fastest request took 7 ms, and the slowest took 914 ms, with an average of 12 ms (SD 17.04 ms). We also evaluated the performance of the CDS controller, which includes service calls to the genome application and OpenCDS and thus closely represents the overall performance of the architecture. Due to hardware limitations of our machine, we limited the

testing to 20 simultaneous users. The CDS controller handled 650 requests over a five-minute period. The fastest request took 356 ms the slowest took 4,243 ms, with an average of 944 ms (SD 621.04). It is important to note that the average response time was under one second. See Table 4 for a performance summary.

Table 4: Performance results of the architecture components using LoadUI

Component	Simultaneous Users	Total requests handled	Min request time (time in ms)	Max request time (time in ms)	Average request time (time in ms)	Standard deviation
Genome app	100	3109	25	697	**40**	47.77
OpenCDS	100	3015	7	914	**12**	17.04
Overall	20	650	356	4243	**944**	621.04

Discussion

As demonstrated by this prototype and demonstration, a service-oriented CDS architecture is able to support the provision of genome-guided CDS at the point of care within the EHR. Indeed, this prototype leverages standards and open-source solutions, and it is capable of integrating with current health IT architectures and workflows.

Issues identified

Throughout the process of developing the prototype of this architecture, several issues were identified. We used ClinVar as the genome variant knowledge base for this prototype because it was easily accessible and publically available. Unfortunately, ClinVar is still in a very early stage with regards to its variant knowledge base, limiting its ability to provide clinical interpretations for relevant variants. Specifically, the available variant knowledge in ClinVar is a small, but growing, subset of all available variant knowledge[36,37]. Nevertheless, it is expected that ClinVar will improve over time as more laboratories begin contributing their variant knowledge. Also, variants reported to ClinVar can range in quality on a scale from one to five stars, ranging from a single submitter's interpretation to variants reviewed and submitted by expert panels. This variable quality of variant interpretation or potential disagreement between submitters can be challenging for CDS to manage. Ideally, variant interpretations used in CDS should meet a minimum threshold of quality and confidence. A recent NIH-funded initiative called ClinGen is expected to improve the data available in ClinVar[38]. Although other genome variant knowledge bases are also available (such as the Human Genome Mutation Database), we did not integrate them with this prototype due to the time and cost associated with each integration. Integrating such variant knowledge bases into the architecture is an anticipated future effort. Nevertheless, once ClinVar becomes more developed, we believe it represents an ideal resource for genome variant knowledge management because of its public financing and its potential to become the largest single repository of genome variant knowledge[12].

Strengths and limitations

An important strength of this work is the use of freely available, open-source components. Indeed, anyone with sufficient training in the technologies used could rebuild this architecture without needing to purchase and use proprietary software or components. Another strength of the architecture is the use of available health IT standards wherever possible. Such standards and terminologies include the HL7 DSS standard, HL7 vMR standard, LOINC, HGNC, and HGVS. As a result, we were able to demonstrate that standards-based approaches could be used to deliver WGS CDS which could be leveraged to support interoperability with other health IT systems in the future.

A limitation of the study is that we only demonstrated this architecture working using a simple scenario in one clinical use case (Lynch Syndrome risk assessment). While, there are many other clinical and genomic scenarios (such as pharmacogenomics) that could be tested and demonstrated with this architecture, the scope of this effort was limited to assessing whether the proposed prototype could deliver standards-based CDS within a hypothetical clinical workflow and within the EHR. Nevertheless, we are planning to demonstrate the extensibility of the architecture with various clinical genomic use cases in future work. Another limitation of the study is that we have only integrated the CDS architecture with a single EHR using a simple CDS event trigger. While the primary aim of the current effort was to demonstrate minimum feasibility, future efforts will need to focus on integrating the architecture with other EHRs with more intuitive event triggers. Of note, the integration of SOA CDS capabilities into various EHRs is part of larger OpenCDS initiatives, outside the scope of the current research on CDS for WGS information.

Future directions

While we were able to demonstrate feasibility of this architecture using open-source solutions and standards, it took a significant amount of configuration of these components to make this possible. As a result, a future direction of this

research is to develop this CDS architecture into an easily deployable format for healthcare organizations to set up and run with minimal modification. Such a solution could potentially consist of a virtual machine image which includes a preconfigured genome database already linked to several genome variant knowledge bases, OpenCDS configured for genome rule authoring, and SwitchYard set up to support WGS-enabled CDS. Furthermore, it will be important to develop and include integration plugins for commercial EHR systems. Indeed, EHR systems are starting to support service capabilities, and future Meaningful Use guidelines may require all EHRs to support them, meaning this architecture has the potential to be used on a widespread basis for WGS CDS in the near future[39]. Likewise, as standards-based CDS integration with EHRs becomes more widespread, it will be important to test the functionality with end-users and conduct clinical trials evaluating the clinical impact of genome-guided CDS using this approach.

Another important future effort is to build the genomic CDS knowledge base and expand the current genome database. As this architecture supports a CDS knowledge base that can serve multiple health care organizations, it may be possible to develop a genomic CDS knowledge base which could be shared among these organizations. Such an effort could involve the collaboration of informaticists, geneticists, and clinical experts to implement published clinical genomics guidelines into a computable CDS knowledge representation. Furthermore, such experts could also review the literature and current clinical practices to develop new genomic CDS knowledge for clinical care. During this processes of developing a genomic CDS knowledge base, additional data requirements for inclusion in the genome database will likely be identified. For example, in the current version of the genome database, only genes and variants are included. While these data may be sufficient for Lynch Syndrome and other autosomal dominant use cases, these data would not be sufficient for autosomal recessive genetic conditions and other types of genomic use cases (e.g. SNP-based testing). As a result, it will be necessary to add additional genomic information for CDS, such as chromosome number, zygosity, tandem repeat number, or sequencing quality scores. These additional genomic data requirements could potentially be identified though a systematic review of the literature and the involvement of various domain experts.

Conclusion

Using WGS information in the clinic for routine clinical care may be challenging for clinicians to manage without assistance. CDS provided within the clinical workflow, at the time of decision-making, provides a feasible solution to enable genetically-guided personalized medicine. To evaluate this potential solution, we developed and tested a functional CDS architecture for WGS information using SOA design principles. Through this effort, we were able to demonstrate that a functional prototype of this approach is capable of providing genome-guided CDS within a hypothetical clinical workflow and within the EHR. While future research and development is necessary before such an approach can be used in a clinical setting, this study demonstrates that the approach is feasible and valid. We therefore speculate that this work will help guide future research and development on the use of WGS-based CDS to support personalized healthcare.

References

1. Collins FS, Green ED, Guttmacher AE, Guyer MS. A vision for the future of genomics research. *Nature*. 2003;422(6934):835–47. doi:10.1038/nature01626.
2. Ashley EA, Butte AJ, Wheeler MT, et al. Clinical assessment incorporating a personal genome. *Lancet*. 2010;375(9725):1525–35. doi:10.1016/S0140-6736(10)60452-7.
3. Talkowski ME, Ordulu Z, Pillalamarri V, et al. Clinical diagnosis by whole-genome sequencing of a prenatal sample. *N Engl J Med*. 2012;367(23):2226–32. doi:10.1056/NEJMoa1208594.
4. Lupski JR, Reid JG, Gonzaga-Jauregui C, et al. Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N Engl J Med*. 2010;362(13):1181–91. doi:10.1056/NEJMoa0908094.
5. President's Council of Advisors on Science and Technology. *Priorities for Personalized Medicine*.; 2008.
6. Wetterstrand K. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). 2013. Available at: http://www.genome.gov/sequencingcosts/. Accessed February 6, 2013.
7. Welch BM, Kawamoto K. The Need for Clinical Decision Support Integrated with the Electronic Health Record for the Clinical Application of Whole Genome Sequencing Information. *J Pers Med*. 2013;3:306–325. doi:10.3390/jpm20x000x.
8. Osheroff JA, Teich JM, Middleton B, Steen EB, Wright A, Detmer DE. A roadmap for national action on clinical decision support. *J Am Med Informatics Assoc JAMIA*. 2007;14(2):141–145. doi:10.1197/jamia.M2334.Introduction.
9. Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *BMJ*. 2005;330(7494):765. doi:10.1136/bmj.38398.500764.8F.
10. Masys DR, Jarvik GP, Abernethy NF, et al. Technical desiderata for the integration of genomic data into Electronic Health Records. *J Biomed Inform*. 2012;45(3):419–22. doi:10.1016/j.jbi.2011.12.005.

11. Welch BM, Eilbeck K, Del Fiol G, Meyer L, Kawamoto K. Technical desiderata for the integration of genomic data with clinical decision support. *J Biomed Inform. 2014 Jun 12. pii: S1532-0464(14)00139-7. doi: 10.1016/j.jbi.2014.05.014*
12. Welch BM, Rodriguez-Loya S, Eilbeck K, Kawamoto K. A Proposed Clinical Decision Support Architecture Capable of Supporting Whole Genome Sequence Information. *J. Pers. Med. 2014, 4(2), 176-199; doi:10.3390/jpm4020176*x.
13. Erl T. *Service-Oriented Architecture (SOA): Concepts, Technology, and Design*. Prentice Hall; 2005:1–792. Available at: http://www.amazon.com/Service-Oriented-Architecture-SOA-Concepts-Technology/dp/0131858580. Accessed December 3, 2013.
14. Kawamoto K, Lobach DF. Design, implementation, use, and preliminary evaluation of SEBASTIAN, a standards-based Web service for clinical decision support. *AMIA Annu Symp Proc*. 2005;(Xml):380–4. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1560495&tool=pmcentrez&rendertype=abstract.
15. Aronson SJ, Clark EH, Varugheese M, Baxter S, Babb LJ, Rehm HL. Communicating new knowledge on previously reported genetic variants. *Genet Med*. 2012;14(8):713–719. doi:10.1038/gim.2012.19.
16. Welch BM, Kawamoto K. Clinical decision support for genetically guided personalized medicine: a systematic review. *J Am Med Inform Assoc*. 2012;20(2):388–400. doi:10.1136/amiajnl-2012-000892.
17. Tarczy-Hornoch P, Amendola L, Aronson SJ, et al. A survey of informatics approaches to whole-exome and whole-genome clinical reporting in the electronic health record. *Genet Med*. 2013;15(10):824–32. doi:10.1038/gim.2013.120.
18. Overby CL, Kohane I, Kannry JL, et al. Opportunities for genomic clinical decision support interventions. *Genet Med*. 2013;15(10):817–23. doi:10.1038/gim.2013.128.
19. Reese MG, Moore B, Batchelor C, et al. A standard variation file format for human genome sequences. *Genome Biol*. 2010;11(8):R88. doi:10.1186/gb-2010-11-8-r88.
20. The Sequence Ontology - Resources - 10Gen Data Set. Available at: http://www.sequenceontology.org/resources/10Gen.html. Accessed January 23, 2014.
21. HeidiSQL - MySQL and MSSQL made easy. Available at: http://www.heidisql.com/. Accessed January 23, 2014.
22. HL7 Standards Product Brief - HL7 Version 3 Standard: Retrieve, Locate, and Update Service (RLUS) Release 1. Available at: http://www.hl7.org/implement/standards/product_brief.cfm?product_id=89. Accessed February 3, 2014.
23. National Center for Biotechnology Information. ClinVar. *US Natl Libr Med*. 2012. Available at: http://www.ncbi.nlm.nih.gov/clinvar/. Accessed February 8, 2013.
24. ClinVar. ClinVarFullRelease_2014-01.xml.gz. 2014. Available at: ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/xml/. Accessed July 1, 2014.
25. Tolven Platform. Available at: http://home.tolven.org/. Accessed February 26, 2014.
26. HL7 Standards Product Brief - HL7 Decision Support Service (DSS). Available at: https://www.hl7.org/implement/standards/product_brief.cfm?product_id=12. Accessed February 3, 2014.
27. Kawamoto K, Del Fiol G, Strasberg HR, et al. Multi-National, Multi-Institutional Analysis of Clinical Decision Support Data Needs to Inform Development of the HL7 Virtual Medical Record Standard. *AMIA Annu Symp Proc*. 2010;2010:377–81. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3041317&tool=pmcentrez&rendertype=abstract.
28. Gray KA, Daugherty LC, Gordon SM, Seal RL, Wright MW, Bruford EA. Genenames.org: the HGNC resources in 2013. *Nucleic Acids Res*. 2013;41(Database issue):D545–52. doi:10.1093/nar/gks1066.
29. dbSNP Home Page. Available at: http://www.ncbi.nlm.nih.gov/SNP/. Accessed February 3, 2014.
30. Horaitis O, Cotton RGH. The challenge of documenting mutation across the genome: the human genome variation society approach. *Hum Mutat*. 2004;23(5):447–52. doi:10.1002/humu.20038.
31. Logical Observation Identifiers Names and Codes (LOINC®) — LOINC. Available at: http://loinc.org/. Accessed February 3, 2014.
32. Lynch syndrome. 2014. Available at: http://ghr.nlm.nih.gov/condition/lynch-syndrome. Accessed March 10, 2014.
33. Lynch Syndrome Management. Available at: http://genefacts.org/index.php?option=com_content&view=article&id=492:management&catid=113:lynch-syndrome&Itemid=665. Accessed March 10, 2014.
34. Marquez E, Geng Z, Pass S, et al. Implementation of routine screening for Lynch syndrome in university and safety-net health system settings: successes and challenges. *Genet Med*. 2013;15(12):925–32. doi:10.1038/gim.2013.45.
35. Levy S, Sutton G, Ng PC, et al. The diploid genome sequence of an individual human. *PLoS Biol*. 2007;5(10):e254. doi:10.1371/journal.pbio.0050254.
36. Peterson T a, Doughty E, Kann MG. Towards precision medicine: advances in computational approaches for the analysis of human variants. *J Mol Biol*. 2013;425(21):4047–63. doi:10.1016/j.jmb.2013.08.008.
37. Riggs ER, Wain KE, Riethmaier D, et al. Towards a Universal Clinical Genomics Database: the 2012 International Standards for Cytogenomic Arrays Consortium Meeting. *Hum Mutat*. 2013;34(6):915–9. doi:10.1002/humu.22306.
38. New NIH-funded resource focuses on use of genomic variants in medical care. Available at: http://www.nih.gov/news/health/sep2013/nhgri-25.htm. Accessed October 8, 2013.
39. Regulations.gov - Proposed Rule Document. Available at: http://www.regulations.gov/#!documentDetail;D=HHS-OS-2014-0002-0001. Accessed March 12, 2014.

Stochastic Gradient Descent and the Prediction of MeSH for PubMed Records

W. John Wilbur MD, PhD and Won Kim PhD
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, U.S.A.

Abstract

Stochastic Gradient Descent (SGD) has gained popularity for solving large scale supervised machine learning problems. It provides a rapid method for minimizing a number of loss functions and is applicable to Support Vector Machine (SVM) and Logistic optimizations. However SGD does not provide a convenient stopping criterion. Generally an optimal number of iterations over the data may be determined using held out data. Here we compare stopping predictions based on held out data with simply stopping at a fixed number of iterations and show that the latter works as well as the former for a number of commonly studied text classification problems. In particular fixed stopping works well for MeSH® predictions on PubMed® records. We also surveyed the published algorithms for SVM learning on large data sets, and chose three for comparison: PROBE, SVMperf, and Liblinear and compared them with SGD with a fixed number of iterations. We find SGD with a fixed number of iterations performs as well as these alternative methods and is much faster to compute. As an application we made SGD-SVM predictions for all MeSH terms and used the Pool Adjacent Violators (PAV) algorithm to convert these predictions to probabilities. Such probabilistic predictions lead to ranked MeSH term predictions superior to previously published results on two test sets.

Introduction

The National Library of Medicine (NLM) produces the PubMed database of biomedical journal article citation records consisting of title, abstract (where available) and appropriate metadata. This data base currently contains over 23 million records and is growing at about 60,000 records per month. Most of these records have approximately a dozen key terms assigned to them from a controlled vocabulary known as Medical Subject Headings (MeSH). There are over 27,000 MeSH terms from which these key terms or index terms are assigned by humans. This involves significant human effort and expense. It is then natural that efforts would be made to mechanize some of this work. Such efforts have been of two types. First, there has been an ongoing effort to develop a system called the Medical Text Indexer (MTI) to predict MeSH assignments as suggestions for MeSH indexers at the NLM[1-9]. Second, there have been a number of investigations of machine learning methods and how they might be applied to the MeSH indexing problem in a more abstract setting with the purpose of understanding and comparing different approaches to this problem[6, 10-15]. The work we report here is of this latter type. We examine the Stochastic Gradient Descent (SGD) method for solving Support Vector Machines (SVMs)[16, 17] and develop an approach which we believe has broad applicability but is especially attractive for solving very large SVM problems. We show how to use this approach to improve MeSH suggestions over prior published work.

SGD has proved to be a very effective method of training machine learning algorithms[16, 17]. It has generally been found to confer a significant decrease in training time without sacrificing accuracy[17-19]. SGD can be applied to standard convex loss functions with regularization terms with good effect, but Zhang[17] suggested using SGD without the usual regularization term and performing the regularization with early stopping. This has become a widely practiced approach[20-22] and is implemented by dividing the training set into disjoint pieces consisting of a new training set and a validation set. Then on each training pass through the new training set one tests on the validation set and that number of iterations over the training data which is found to be optimal is recorded. One then trains on the original training set with this number of iterations and evaluates the results on the test set to rate the method. One of our contributions in this work is to show that on a number of text classification problems one can implement early stopping by just using a constant number of iterations and obtain the same performance as one obtains using the validation set approach.

In particular we find that SGD with a fixed number of eight iterations over the training data provides MeSH classification results as good as early stopping based on held out data and as good as several other popular methods which take much longer to train. Given the efficiency of the method we are able to apply it to each MeSH heading to make predictions. It would then be possible to use the SVM scores of all the MeSH terms for a given document to

rank the predictions for that document. However, different MeSH terms have different frequencies in the training data and this leads to classifier scores that are not directly comparable, i.e., one does not obtain the best result with such a ranking approach based on raw scores. This leads us to take a slightly different approach. We first divide the training data into two equal halves. For a given MeSH term we train a classifier on each half. We then apply the Pool Adjacent Violators (PAV) algorithm[23-26] to each classifier's scoring of the half of the training data on which it was not trained. The PAV algorithm converts raw scores to probabilities. As a consequence, for any previously unseen test document, we can apply the classifiers to obtain two raw scores and use the PAV models to convert these raw scores to probabilities. We average the two probabilities to obtain a single probability estimate for that MeSH term for the test document. Since probabilities are optimally comparable we use the resulting probabilities over all MeSH terms to make ranked MeSH term predictions. We find such predictions superior to predictions obtained by raw score ranking and also superior to previous published predictions.

Methods

The SGD Algorithm with Early Stopping Assume we are given N training points $\{(x_i, y_i)\}_{i=1}^{N}$ randomly and independently sampled from the same source distribution, where for each i, $x_i \in R^n$ and $y_i \in \{-1,1\}$. The objective is to learn a weight vector $w \in R^n$ so that the function

$$f_w(x) = \text{sgn}(w \cdot x) \tag{1}$$

has a high probability of agreeing with y if (x, y) is randomly sampled from the same source as the training data. Here the function sgn is known as the sign function and has the definition

$$\text{sgn}(x) = \begin{cases} 1, & x > 0 \\ 0, & x = 0 \\ -1, & x < 0 \end{cases}. \tag{2}$$

Then pseudocode for the SGD algorithm without regularization, but with early stopping is as follows.

SGD without regularization

Input: training data $\{(x_i, y_i)\}_{i=1}^{N}$; learning rate $\lambda > 0$.

Initialize: $w_0 = 0$; $t = 0$.

A. Randomly sample j, $1 \le j \le N$

B. If $y_j(w_t \cdot x_j) < 1$ set $w_{t+1} = w_t + \lambda y_j x_j$ and $t = t + 1$.

C. if stopping criterion satisfied return w_t else return to A.

Several issues require comment here. First, we do not deal with a threshold independent of the vectors x_i but assume that all vectors have a common added dimension which is also present in w and functions as threshold. Second, we do not randomly sample integers from the interval $[1, N]$ as in step A of the algorithm. Instead we do the training in rounds or passes over the training data and before each pass we re-randomize the order in which the integers in $[1, N]$ are visited. Third, we use a learning rate $\lambda = 0.002$. On all three of these points we are simply following[17]. Our pseudocode does not define the stopping criterion. Generally cross validation or held out data is used to determine a stopping point, but our purpose is to compare that approach to simply stopping at a fixed number of passes over the data without doing any cross validation.

Method of SGD Experiments By holding out training data, we may experimentally determine the optimal number of iterations of SGD for SVM with early stopping (L1 loss Minimization without regularization, hereafter we call it SGD-SVM)[20-22]. Let $D = \{(x_n, y_n), n = 1, \cdots, N\}$ be our data set, where x_n is a vector representing the

attributes of the nth instance and y_n is the binary class value of x_n. We randomly split the data into J equal parts $D_1, D_2 \cdots D_J$. Let D_j and $D^{(-j)} = D - D_j$ be the test and training sets for the jth fold of a J fold cross validation. For the jth fold we again randomly split the training set $D^{(-j)}$ into K equal parts $D_1^{(-j)} \cdots D_K^{(-j)}$. Defining $D_K^{(-j)}$ to be the held out or validation set, we may determine the optimal number of iterations of SVM-SGD. We train SGD-SVM on $D^{(-j)} - D_K^{(-j)}$ and let $Iter_j$ be the optimal iteration number for SGD-SVM evaluated on $D_K^{(-j)}$. We then apply SVM-SGD learning with $Iter_j$ iterations on $D^{(-j)}$, and measure the success of this learning on D^j. Note that random splits are done in such a way as to keep the number of negative records and the number of positive records the same or as close to the same as possible over all splits. The J different test results from the J different folds are combined by macro-averaging , which we believe is appropriate when all J experiments are very similar. We report mean average precision (MAP) and break even (BE) values[27] and macro-averaging means we compute the measure for each fold and then take a simple average over all folds for the final value reported. The break even value is sometimes also known as R-precision[27].

When we do the experiment without using held out data the protocol is simpler. As before, we randomly split the data into J equal parts $D_1, D_2 \cdots D_J$ and let D_j and $D^{(-j)} = D - D_j$ be the test and training sets for the jth fold of a J fold cross validation. But now $Iter_j$ is a fixed number for all j. One may ask how do we decide what this number should be. Our answer is empirical. We have observed what works well on the data from much experience based on cross validations and other experiments where we simply run the learning algorithm for a number of iterations and watch the performance. Our conclusions are surprisingly simple. For all the MeSH experiments we use 8 iterations and for all the other classification problems we study in this work we find 9 iterations give good results. We suspect this difference is due to the much larger size of the training data for MeSH than for any of the other problems we study.

Data Sets In addition to studying classification for MeSH assignment, we also study six smaller textual databases, each of which is associated with one or more classification problems. The databases and their sizes are listed in Table 1. Other than the PubMed database these sources were prepared previously by the authors and documented in[28] and to save space we refer the reader to this source for details. The PubMed database used is a March 2013 copy.

Table 1. Corpora used in this study. Given for each corpus is the number of defined sub-problems, the number of examples, and the number of features.

Database	classes	Examples	Features
REBASE	2	102,997	2,032,075
MDRDataset	3	620,119	738,104
Newsgroups	20	20,000	98,587
IndustrySectors	104	9,555	55,056
WebKB	7	8,280	69,911
Reuters	10	27,578	46,623
PubMed	27	22,411,501	77,040,540

PubMed We study a March 2013 version of the PubMed database. As noted above, there are over 23 million records in PubMed representing as many journal articles mostly on diverse biomedical topics. More detail can be obtained online[29]. Each record in the database is represented by features consisting of the words and two word colocations contained in the title and abstract. A stoplist of function words are not allowed in these features, but no stemming is done. We divide this corpus into two thirds for training (14,941,100) and one third for testing (7,470,501). Most records in PubMed are manually assigned one or more MeSH headings. There are over 27 thousand terms in the MeSH vocabulary, giving rise to as many possible classification problems on this corpus. In a previous work[14], we studied 20 MeSH terms representing a range of frequencies. Each of these MeSH terms is listed

in Table 2 along with its frequency in the studied version of PubMed. Also shown in Table 2 are seven additional higher frequency MeSH terms that were selected for algorithm timing.

Table 2. MeSH terms used in this study of SGD-SVM on the PubMed corpus. Given for each MeSH term is the number of records indexed with that MeSH term and the percent this is of the total.

Set	MeSH Terms	Freq	Percent
M1	rats, wistar	180,179	0.80%
M2	myocardial infarction	131,491	0.59%
M3	blood platelets	63,061	0.28%
M4	serotonin	60,164	0.27%
M5	state medicine	43,787	0.20%
M6	urinary bladder	39,795	0.18%
M7	drosophila melanogaster	31,502	0.14%
M8	tryptophan	25,803	0.12%
M9	laparotomy	14,554	0.06%
M10	crowns	12,862	0.06%
M11	streptococcus mutans	6,623	0.03%
M12	infectious mononucleosis	6,593	0.03%
M13	mentors	6,425	0.03%
M14	blood banks	5,753	0.03%
M15	humeral fractures	5,447	0.02%
M16	tuberculosis, lymphnode	4,471	0.02%
M17	tooth discoloration	2,570	0.01%
M18	pentazocine	2,128	0.01%
M19	hepatitis e	1,845	0.01%
M20	genes, p16	1,782	0.01%
M21	pathological condition, signs and symptoms	3,819,888	17.04%
M22	therapeutics	2,898,880	12.93%
M23	pharmacologic actions	2,404,751	10.73%
M24	enzymes	2,167,158	9.67%
M25	population characteristics	1,242,430	5.54 %
M26	reproductive physiological phenomena	1,019,313	4.55%
M27	terpenes	217,959	0.97%

Results

SGD-SVM Experiments Three folds were used in all cross-validation experiments so that we set $J = 3$ and, where used, $K = 3$. Cross validation was used for all databases except WebKB and 20 Newsgroups which are already partitioned into training and test sets. In these latter two cases we still did the learning with an optimal number of iterations based on holding out a third of the training set to determine it versus doing the learning based on a fixed number of nine iterations over the training data.

Table 3. Results on small corpora for SGD-SVM with cross-validation and SGD-SVM with a fixed 9 iterations. Each row in the table contains the averages over all classification problems defined in that corpus. The last row averages over the corpora.

Set	SVM-SGD Cross Validation			SVM-SGD with Fixed 9 iterations	
	AP	BE	$Iter_{opt}$	AP	BE
Rebase	0.84	0.80	5.3	0.85	0.80
MDR	0.95	0.92	20	0.94	0.91
20News	0.84	0.80	12.2	0.85	0.81
Reuters	0.93	0.89	14	0.93	0.90
WebKB	0.77	0.73	3.8	0.77	0.72
Industry	0.86	0.82	9.7	0.92	0.88
Ave	0.87	0.83	10.8	0.88	0.84

Table 4. Results on the PubMed corpus for SGD-SVM with cross-validation and SGD-SVM with a fixed 8 iterations. The last row averages over the twenty MeSH terms.

MeSH Set	SVM SGD Cross Validation			SVM SGD with 8 iterations	
	AP	BE	$Iter_{opt}$	AP	BE
M1	0.502	0.510	11	0.495	0.508
M2	0.714	0.716	5	0.720	0.715
M3	0.662	0.691	5	0.659	0.690
M4	0.670	0.682	6	0.666	0.682
M5	0.284	0.358	10	0.284	0.356
M6	0.549	0.584	6	0.544	0.585
M7	0.667	0.659	8	0.667	0.659
M8	0.603	0.612	9	0.602	0.613
M9	0.253	0.330	13	0.250	0.332
M10	0.593	0.603	18	0.591	0.597
M11	0.794	0.805	4	0.787	0.799
M12	0.711	0.734	3	0.701	0.724
M13	0.410	0.477	8	0.410	0.477
M14	0.372	0.432	5	0.381	0.433
M15	0.585	0.613	11	0.592	0.620
M16	0.462	0.496	9	0.464	0.499
M17	0.465	0.508	12	0.469	0.516
M18	0.702	0.718	13	0.701	0.714
M19	0.737	0.772	15	0.735	0.784
M20	0.316	0.421	3	0.304	0.414
Ave	0.552	0.586	8.7	0.551	0.586

The results in Table 3 and Table 4 strongly support our contention that a fixed number of iterations gives as good performance for SGD-SVM as using held out data to determine an optimal number of iterations. We also compared SGD-SVM with three published SVM algorithms designed to work well on large training sets. These algorithms are PROBE[30], SVMPerf[31] and LibLinear[32]. Results for all algorithms are in Table 5. No single algorithm is superior to the others on all MeSH terms and SGD-SVM with a fixed 8 iterations is competitive with the others.

Table 5. Results on the PubMed corpus for the four algorithms: SGD-SVM with a fixed 8 iterations, PROBE, SVMperf and LibLinear. For each of the MeSH terms the MAP and BE are reported for a three-fold cross validation and the same folds were used for all the algorithms. The last row averages over the twenty MeSH terms.

	SGD-SVM		PROBE		SVMPerf		LibLinear	
	MAP	BE	MAP	BE	MAP	BE	MAP	BE
M1	0.495	0.508	0.482	0.495	0.472	0.495	0.512	0.507
M2	0.720	0.715	0.723	0.711	0.707	0.697	0.753	0.711
M3	0.659	0.690	0.671	0.689	0.648	0.673	0.693	0.683
M4	0.666	0.682	0.680	0.684	0.651	0.664	0.688	0.671
M5	0.284	0.356	0.308	0.375	0.285	0.350	0.291	0.338
M6	0.544	0.585	0.536	0.570	0.521	0.562	0.553	0.548
M7	0.667	0.659	0.665	0.645	0.633	0.635	0.693	0.642
M8	0.602	0.613	0.591	0.609	0.575	0.599	0.606	0.594
M9	0.250	0.332	0.276	0.340	0.255	0.326	0.308	0.355
M10	0.591	0.597	0.594	0.599	0.583	0.603	0.582	0.569
M11	0.787	0.799	0.777	0.790	0.751	0.773	0.791	0.791
M12	0.701	0.724	0.706	0.729	0.666	0.706	0.725	0.727
M13	0.410	0.477	0.418	0.484	0.389	0.451	0.425	0.486
M14	0.381	0.433	0.381	0.437	0.339	0.424	0.378	0.409
M15	0.592	0.620	0.577	0.606	0.557	0.590	0.568	0.586
M16	0.464	0.499	0.479	0.499	0.428	0.467	0.461	0.495
M17	0.469	0.516	0.443	0.502	0.399	0.463	0.400	0.452
M18	0.701	0.714	0.683	0.690	0.571	0.570	0.701	0.687
M19	0.735	0.784	0.714	0.764	0.717	0.772	0.716	0.737
M20	0.304	0.414	0.311	0.401	0.319	0.401	0.374	0.433
Ave	0.551	0.586	0.551	0.581	0.523	0.561	0.561	0.571

In doing the computations for Table 5 it became evident that SGD-SVM was much faster than the other algorithms. To investigate the issue of computation time further we chose seven additional MeSH terms with high frequencies (thus large positive sets) where the calculation is more difficult. Statistics on these MeSH terms are listed in Table 2. Both performance and elapsed time (wall clock) are shown in Table 6. Two things stand out in Table 6. First, SGD-SVM performs better than PROBE or SVMPerf on these problems and at least as good as LibLinear. Second, PROBE averages between 4 and 5 hours, LibLinear between 8 and 9 hours and SVMPerf may take much longer (we stopped any calculation over 30 hours) for these problems. This in contrast with SGD-SVM which takes 10 minutes or less. We conclude from these experiments that in all circumstances, particularly for large training sets with positive sets of significantly large size, the SGD-SVM algorithm with a fixed 8 iterations is the preferred method.

Table 6. BE performance and elapsed time for classification of high frequency MeSH terms. Here SGD-SVM is with a fixed 8 iterations. All timings were conducted using a standalone multi-core computer with Intel Xeon CPUs with a clock speed of 2.67 GHz and 48 gigabytes of RAM. NA=not available.

	SVM-SGD		PROBE		SVMPerF		LibLinear	
	BE	Time	BE	TIME	BE	TIME	BE	TIME
M21	0.637	9 Min	0.592	284 Min	NA	>30 Hr	0.636	564 Min
M22	0.616	10 Min	0.584	283 Min	NA	>30 Hr	0.619	553 Min
M23	0.616	10 Min	0.574	291 Min	NA	>30 Hr	0.614	547 Min
M24	0.768	9 Min	0.722	256 Min	NA	>30 Hr	0.762	557 Min
M25	0.507	9 Min	0.499	250 Min	NA	>30 Hr	0.521	535 Min
M26	0.709	9 Min	0.690	290 Min	0.689	737 Min	0.702	534 Min
M27	0.714	9 Min	0.671	308 Min	0.698	191 Min	0.689	516 Min
Ave	0.652	9.3 Min	0.619	208 Min	NA	NA	0.649	543 Min

Application to MeSH Assignment Given the speed and accuracy of SGD-SVM, we were led to contemplate calculations which would have been unthinkable in the past. There are over 27 thousand MeSH terms, but it is not difficult to train classifiers for all of them. The question is whether such a set of classifiers will yield superior performance in predicting MeSH assignments for unseen documents. To answer this question we decided to perform an experiment. There are two sets of PubMed documents on which experiments have been done in the past and with which we can compare our results. One set is known as NLM2007, is available at[33] and consists of 200 PubMed records originally selected in 1999 and used as a bench mark by several studies[2, 34, 35]. The second set is L1000, a set of 1,000 randomly selected PubMed records, created and studied in[34] and available at[36].

Based on these considerations we held out the two test sets and used the remainder of the March 2013 MEDLINE as training data. For each MeSH term we randomly divided the training set into two disjoint sets with each half having, as nearly as possible, the same number of documents with the MeSH term assigned and without the MeSH term assigned. We then trained one SGD-SVM classifier on each half. Once a classifier was trained we applied it to produce scores on the half of the documents where it had not been trained. Finally, we applied the Pool Adjacent Violators (PAV) algorithm to these scores to convert them to probabilities. This algorithm produces a probability function of scores that is non-decreasing as a function of score and gives a best fit to the actual data. Best fit here means if the probabilities are looked at as predictions of whether the MeSH term is assigned to a document or not, then the actual assignments observed in the data have maximum probability. No other non-decreasing probability function of score could assign a higher probability to the data. Since we have two trainings, one on each half of the training data, we have two probability functions. We then score the test documents for each training and convert each into a probability based on its PAV function and average these two probabilities to obtain our probability estimate that the MeSH term will be assigned to the test document. Space does not allow more detail regarding PAV, but we refer the reader to[25].

Evaluation. We give four measures of performance: precision, recall, F_1-score, and mean average precision (MAP) which evaluate success in automatic MeSH term assignment to MEDLINE citations. We use these metrics because they have been used for the same task in previous studies. These metrics are standards in the field and are described in[27]. For a given test document precision at rank 25 is the fraction of MeSH terms in the top ranked 25 which were assigned to that document by the human indexers and recall at rank 25 is the fraction of MeSH terms assigned by human indexers to that document which appear in the top 25 ranks of the prediction. The F_1-score at rank 25 is the harmonic mean of the precision and recall at rank 25. The average precision of a ranking for a document is the average of all precisions computed at ranks containing a correctly assigned MeSH term for that document. Finally, for a given set of test documents we compute these four measures for each document and average the results for each measure over all documents in the set and report this average. For average precisions this final average has been given the name of mean average precision or MAP.

We compare our approach to five other methods which were previously published in[34]. The first one is NLM's MTI system[2]. The second method is known as reflective random indexing[35]. The third and fourth are k-nearest neighbor methods. For these methods the neighbors are computed using the algorithm reported in[37]. The value of k determined to be optimal for both frequency and similarity approaches was studied in[34] and found to be 20. The fifth method is based upon a learning-to-rank algorithm which begins with the features used for ranking in the k-nearest neighbor methods and adds a number of other features and learns a weighting for these features that allows an improved ranking of the MeSH terms assigned to the k-nearest neighbors for all training documents at once. For more details see[34]. Table 7 and Table 8 list the results for the databases NLM2007 and L1000, respectively.

Table 7. Precision, recall, F-score over top 25 predictions, and MAP for different methods on the set NLP2007.

	Prec	Recall	F-Score	MAP
MTI	0.318	0.574	0.409	0.450
Reflective random indexing	0.372	0.575	0.451	N/A
Neighborhood frequency	0.369	0.674	0.476	0.598
Neighborhood familiarly	0.376	0.677	0.483	0.604
Learning-to-rank algorithm	0.390	0.712	0.504	0.626
SGD-SVM	0.390	0.712	0.504	0.640
SGD-SVM with PAV	0.405	0.740	0.524	0.681

Table 8. Precision, recall, F1-score over top 25 predictions, and MAP for different methods on the L1000 set.

	Prec	Recall	F-Score	MAP
MTI	0.302	0.583	0.398	0.462
Neighborhood frequency	0.329	0.679	0.443	0.584
Neighborhood similarity	0.333	0.687	0.449	0.591
Learning-to-rank algorithm	0.347	0.714	0.469	0.615
SGD-SVM	0.346	0.712	0.465	0.638
SGD-SVM with PAV	0.368	0.756	0.495	0.676

In these tables all results compared with SGD-SVM are taken from[34]. Two things need to be noted regarding these comparisons. First, for both NLM2007 and L1000 we used the PubMed IDs to find the current forms of the documents and our results are based on the current indexing of these documents. This is the only reasonable thing to do as our training data is all based on the current indexing. This current indexing could conceivably involve some changes that could affect the difficulty in assigning MeSH terms to a document. Since we produced the k-nearest neighbor results used in[34] and quoted in these tables, we recomputed these numbers as a check on any such possible change. We found the four measures to be identical to the numbers given here for the neighborhood frequency method for both NM2007 and L1000. For the neighborhood similarity method we found small variations. For NLM2007 the current F1-score is 0.480 (down 0.003) and the current MAP is 0.605 (up 0.001) and for L1000 the current F1-score is 0.449 (unchanged) and the current MAP is 0.592 (up 0.001). We present these results as evidence that the indexing task for NLM2007 and L1000, as they appear in the 2013 data used for this study, is substantially what it was when studied in[34] and in any case has not gotten easier. The second point we wish to make is that previous studies were limited to rankings of MeSH terms coming from a short list of neighbor documents (generally the top 20) and this limited the number of participants in the ranking to compute the MAP values reported in the tables. In our approach all MeSH are ranked for each document and this allows for a slight improvement in the MAP numbers.

We applied statistical tests to compare the MAP values produced by SGD-SVM with PAV with the Learning-to-rank results. Each method gives a figure for the average precision of MeSH assignment to a document. These pairs over all the documents in a set are then tested. The test methods and the corresponding p-values are given in Table 9.

Table 9. Comparison of SGD-SVM with PAV with Learning-to-rank. We applied the sign test, the Wilcoxon signed rank test, and the paired t test.

	NLM 2007	L1000
Sign test	1.2×10^{-9}	1.4×10^{-39}
Wilcoxon	6.0×10^{-12}	7.9×10^{-59}
Paired t test	2.5×10^{-4}	$<10^{-4}$

Discussion

From a theoretical point of view it is interesting to ask why SGD with early stopping works. In this regard it is useful to point out that what Zhang[17] calls SGD with early stopping is called by Collobert and Bengio[38] the margin perceptron algorithm. The latter authors point out that the wider the margin the better the performance is likely to be and that the margin is preserved by a small value of λ and by a small number of iterations. The small number of iterations is achieved by early stopping. However, as far as we can discern, it has not previously been appreciated that early stopping can be achieved with a fixed number of iterations for a whole class of problems.

The practical consequence of SGD with early stopping at a fixed number of passes over the data is an approximate halving of the training time. The first run with held out data to determine the optimal number of passes over the data proves unnecessary as documented in Table 3 and Table 4. SGD with early stopping is already a very useful approach on problems with large training data and using fixed iterations only adds to the benefit. With this approach there is no need to reduce the size of the training set to make the calculations feasible as done in[39]. We

used unigram and bigram features based on our own experience that this gives good performance[28]. We see that[39] also used unigram and bigram features, but due to our much larger training set sizes we had to deal with a much larger feature set (over 70 million features as compared with their approximately 2 million). In spite of this we can perform a single training run in 10 minutes and perform training over all MeSH terms in a couple of days on our compute farm using approximately 200 cpu's. While at this point we have not been able to directly compare our approach to that of[39], given that they use the same feature types and use a standard SVM approach, we would expect the two methods to perform at a similar level. The results of Table 5 and Table 6 support this conclusion.

We tested our approach on the test sets NLM2007 and L1000 because there are published results for several different methods on these sets[34] and because the particular method of testing on the top 25 predicted MeSH terms is very relevant to the MTI system at NLM. The suggestions to the MeSH indexers by the MTI system[6] are generally on the order of the top 25 terms. The precision, recall and F1 scores given in Table 7 and Table 8 are all computed for the top 25 predicted MeSH terms.

Conclusions

First, SGD with early stopping at a fixed number of iterations is an accurate and fast way to train a SVM classifier for large training sets. Second, when SGD-SVM is combined with the PAV algorithm the results on two previously studied test sets are superior to published results. We are currently collaborating with James Mork and Alan Aronson at the NLM testing whether results from SGD-SVM can be used to improve the MTI system.

Acknowledgement This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

References

1. Aronson AR, Bodenreider O, Chang HF, Humphrey SM, Mork JG, Nelson SJ, et al., The nlm indexing initiative. American Medical Informatics 2000 Annual Symposium; 2000; Los Angeles, CA: American Medical Informatics Association.
2. Aronson AR, Mork JG, Gay CW, Humphrey SM, Rogers WJ. The nlm indexing initiative's medical text indexer. Stud Health Technol Inform. 2004; 107:268-72.
3. Herskovic J, Cohen T, Subramanian D, Iyengar M, Smith J, Bernstam E. Medrank: Using graph-based concept ranking to index biomedical texts. Int J Med Inform. 2011; 80(6):431-41.
4. Jimeno-Yepes A, Plaza L, Mork J, Aronson A, Díaz A. Mesh indexing based on automatically generated summaries. BMC Bioinformatics. 2013; 14(208).
5. Kim W, Aronson AR, Wilbur WJ, Automatic mesh term assignment and quality assessment. Proc AMIA Symp; 2001; Washington, D.C.
6. Mork JG, Jimeno Yepes AJ, Aronson AR. The nlm medical text indexer system for indexing biomedical literature. BioASQ; Valencia, Spain; 2013.
7. Neveol A, Shooshan S, Humphrey S, Rindflesh T, Aronson A, Multiple approaches to fine-grained indexing of the biomedical literature. Pacific Symposium on Biocomputing; 2007.
8. Névéol A, Shooshan S, Mork J, Aronson A, Fine-grained indexing of the biomedical literature: Mesh subheading attachment for a medline indexing tool. AMIA Annu Symp Proc; 2007.
9. Neveol A, Shooshan SE, Humphrey SM, Mork JG, Aronson AR. A recent advance in the automatic indexing of the biomedical literature. Journal of biomedical informatics. 2009; 42(5).
10. Jimeno-Yepes A, Mork J, Demner-Fushman D, Aronson A, Automatic algorithm selection for mesh heading indexing based on meta-learning. Fourth International Symposium on Languages in Biology and Medicine; 2011; Singapore.
11. Jimeno-Yepes A, Mork J, Demner-Fushman D, Aronson A. A one-size-fits-all indexing method does not exist: Automatic selection based on meta-learning. Journal of Computing Science and Engineering. 2012; 6(2):151-60.
12. Jimeno-Yepes A, Mork J, Demner-Fushman D, Aronson A, Comparison and combination of several mesh indexing approaches. AMIA Annual Symposium; 2013; Washington, D.C.: American Medical Informatics Association.
13. Jimeno-Yepes A, Mork J, Wilkowski B, Demner-Fushman D, Aronson A, Medline mesh indexing: Lessons learned from machine learning and future directions. ACM International Health Informatics (IHI) Symposium; 2012.

14. Sohn S, Kim W, Comeau DC, Wilbur WJ. Optimal training sets for bayesian prediction of mesh assignment. J Am Med Inform Assoc. 2008 Jul-Aug; 15(4):546-53.
15. Trieschnigg D, Pezik P, Lee V, de Jong F, Kraaij W, Rebholz-Schuhmann D. Mesh up: Effective mesh text classification for improved document retrieval. Bioinformatics. 2009; 25(11):1412-8.
16. Wikipedia_SGD. Stochastic gradient descent. 2013; Available from: http://en.wikipedia.org/wiki/Stochastic_gradient_descent#cite_ref-6.
17. Zhang T, Solving large scale linear prediction problems using stochastic gradient descent algorithms. Twenty-first International Conference on Machine learning; 2004: Omnipress.
18. Finkel JR, Kleeman A, Manning CD, Efficient, feature-based, conditional random field parsing. ACL-08: HLT; 2008; Columbus, Ohio.
19. Tsuruoka Y, Tsujii Ji, Ananiadou S, Stochastic gradient descent training for l1-regularized log-linear models with cumulative penalty. 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP; 2009; Suntec, Singapore: ACL.
20. Bandos TV, Camps-Valls G, Soria-Olivas E. Letters: Statistical criteria for early-stopping of support vector machines. Neurocomput. 2007; 70(13-15):2588-92.
21. Prechelt L. Automatic early stopping using cross validation: Quantifying the criteria. Neural Networks. 1998; 11:761-7.
22. Raskutti G, Wainwright MJ, Yu B. Early stopping and non-parametric regression: An optimal data-dependent stopping rule. J Mach Learn Res. 2014; 15(1):335-66.
23. Ayer M, Brunk HD, Ewing GM, Reid WT, Silverman E. An empirical distribution function for sampling with incomplete information. Annals of Mathematical Statistics. 1954; 26:641-7.
24. Barlow RE, Bartholomew DJ, Bremner JM, Brunk HD. Statistical inference under order restrictions. The theory and application of isotonic regression. New York: John Wiley & Sons; 1972.
25. Wilbur WJ, Yeganova L, Kim W. The synergy between pav and adaboost. Machine Learning. 2005; 61:71-103.
26. Zadronzny B, Elkan C. Transforming classifier scores into accurate multiclass probability estimates. Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Edmonton, Alberta, Canada: ACM; 2002. p. 694-9.
27. Manning CD, Raghavan P, Schütze H. Introduction to information retrieval. Cambridge, England: Cambridge University Press; 2009.
28. Wilbur WJ, Kim W. The ineffectiveness of within-document term frequency in text classification. Information Retrieval. 2009; 12:509-25.
29. PubMed H. Pubmed help [internet]. Bethesda, MD: National Center for Biotechnology Information (US); 2013. Available from: http://www.ncbi.nlm.nih.gov/books/NBK3830/.
30. Smith L, Kim W, Wilbur WJ. Probe: Periodic random orbiter algorithm for machine learning2012.
31. Joachims T. Training linear svms in linear time. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining; Philadelphia, PA, USA: ACM; 2006.
32. Fan R-E, Chang K-W, Hsieh C-J, Wang X-R, Lin C-J. Liblinear: A library for large linear classification The Journal of Machine Learning Research. 2008; 9.
33. Aronson AR. 200 medline citations test collection Bethesda, MD: National Library of Medicine; 2013 [cited 2013]; Available from: http://ii.nlm.nih.gov/DataSets/index.shtml#200MEDLINE.
34. Huang M, Neveol A, Lu Z. Recommending mesh terms for annotating biomedical articles. J Am Med Inform Assoc. 2011 Sep-Oct; 18(5):660-7.
35. Vasuki V, Cohen T. Reflective random indexing for semi-automatic indexing of the biomedical literature. J Biomed Inform. 2010; 43(5):694-700.
36. Lu Z. Recommending mesh data sets. Bethesda, MD: National Library of Medicine; 2013; Available from: http://www.ncbi.nlm.nih.gov/CBBresearch/Lu/indexing/.
37. Lin J, Wilbur WJ. Pubmed related articles: A probabilistic topic-based model for content similarity. BMC Bioinformatics. 2007; 8:423.
38. Collobert R, Bengio S. Links between perceptrons, mlps and svms. 21st International Conference on Machine Learning; Banff, CA; 2004.
39. Tsoumakas G, Laliotis M, Markantonatos N, Vlahavas I. Large-scale semantic indexing of biomedical publications at bioasq. BioASQ 2014 Valencia, Spain; 2013.

Using String Metrics to Identify Patient Journeys through Care Pathways

Richard Williams, BA[1,2], Iain E. Buchan, MD, FACMI[1,2], Mattia Prosperi, M.Eng., Ph.D[1], John Ainsworth, BSc, MSc[1,2]
[1]Centre for Health Informatics, [2]Greater Manchester Primary Care Patient Safety Translational Research Centre, University of Manchester, Manchester, UK.

Abstract

Given a computerized representation of a care pathway and an electronic record of a patient's clinical journey, with potential omissions, insertions, discontinuities and reordering, we show that we can accurately match the journey to a particular route through the pathway by converting the problem into a string matching one. We discover that normalized string metrics lead to more unique pathway matches than non-normalized string metrics and should therefore be given preference when using these techniques.

Introduction

When faced with a patient's electronic health record (EHR) and a prescribed care pathway it is useful to know if that patient's care has deviated from the expected route through the pathway[1]. The degree of deviation from a pathway calculated with a distance metric, when combined with outcome data, could lead to the discovery of instances where the standard of care has been suboptimal leading to adverse outcomes, and also to instances of localized practice that lead to better outcomes.

However, before determining distance from a given route, we need accurately to determine which route through the pathway was traversed by the patient. This is a problem because routinely collected patient information is often poorly recorded with missing data, incorrect coding practice and data recorded out of sequence.

String metrics provide the distance between two strings and are usually based on algorithms for matching strings to patterns, with various degrees of approximation. They typically involve performing operations such as insertion, deletion and substitution. The string metric can be normalized[2,3] or non-normalized[4–6].

We attempt to discover the routes patients took through a care pathway by using string matching methods in a novel way with electronic health records from Salford, UK.

Related Work

Representing a care pathway in a format that can be readily interpreted by a computer is essential for analysis and also enables health information systems to provide decision support to health care professionals[7]. Computer-interpretable guidelines (CIGs) are computer representations of the clinical knowledge in a clinical guideline and are usually networks of tasks that occur over time[8]. A recent review of CIGs shows there is ongoing work on CIG modelling languages, their integration with EHRs, validation and verification of CIGs, compliance monitoring and sharing[9]. Most CIG modelling is based on Task-Network Models[8,9] of which our graph-based approach is a general case.

There is also a large body of work on process mining[10,11], frequent pattern mining, and the use of hidden Markov models for trajectory clustering[12] for healthcare data, which has been reviewed by Lakshmanan et al.[13] However, each of these techniques begin with the healthcare data and attempts to interpolate the pathways taken, whereas our approach differs by starting with a well-defined care pathway and attempts to discover the route taken.

Background

Care Pathways

Care pathways are structured guidelines for the assessment, diagnosis, and treatment of patients with a given condition[1,14–16]. They provide the ideal care that a patient should receive and are often represented as a flow chart[1,14]. In the UK, "NICE Pathways" (National Institute for Health and Care Excellence) offers pathways for over 150 conditions[17].

More formally, a care pathway flow chart can be represented as a directed graph, $G = (V, E)$, with V a set of nodes that represent clinical events such as diagnoses, measurements, procedures and treatments, and E a set of directed

edges that correspond to the permitted transitions between nodes. A transition can occur in a determined amount of time. Figure 1 shows an example of a care pathway represented as a directed graph, defined a priori by experts.

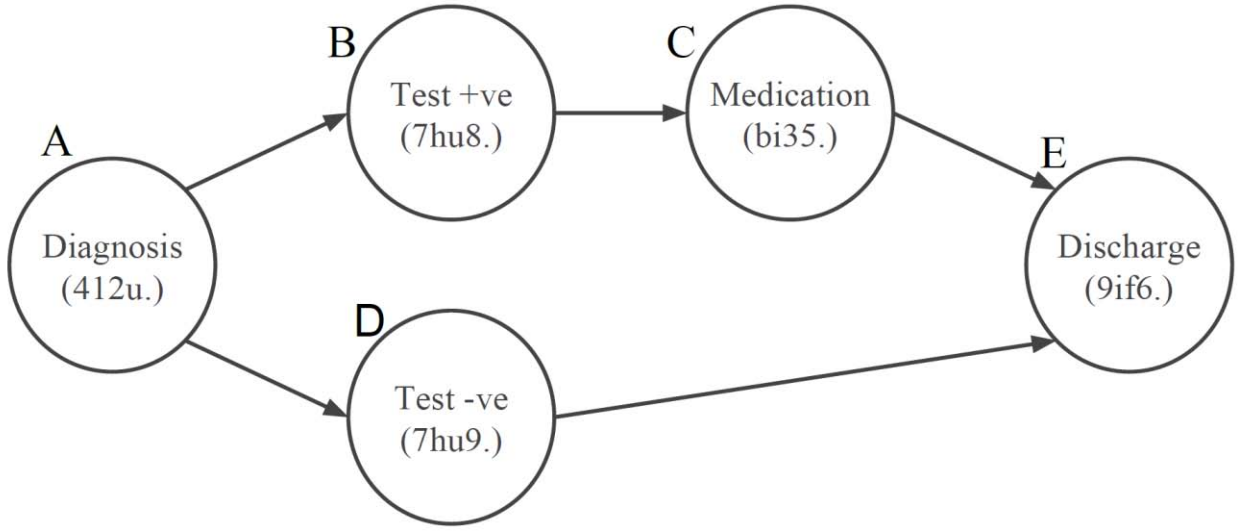


Figure 1: A graphical model of a simplified, coded care pathway. Clinical codes in parentheses.

SINAP

The Stroke Improvement National Audit Programme (SINAP)[18] is a data collection process for the purposes of clinical audit. It collects data about the care provided to stroke patients and includes several index events and the times they occurred. Here we examine data from Salford Royal Foundation Trust (SRFT) on 1078 patients with suspected strokes between 2010 and 2011. Figure 2 shows the approximate pathways that can be followed when a patient is admitted to hospital with a suspected stroke, covering the events recorded in the SINAP dataset. This is a simple pathway with only two decision points following when the patient is first seen and also after the patient has undergone brain imaging. The alphanumeric characters associated with each node in the pathway will be used later.

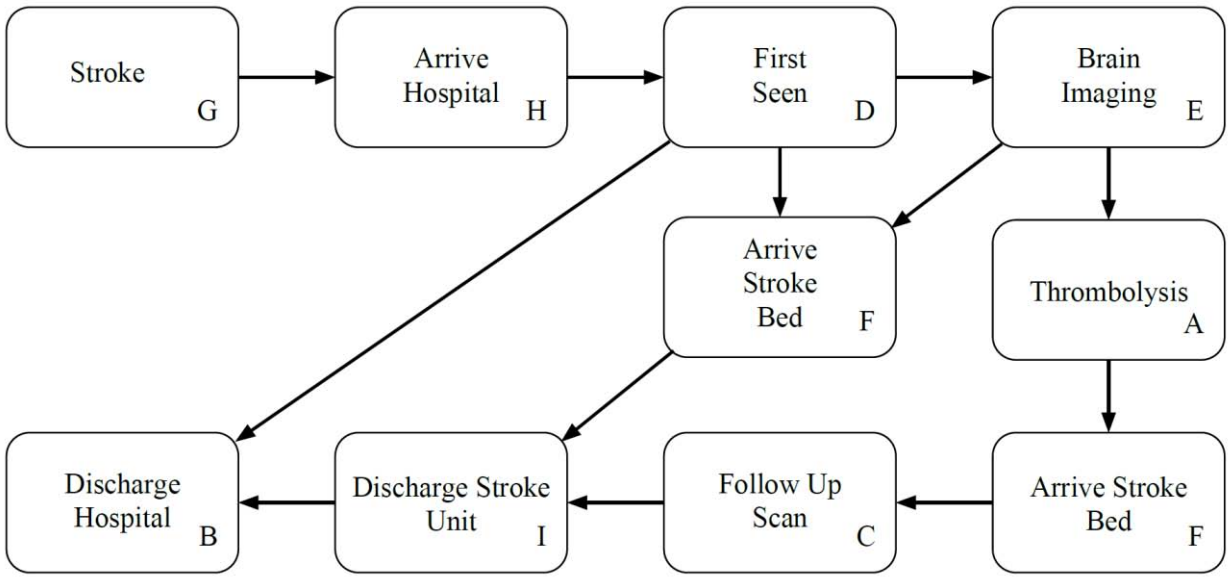


Figure 2: Stroke Improvement National Audit Programme (SINAP) pathway nodes as characters.

Electronic Health Record

A patient's EHR is typically a list of coded events and states describing their care. In the UK a variety of coding schemes are used, such as Read Codes v2[19], CTV3[19], ICD-10[20] and SNOMED[21]. The processes described in this paper can be used with any coding system: here we use the SINAP dataset that employs custom codes.

Method

Process

We first assign an alphanumeric character to each node in the graph. By using the Unicode[22] character set we can manage care pathways with up to 65,536 nodes. We then extract every possible route through the pathway as a string made up of the characters assigned to each node. For a graph G with n possible routes we construct the set $R = \{R_1, R_2, \cdots, R_n\}$, where each R_i is a string representing one of the n possible routes. For acyclic graphs such as the stroke pathway for the SINAP dataset this is straightforward via recursion. For a directed graph with cycles it is possible to repeat a cycle indefinitely so the number of possible routes is infinite. To avoid this we only allow each cycle to be repeated a finite number of times.

Due to the nature of our data, the events recorded are all covered by the pathway. In general, however, when using records from primary or secondary care, they may not be consistent with a care pathway event/transition graph. For a single patient we therefore extract all timed events from their record that occur on the pathway of interest, convert the events to characters, and concatenate the characters into strings according to their date-time order. The strings then represent the patient's journey through the care pathway.

If our dataset contains patients with multiple interactions with the pathway, we must then distinguish between distinct interactions with the care pathway by specifying a cut-off time. If ever the gap between adjacent patient events is greater than the cut-off, then we assume that the patient has left the pathway and any subsequent events form part of the patient's next visit to the pathway. This works well when the timescale of a pathway is shorter than the distances between them.

We then use the following string metrics to determine the distance between a patient pathway and each possible route through a care pathway.

Longest Common Subsequence

Formally, given two sequences $A = a_1 a_2 \cdots a_m$ and $B = b_1 b_2 \cdots b_n$ $(m \leq n)$ we say that A is a subsequence of B if there are indices $0 < j_1 < j_2 < \cdots < j_m \leq n$ such that $a_i = b_{j_i}$ is true for $i = 1,2, \ldots, m$.

Given two sequences $\boldsymbol{X}$ and $\boldsymbol{Y}$, $\boldsymbol{Z}$ is a common subsequence if it is a subsequence of both $\boldsymbol{X}$ and $\boldsymbol{Y}$. $\boldsymbol{Z}$ is the longest common subsequence (LCS) if $|\boldsymbol{Z}| >= |\boldsymbol{Z'}|$ for all common subsequences $\boldsymbol{Z'}$, where $|\boldsymbol{X}|$ is the length of $\boldsymbol{X}$. The LCS is not necessarily unique.

We are interested in which route through the pathway a patient took so we need to decide on a distance metric to convert the LCS into something more meaningful. An initial algorithm for a single patient is as follows:

1. Create a list of all the possible routes $\boldsymbol{R_1}, \ldots, \boldsymbol{R_n}$ through the care pathway
2. Filter the patient's events to just include pathway events and apply the time cut-off to give an event sequence $\boldsymbol{E} = \boldsymbol{E_1} \ldots \boldsymbol{E_m}$
3. For each route $\boldsymbol{R_i}$ calculate $\boldsymbol{L_i} = \boldsymbol{LCS}(\boldsymbol{R_i}, \boldsymbol{E})$
4. If $\boldsymbol{L_i} > \boldsymbol{0}$ calculate the distance $\boldsymbol{d_i} = \boldsymbol{max}(|\boldsymbol{R_i}|, |\boldsymbol{E}|) - |\boldsymbol{L_i}|$
5. Return the set of routes with the smallest distance

However, this only considers the discrepancy between the LCS and the pathway route; it doesn't take into account the length of the LCS. We can normalize the distance by either dividing by the LCS, or by dividing by the combined length of the two strings and step 4 above becomes either:

4. If $\boldsymbol{L_i} > \boldsymbol{0}$ calculate the distance $\boldsymbol{d_i} = \frac{\boldsymbol{max}(|\boldsymbol{R_i}|, |\boldsymbol{E}|) - |\boldsymbol{L_i}|}{|\boldsymbol{L_i}|}$

or

4. If $\boldsymbol{L_i} > \boldsymbol{0}$ calculate the distance $\boldsymbol{d_i} = \frac{\boldsymbol{max}(|\boldsymbol{R_i}|, |\boldsymbol{E}|) - |\boldsymbol{L_i}|}{|\boldsymbol{R_i}| + |\boldsymbol{E}|}$

We call these two methods LCS1 and LCS2 respectively.

Simple Edit Distance (Levenshtein Distance)

An alternative to the LCS is to consider the edit distance or Levenshtein distance[4]. The edit distance between two strings $\boldsymbol{X}$ and $\boldsymbol{Y}$ is the minimum number of operations required to convert $\boldsymbol{X}$ into $\boldsymbol{Y}$ where an operation is either: insert a character, delete a character or replace a character. When switching is allowed ($\boldsymbol{ab \rightarrow ba}$) the algorithm is the Damerau-Levenshtein[5,6]. The costs of inserting, deleting and replacing are given as $\boldsymbol{W_I}$, $\boldsymbol{W_D}$, and $\boldsymbol{W_R}$ respectively. It holds that $\boldsymbol{W_R \leq W_D + W_I}$, as we can always delete and then insert instead of substituting. By default the cost of each operation is 1.

The algorithm for our problem would be:

1. Create a list of all the possible routes $\boldsymbol{R_1, \ldots, R_n}$ through the care pathway
2. Filter the patient's events to just include pathway events and apply the time cut-off to give an event sequence $\boldsymbol{E = E_1 \ldots E_m}$
3. For each route $\boldsymbol{R_i}$ calculate the distance $\boldsymbol{d_i = LEV(R_i, E)}$
4. Return the set of routes with the smallest distance

Similarly we can do this for the Damerau-Levenshtein distance which we will notate as $\boldsymbol{d_i = DAM(R_i, E)}$.

Levenshtein Variants

Several versions of the Levenshtein Distance normalized to the length of the strings have been suggested. We notate the following as $\boldsymbol{NLEV}$[2].

$$\boldsymbol{NLEV(X,Y) = \frac{LEV(X,Y)}{|X| + |Y|}}$$

Also a normalized Levenshtein distance that satisfies the triangle equality and is therefore a true distance metric:

$$\boldsymbol{NLD(X,Y) = d_{N-GLD}(X,Y) = \frac{2 \cdot LEV(X,Y)}{\alpha \cdot (|X| + |Y|) + LEV(X,Y)}}$$

where α is whichever cost is greater out of insertion and deletion[3]. However, when $\boldsymbol{\alpha = 1}$, as is the case when all the weights are set to 1 by default, although the distances produced by NLD and NLEV will differ, the ordering of the matches will always be the same.

Finally, we consider a normalized version of the Damerau Levenshtein distance.

$$\boldsymbol{NDAM(X,Y) = \frac{DAM(X,Y)}{|X| + |Y|}}$$

We compare and contrast the different distance measures: LCS1, LCS2, LEV, DAM, NLEV, NLD and NDAM.

Data cleaning

Right censoring of the data is unlikely as once in hospital all end points are recorded. Most times in the data seem to be rounded to the nearest 10 or 15 minutes. This may potentially result in events appearing simultaneously or even out of order. There is also a risk of recollection or estimation bias as the data is often captured after the event.

When events occur at the same time there are several options available. The patient can be ignored, but this would result in a lot of data being excluded from the analysis. An alternative would be to perform the analysis on the data ordered randomly and let the string matching methods correct any discrepancies. However as we are interested in discovering the actual path the patient took, we can assume where possible the events occurred in the correct order.

For two events A and B on a pathway there is either: a one-way path from A to B, a one-way path from B to A, a path from A to B and B to A, or it is impossible to get from one to the other. For a group of events occurring at the same time if it is possible to order them in a unique way then we choose that as the order of the events. If it is not possible, because of a cycle or an unreachable node, then we discard that patient. For datasets where this is commonplace it may be better to include the patients discarded here and randomise the order of the cotemporaneous events. Alternatively we could just discard the events rather than the patient.

Similarly, events of unknown time, or those with just a date and not a time, can be inserted at the correct point of a patient record, if possible, or discarded if contradictions arise.

Data Management and Analysis Environment

The SINAP dataset was transferred to us via an encrypted external hard drive in CSV format. This was then uploaded to a Microsoft SQL Server 2008 database for analysis. Sequence matching was performed with C#.NET and all statistical analysis was done using R[23]. The sm library[24] was used for plotting density curves and the pROC[25] package was used for comparing Receiver Operating Characteristic (ROC) curves.

Results

Data Characteristics

The SINAP dataset contains 1078 patients of which 549 are female and 529 are male.

Table 1 shows the number of records that were cleaned using the above data cleaning process. Only 1 patient's route could not be uniquely re-ordered.

Table 1. Data cleaning results

Total patients	1078
Midnight events – able to insert	424
Simultaneous events – able to order	3
Midnight and simultaneous events – able to order	648
No midnight or simultaneous events – no need to order	2
Midnight events – unable to insert	1

There are 46 distinct pathways taken by the 1077 patients following time reordering. Table 2 shows the frequency of the top 10 patient pathways. The pathways that match the ICP are in bold. The route of GHDB should be a valid route however there are no patients in our cohort who followed this – suggesting this is not a valid route and the care pathway could be altered.

Table 2. Top 10 pathways – character sequences from figure 2.

Patient Record	Count	Comments
GHDEFIB	**275 (26%)**	**Valid route**
GHDFIB	**275 (26%)**	**Valid route**
GHDFEIB	122 (11%)	Valid route with E/F switched – lots of people so maybe a valid route.
GDHEFIB	63 (6%)	Valid route with D/H switched – can't be seen before you arrive.
GDHFIB	60 (6%)	Valid route with D/H switched – as above.
GHDEAFCIB	**56 (5%)**	**Valid route**
GHEDFIB	39 (4%)	Valid route with E/D switched – can't be imaged before first seen.
GHDEFACIB	37 (3%)	Valid with A/F switched.
GHFDIB	24 (2%)	D/F switched – can't arrive in specialist bed before being seen.
GDHFEIB	24 (2%)	D/H and E/F switched

It appears that there are some valid routes that aren't in our pathway. For those who don't get thrombolysed there are many people who arrive in a specialist stroke bed prior to their brain scan. Also there are many people who get "First Seen" before they arrive at the hospital. This seems nonsensical but could be valid if "First Seen" applied to GPs or ambulance staff. Finally there are patients who receive thrombolysis after getting to a specialist stroke bed which could also be a valid route. All other switches appear to be mistakes – for example having a brain scan prior to being first seen.

In order to determine how well each method works we must determine for each patient the most probable route taken. As our dataset is small we can do this manually by defining rules based on the data. We first assume that events that don't happen are rarely inserted and then classify the patients according to the following rules:

1. If a patient has thrombolysis or a follow up scan then assumes route GHDEAFCIB
2. Of those remaining, for any with a brain scan we assume route GHDEFIB

3. Of those remaining, for any with a stroke unit arrival or discharge we assume route GHDFIB
4. Of those remaining we assume GHDB

In addition to returning the correct result it is also of use if the distance measure returns a unique result. There will be situations where this isn't possible but in general string matching methods that return more unique results are preferable.

For each method, Table 3 gives the number of unique matches and the number of correct matches where a correct match is one that is both unique and matches with the routes we assume the patients actually followed.

Table 3. Number of unique and correct matches

Method	Unique Matches	Correct Matches	Correct
LEV	818 (75.95%)	645 (78.85%)	59.89%
DAM	853 (79.20%)	849 (99.53%)	78.83%
LCS1	882 (81.89%)	878 (99.55%)	81.52%
LCS2	1077 (100.00%)	1070 (99.35%)	99.35%
NLEV	1076 (99.91%)	841 (78.16%)	78.09%
NLD	1076 (99.91%)	841 (78.16%)	78.09%
NDAM	1076 (99.91%)	1068 (99.26%)	99.16%

The NLEV and NLD methods produce the same results as predicted. The ratio of correct matches to unique matches shows that the Damerau-Levenshtein and the longest common subsequence methods work excellently with >99% correct, whereas the Levenshtein variants only achieve 78-79%. It can also be seen that normalized methods are better at producing unique matches with LCS2 matching all pathways uniquely, while NLEV, NLD and NDAM only fail to give a unique answer for a single patient - actually a different patient for each method. Examining the difference between NLEV and NDAM shows that NDAM is correctly identifying pathways where events have been recorded out of sequence. As an example the patient record of GHDFEIB is correctly matched to GHDEFIB by NDAM, while NLEV matches it to GHDFIB.

When the values for unique correct matches are combined the normalized Damerau-Levenshtein and the second Longest Common Subsequence methods are best, correctly matching >99% of the patient pathways.

For these two methods we can split the pathways into two groups: correct and incorrect matches, where a correct match is when the algorithm uniquely identifies the route the patient traversed through the pathway. We then compare the groups under the null hypothesis that the mean 'string' distance between them is equal. The density plots in Figure 3 demonstrate the data we want to contrast are not drawn from normal or symmetrical distributions, indeed the distributions of string distances are quite different for matches compared with non-matches. Thus we make the contrast with a non-parametric (Matt-Whitney) method[26], demonstrating statistically highly significant differences for both NDAM ($P < 0.0001$) and LCS2 ($P < 0.0001$) metrics.

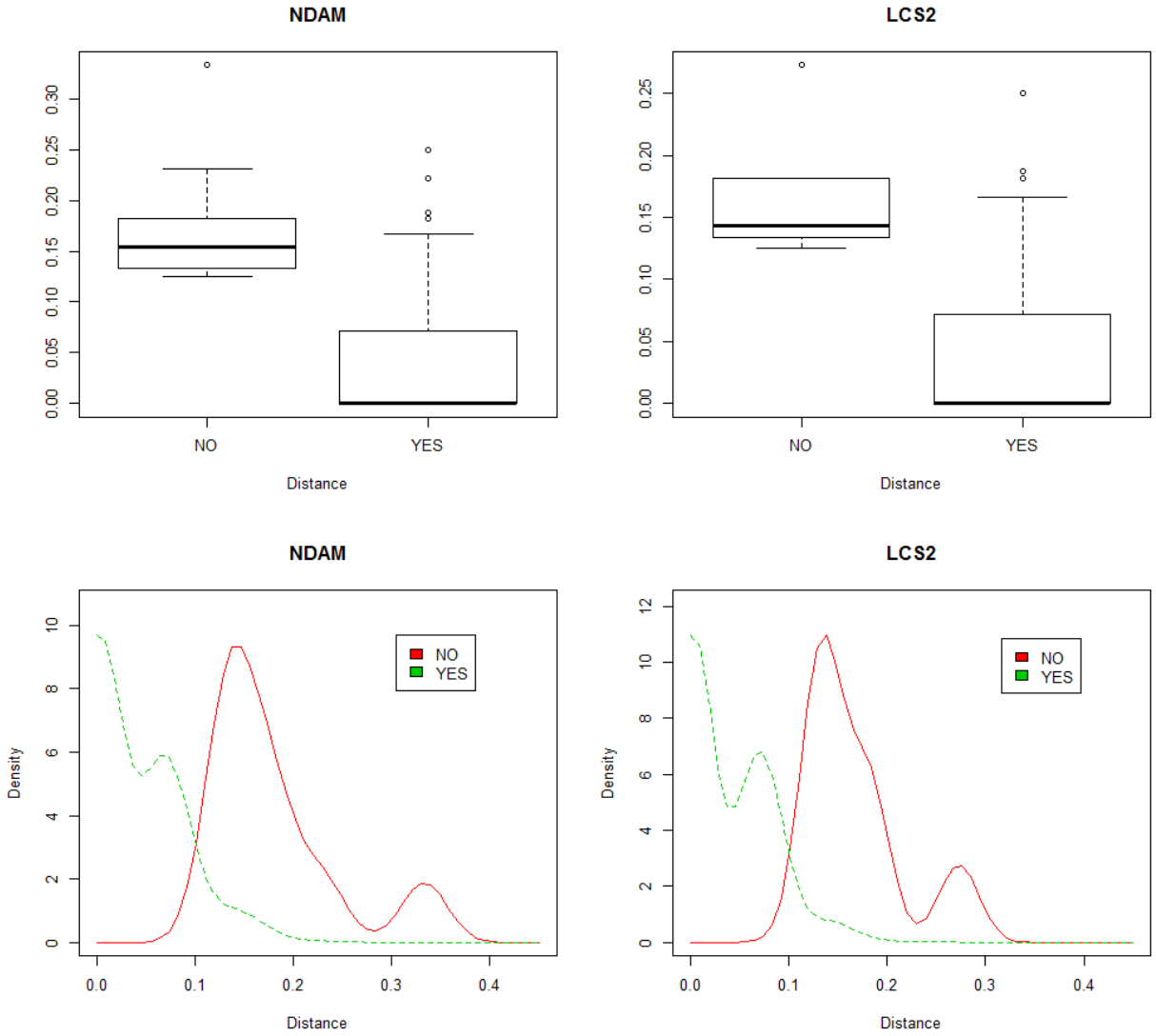


Figure 3: Box (top) and density (bottom) plots of string distances for matches (YES) and non-matches (NO) for NDAM and LCS2 metrics.

Finally, we compare NDAM, LCS2 and NLEV string distance metrics with regard to their classification accuracy for our care pathway journeys. Figure 4 shows the ROC curves for each metric with our test dataset, and the 95% confidence intervals for the areas under the curves: the more detailed comparison of the two most accurate metrics (NDAM and LCS2) is the Mann-Whitney result above.

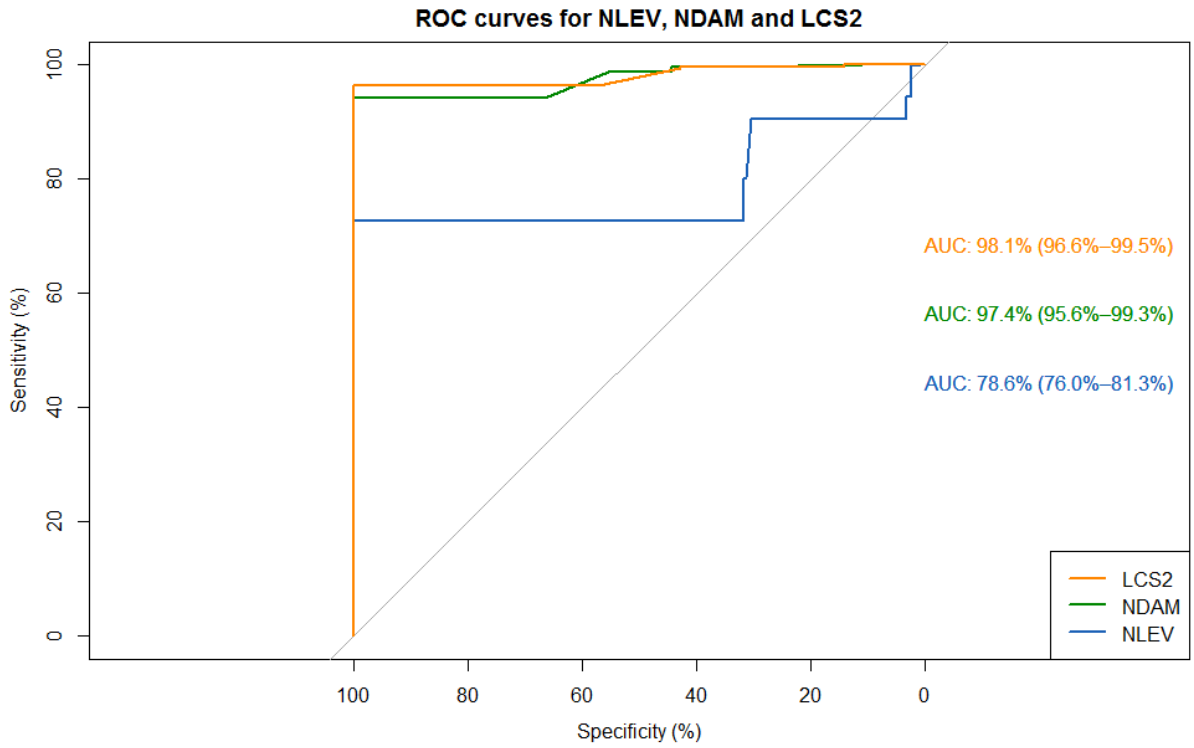


Figure 4: Receiver Operating Characteristic (ROC) curves for NDAM, NLEV and LCS2 string distance metrics with 95% confidence intervals for the areas under the curves.

Discussion

Distance Weighting

The operations in the Damerau-Levenshtein string metric can be weighted. Given the nature of our dataset it is more likely that records were omitted or out of order, than miscoded. If we are sure of this we can change the weighting of the operations accordingly – an option that is possible with the NDAM and not the LCS2 method. By doubling the weight associated with deleting a character, therefore making it less likely that matches will feature deletions, of the 1077 patients we yield 1077 unique matches of which 1074 are correct. Weighted NDAM then becomes the most accurate way of predicting a patient's route.

Generalization

The string matching process described here operates on a graph based representation of a care pathway. Therefore the methodology is theoretically applicable, although untested, to any process or workflow that can be represented as a graph, in healthcare and beyond.

Future work

There are several factors unstudied in this paper that will affect the overall success of the method. The size and shape of the graph is a factor, as is the quality of the data. Further work is needed to determine which graph shapes work well with this method. Finally, the next stage of our work is to determine how the distance a patient is from their care pathway predicts their outcomes.

Conclusion

String matching would seem to be a highly successful way to determine which route a patient followed in a care pathway. Normalized distance functions should be used to ensure high numbers of unique matches. For clinical data where the chance of events occurring, or being recorded, in the wrong order is high, the Damerau-Levenshtein or Longest Common Subsequence methods should be used in preference to the Levenshtein distance.

Acknowledgements

Funded by the National Institute for Health Research Greater Manchester Primary Care Patient Safety Translational Research Centre (NIHR GM PSTRC). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

References

1. Ainsworth J, Buchan I. COCPIT: A Tool for Integrated Care Pathway Variance Analysis. *Stud Health Technol Inform*. 2012;180:995-9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22874343.

2. Marzal A, Vidal E. Computation of normalized edit distance and applications. *IEEE Trans Pattern Anal Mach Intell*. 1993;15. doi:10.1109/34.232078.

3. Yujian L, Bo L. A normalized Levenshtein distance metric. *IEEE Trans Pattern Anal Mach Intell*. 2007;29:1091-1095. doi:10.1109/TPAMI.2007.1078.

4. Levenshtein VI. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Sov Phys Dokl*. 1966;10:707-710. Available at: http://adsabs.harvard.edu/abs/1966SPhD...10..707L.

5. Damerau FJ. A technique for computer detection and correction of spelling errors. *Commun ACM*. 1964;7:171-176. doi:10.1145/363958.363994.

6. Oommen BJ, Loke RKS. Pattern recognition of strings with substitutions, insertions, deletions and generalized transpositions. *Pattern Recognit*. 1997;30(5):789-800. doi:10.1016/S0031-3203(96)00101-X.

7. Gooch P, Roudsari A. Computerization of workflows, guidelines, and care pathways: a review of implementation challenges for process-oriented health information systems. *J Am Med Inform Assoc*. 2011;18(6):738-48. doi:10.1136/amiajnl-2010-000033.

8. Peleg M, Tu S, Bury J, et al. Comparing Computer-interpretable Guideline Models: A Case-study Approach. *J Am Med Informatics Assoc* . 2003;10 (1):52-68. doi:10.1197/jamia.M1135 .

9. Peleg M. Computer-interpretable clinical guidelines: a methodological review. *J Biomed Inform*. 2013;46(4):744-63. doi:10.1016/j.jbi.2013.06.009.

10. Huang Z, Dong W, Ji L, Gan C, Lu X, Duan H. Discovery of clinical pathway patterns from event logs using probabilistic topic models. *J Biomed Inform*. 2014;47:39-57. doi:10.1016/j.jbi.2013.09.003.

11. Kaymak U, Mans R, Steeg T Van De, Dierks M. On process mining in health care. *2012 IEEE Int Conf Syst Man, Cybern*. 2012:1859-1864. doi:10.1109/ICSMC.2012.6378009.

12. Poelmans J, Dedene G. Combining business process and data discovery techniques for analyzing and improving integrated care pathways. *Adv Data ...*. 2010. Available at: http://link.springer.com/chapter/10.1007/978-3-642-14400-4_39. Accessed July 8, 2014.

13. Lakshmanan G, Rozsnyai S, Wang F. Investigating clinical care pathways correlated with outcomes. *Bus Process Manag*. 2013:323-338. Available at: http://link.springer.com/chapter/10.1007/978-3-642-40176-3_27. Accessed July 8, 2014.

14. Schrijvers G, van Hoorn A, Huiskes N. The care pathway: concepts and theories: an introduction. *Int J Integr Care*. 2012;12(Spec Ed Integrated Care Pathways):e192. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3602959&tool=pmcentrez&rendertype=abstract. Accessed March 4, 2014.

15. Campbell H, Hotchkiss R, Bradshaw N, Porteous M. Integrated care pathways. *BMJ*. 1998;316(7125):133-7. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2665398&tool=pmcentrez&rendertype=abstract. Accessed March 4, 2014.

16. Riley K. Care pathways. Paving the way. *Health Serv J*. 1998;108(5597):30-1. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10177611. Accessed March 4, 2014.

17. NICE Pathways. Available at: http://pathways.nice.org.uk/. Accessed March 4, 2014.

18. SINAP (Stroke Improvement National Audit Programme) | Royal College of Physicians. Available at: http://www.rcplondon.ac.uk/projects/stroke-improvement-national-audit-programme-sinap. Accessed March 4, 2014.

19. NHS Connecting for Health. NHS Connecting for Health - Read Codes. *{NHS} Connect Heal*. 2013. Available at: http://www.connectingforhealth.nhs.uk/systemsandservices/data/uktc/readcodes.

20. ICD-10 Classification — NHS Connecting for Health. Available at: http://www.connectingforhealth.nhs.uk/systemsandservices/data/clinicalcoding/codingstandards/icd10. Accessed March 8, 2014.

21. Release I, International T, Terminology H, Development S. SNOMED Clinical Terms Technical Reference Guide. *Development*. 2008:164. Available at: http://htg.his.uvic.ca/index.php?ContentFileId=57.

22. Needleman M. The Unicode Standard. *Ser Rev*. 2000;26:51-54. doi:10.1016/S0098-7913(00)00059-9.

23. R Core Team. R: A Language and Environment for Statistical Computing. 2013. Available at: http://www.r-project.org/.

24. Bowman AW, Azzalini A. *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-plus Illustrations*.; 1997:982. doi:10.2307/2670015.

25. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12:77. doi:10.1186/1471-2105-12-77.

26. With S. Wilcoxon – Mann – Whitney. *Stat Surv*. 1945;4:1-3. doi:10.1214/09-SS051.

Desiderata for an authoritative Representation of MeSH in RDF

Rainer Winnenburg, PhD, Olivier Bodenreider, MD, PhD
National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA
{rainer.winnenburg|olivier.bodenreider}@nih.gov

The Semantic Web provides a framework for the integration of resources on the web, which facilitates information integration and interoperability. RDF is the main representation format for Linked Open Data (LOD). However, datasets are not always made available in RDF by their producers and the Semantic Web community has had to convert some of these datasets to RDF in order for these datasets to participate in the LOD cloud. As a result, the LOD cloud sometimes contains outdated, partial and even inaccurate RDF datasets. We review the LOD landscape for one of these resources, MeSH, and analyze the characteristics of six existing representations in order to identify desirable features for an authoritative version, for which we create a prototype. We illustrate the suitability of this prototype on three common use cases. NLM intends to release an authoritative representation of MeSH in RDF (beta version) in the Fall of 2014.

1 Introduction

In their seminal paper in 2001 [1], Berners-Lee et al. offer a vision of the Semantic Web featuring use cases in healthcare and the life sciences, such as accessing treatment information, finding healthcare providers and scheduling appointments. Later, Ruttenberg and members of the Health Care and Life Sciences Interest Group (HCLSIG) of the World Wide Web Consortium have highlighted the potential of the Semantic Web for supporting translational research [2].

In the era of Linked Open Data, the biomedical domain represents a significant portion of the Linked Open Data cloud [3], a growing collection of interoperable resources supported by Semantic Web technologies. As shown in Figure 1, the biomedical portion of the Linked Open Data cloud (depicted in pink) included over 40 datasets in 2011 and is still growing as datasets become available in formats suitable for Linked Data (e.g., RDF – the Resource Description Framework).

Some data providers have made their resources available in RDF (e.g., UniProt [4]). In many cases, however, the Semantic Web community has stepped up and transformed existing resources to RDF so they can participate in the Linked Open Data (LOD) cloud. According to the statistics published on the LOD cloud website [5], as of August 2011, out of the 295 datasets in the LOD cloud only 113 (39 %) were published by the data producers themselves, while 180 (61 %) were published by third-parties. For example, LinkedCT is a Linked Data version of the National Library of Medicine's registry of clinical trials, ClinicalTrials.gov, created and maintained by researchers at the University of Toronto [6], and used in several projects, including clinical registries [7].

While Linked Open Data is arguably the most visible part of the Semantic Web, Semantic Web technologies have permeated many industries, including libraries. For example, the Library of Congress has recently initiated the Bibliographic Framework Initiative (BIBFRAME) [8], an attempt to replace the legacy MARC 21 format [9] with Semantic Web technologies for the representation and exchange of bibliographic data [10]. This new framework could leverage RDF representations of legacy authority files, such as the Library of Congress Subject Headings and the National Library of Medicine's Medical Subject Headings (MeSH) [11], for the annotation of bibliographic records. However, MeSH is made available by its developer in XML, MARC, and ASCII flat files, as well as through the Unified Medical Language System (UMLS) Metathesaurus [12], but not in RDF.

This initiative prompted us to revisit our earlier attempt to produce an RDF version of MeSH, in the objective of establishing desiderata for an authoritative representation of MeSH in RDF. More specifically, we explore the Linked Open Data cloud for RDF versions of MeSH contributed by the community, including our own, and we analyze their characteristics in order to identify desirable features for an authoritative version. We propose a prototype RDF representation of MeSH that meets these criteria, and we illustrate its usefulness through three common use cases. NLM intends to release an authoritative representation of MeSH in RDF (beta version) in the Fall of 2014.

2 Background

2.1 Semantic Web technologies

The Semantic Web is an extension of the current Web [1, 13]. Underlying the Semantic Web are a set of technologies, including Uniform Resource Identifiers (URIs) – identifiers for resources on the Web [14], the Resource Description Framework (RDF) – a format for representing (and making statements about) Web resources [15], and the SPARQL query language for RDF repositories [16]. Ontologies provide the vocabulary and shared semantics required for annotating resources and to support inference. RDF and the Web Ontology Language (OWL) are the W3C standards for encoding data/knowledge [17]. The Simple Knowledge Organization System (SKOS) is recommended for the representation of thesauri and similar artifacts [18].

RDF describes information in the form of subject-predicate-object triples. This enables information to be represented in the form of a graph. The graph can then be queried using SPARQL. RDF has multiple serialization formats, including RDF/XML, N-Triples, Turtle and, most recently, JSON-LD. The two main distribution mechanisms for RDF data are making the RDF datasets available for download and providing a "SPARQL endpoint", i.e., a live service to which queries can be made.

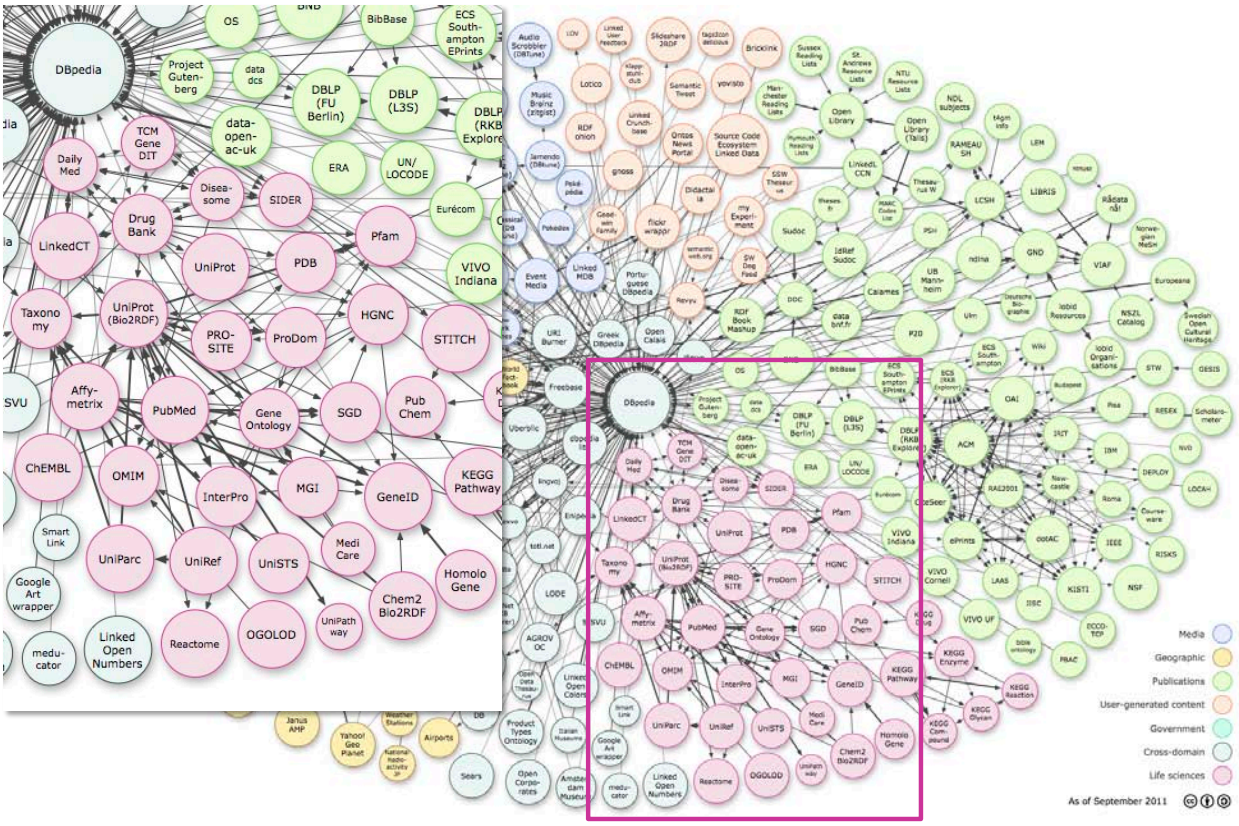


Figure 1. Linked Open Data cloud as of September 2011 (with close-up view on the life sciences portion)

2.2 Medical Subject Headings (MeSH)

The MeSH thesaurus is a controlled vocabulary produced by the National Library of Medicine (NLM) and used for indexing, cataloging and searching for biomedical and health-related information and documents [11]. MeSH consists of three main record types: Descriptor records, Qualifier records and Supplementary Concept records (SCRs). Each record has a unique identifier. Descriptors, also known as Main Headings, are mostly used to indicate the subject of an indexed item in NLM's MEDLINE bibliographic database and other databases. *Acquired immunodeficiency Syndrome* (D000163) is an example of a descriptor. Qualifiers, also known as subheadings, are used for indexing and cataloging in conjunction with descriptors, to indicate which specific aspect of a descriptor is discussed. For example, the qualifier *adverse effects* can be used with drug descriptors to index adverse drug events. In addition to de-

scriptors, SCRs are used by annotators to index new or less frequently occurring terms in the literature. All SCRs are connected to at least one descriptor ("heading mapped to" in MeSH parlance). In some cases, SCRs can be mapped to multiple descriptors or to descriptor-qualifier combinations. MeSH is easily accessible via the MeSH Browser [19] and made available by NLM for download in various formats, including XML, MARC, and ASCII flat files, as well as through the Unified Medical Language System (UMLS) Metathesaurus [12].

Three features of MeSH make it non-standard. These idiosyncratic features are: a 3-level structure (descriptor / concept / term), a contextual hierarchical structure and the need in common use cases for entity combinations (descriptor-qualifier combinations) that are not materialized in MeSH.

3-level structure. Instead of the traditional concept-term terminological model of most thesauri, MeSH uses a 3-level structure. In addition to concepts and terms (i.e., concept names), MeSH also defines descriptors, i.e., small aggregates of concepts grouped together as needed to support indexing and retrieval. For example, the MeSH descriptor *Ofloxacin* (D015242) groups several concepts, including the main concept identified by M0023430 (with terms *Ofloxacin* and *Ofloxacine*), the concept for a salt of this drug, M0329515 (with term *Ofloxacin Hydrochloride*), and the concept for the experimental form of this drug before it was marketed, M0023432 (with name *Ru-43280*). This 3-level structure is not amenable to representation with standard terminological models, such as SKOS, which only accommodates concepts and terms.

Contextual hierarchical structure. In addition to a non-standard terminological model, MeSH also uses a non-standard hierarchical organization. The hierarchy among MeSH descriptors is indicated through "tree numbers" assigned to descriptors. Tree number inclusion reflects that the descriptor with the longer tree number is narrower than that with the shorter tree number. For example, the tree number for *Liver* [A03.620] has an additional node (.620) compared to that of *Digestive System* [A03], indicating the narrower relation between the two. Note that tree numbers are not the unique identifiers of descriptors. Descriptors often have multiple tree numbers reflecting particular aspects of the descriptors, each aspect being assigned specific broader and narrower descriptors. For example, the descriptor *Eye* (D005123) has two tree numbers, A01.456.505.420 and A09.371. In the A01 tree, *Eye* is narrower than *Head* [A01.456] and broader than *Eyebrows* [A01.456.505.420.338] and *Eyelids* [A01.456.505.420.504], whereas, in the A09 tree, *Eye* is narrower than *Sense Organs* [A09] and broader than *Eyelids* [A09.371.337], *Retina* [A09.371.729], *Uvea* [A09.371.894] and nine other descriptors. Note that, although *Head* is broader than *Eye* (A01 tree), some descendants of *Eye* in the A09 tree (e.g., *Retina*) do not have *Head* as their ancestor. For all practical purposes, the broader/narrower relationship among MeSH descriptors is not transitive.

Descriptor-qualifier combinations. Finally, although MeSH represents descriptors and qualifiers as separate entities, common use cases of MeSH require combinations of descriptors and qualifiers. Chief among them is MEDLINE indexing, where descriptor-qualifier combinations are assigned to articles from the biomedical literature by indexers. For example, the descriptor-qualifier combination *Levofloxacin/adverse effects* is found as an index term for this report of a dermatological adverse drug event titled "Case of drug-induced bullous pemphigoid by levofloxacin" [20]. As mentioned earlier, MeSH itself uses descriptor-qualifier combinations to relate supplementary concept records to descriptors (and qualifiers). For example, the SCR for the drug *antofloxacin* (C522674) is mapped to *Ofloxacin/analogs & derivatives*, combining the descriptor *Ofloxacin* and the qualifier *analogs & derivatives*.

3 Related work

In the absence of an authoritative representation of MeSH in RDF from the NLM, there have been several efforts over the past few years to make MeSH available for the Sematic Web, starting from various sources, making use of different transformation techniques, and adopting different schemas and models. In the following, we review these existing representations of MeSH in RDF critically, in order to examine how the developers of RDF representations of MeSH have coped with the challenges associated with representing the three idiosyncratic features of MeSH discussed earlier. References for the six sources can be found in Table 1.

3.1 Original MeSH-SKOS

In 2004, Van Assem et al. generated what is probably the first representation of MeSH in RDF, leveraging the Simple Knowledge Organization System (SKOS) RDF Schema [21]. In addition to the RDF resource itself, these researchers made available the script they had developed to create it. Although the representation they produced was essentially for proof-of-concept purposes and was never updated, interested users can still apply their transformation to more recent versions of MeSH.

Because SKOS is a concept-based model, it cannot do justice to the distinction between descriptors and concepts in MeSH. As a consequence, all terms are directly attached to the descriptor in the SKOS representation, which may constitute a limitation for some applications. Hierarchical relations among MeSH descriptors are provided through *skos:broader* relations between descriptor URIs. Descriptor-qualifier combinations are not materialized. Of note, this proof-of-concept version does not provide a complete representation of MeSH (e.g., supplementary concept records are omitted).

3.2 Science Commons MeSH-SKOS and qualified-headings

In 2006, the Van Assem transformation was slightly modified by Science Commons researchers for use in the HCLSIG 2007 demo and was part of a mashup system that seeks to help the process of bioinformatics knowledge integration [22]. Additionally, Science Commons provides a companion resource (mesh/qualified-headings), in which MeSH descriptor-qualifier combinations are materialized, e.g., *mesh:D000001Q000008 skos:prefLabel "Calcimycin - administration & dosage"*. Although the data on the Science Commons website is for MeSH 2008, the script for creating the resources is available from the website.

The main MeSH-SKOS has the limitations of representations of MeSH in SKOS discussed earlier. Although descriptor-qualifier combinations are not materialized in the main SKOS representation, they are made available as a separate resource. This feature makes it possible to refer to these combinations in the representation of MEDLINE citations developed by the same researchers, in which *has-as-major-mesh* and *has-as-minor-mesh* relations are asserted between MEDLINE records and descriptor-qualifier combinations in MeSH. This representation is interesting as it reflects a specific use case, i.e., MEDLINE indexing, where citations are often indexed with descriptor-qualifier combinations. This representation is the only one we found that treats descriptor-qualifier combinations as first-class entities, with their own URIs.

3.3 Bio2RDF MeSH

Started in 2006, the Bio2RDF project uses Semantic Web technologies to provide linked data from publicly available databases in the life sciences [23, 24]. As of early 2014, there are 28 datasets available, including MeSH, linked together with normalized URIs, and sharing a common ontology. The current representation of MeSH in Bio2RDF is derived from the original MeSH 2014 ASCII flat files, containing all descriptor, supplementary concept, and qualifier records, their relations, and metadata.

Bio2RDF MeSH does not distinguish between descriptors and concepts in MeSH and all terms are directly attached to the descriptor as literals through proprietary *entry-term* relations. Hierarchical relations among MeSH descriptors are provided through *rdfs:subClassOf* relations, while MeSH only asserts broader/narrower relations. Moreover, Bio2RDF hierarchical relations are between materialized "tree-number classes" (i.e., specific aspects of the descriptors), e.g., *mesh:A08.186.211.132.93 rdfs:subClassOf mesh:A08.186.211.132*. Descriptors are linked to these tree-number classes through *mesh_vocabulary:mesh-tree-number* relationships. Descriptor-qualifier combinations are not materialized and supplementary concept records are mapped to a literal representation of the combination instead (e.g., *mesh:C014481 mesh_vocabulary:heading-mapped-to "Codeine/*analogs & derivatives"*), which is suboptimal from a Linked Data perspective.

3.4 NCBO BioPortal UMLS-MESH

Developed by the National Center for Biomedical Ontology at Stanford University since 2006, BioPortal is an open repository of biomedical ontologies made accessible via web services and web browsers [25]. BioPortal now offers RDF versions of all its ontologies. It actually provides two versions of MeSH. The main version (MESH) is derived from various files from the Unified Medical Language System (UMLS) [12, 26] Metathesaurus distribution and is currently being used in more than ten projects (e.g., the Drug Interaction Knowledge Base). UMLS-MeSH is available for the 2014 version of MeSH.

Like most MeSH representations, the version in BioPortal does not distinguish between descriptors and concepts in MeSH. MeSH descriptors are subclasses (*rdfs:subClassOf*) of their broader descriptors according to the MeSH tree number hierarchy. The tree numbers themselves are linked to the descriptors through annotation properties. Descriptor-qualifier combinations are not materialized. Supplementary concept records (SCRs) are mapped to descriptors through *mapped_to* relations and, independently, to qualifier through *has_mapping_qualifier* relations, where applicable. Decoupling the mapping of SCRs to descriptors and qualifiers is problematic when an SCR is mapped to multiple descriptor-qualifier combinations, because there is no explicit statement of the association between descriptors and qualifiers in this case. Additionally, supplementary concept records are mapped to literals

created for descriptor-qualifier combinations (e.g., *"D007830/Q000002"*), which, here again, is suboptimal from a Linked Data perspective.

3.5 NCBO BioPortal RH-MeSH

The "Robert Hoehndorf version of MeSH" (RH-MeSH) is the second version of MeSH found in BioPortal. Unlike the version of MeSH derived from the UMLS presented earlier, this version was created for a specific purpose, i.e., to facilitate the use of the descriptor and SCR hierarchy. It is used in the cross-species phenotype network, PhenomeNet. RH-MESH is represented in OWL. All MeSH descriptors and SCRs are classes as expected. Additionally, tree-numbers (representing specific aspects of descriptors) are also treated as classes and linked to other tree number classes through *rdfs:subClassOf* relationships. SCRs are represented as subclasses of descriptors. MeSH descriptors for drugs are also subclasses of the descriptors corresponding to their pharmacological actions. Except for class labels, this version of MeSH does not expose any other properties of the descriptors (e.g., definition). RH-MeSH is available for the 2014 version of MeSH.

While blurring the distinction between descriptors and SCRs (and even between descriptors and their tree numbers), this representation of MeSH provides an easy way for traversing the tree of MeSH entities. However, because it assumes that the descriptors are linked through subclass relationships, which is not what MeSH asserts, it contains inaccurate assertions (e.g., *Liver* subclass of *Digestive system*). Moreover, it links SCRs to descriptors (and not descriptor-qualifier combinations), and considers these links subclass relations (not mapping relations as indicated in MeSH), which also results in inaccurate assertions. For example, the MeSH assertion *Acrorenal Syndrome* mapped to *kidney/abnormalities* (descriptor-qualifier combination), is wrongly translated into *Acrorenal Syndrome* subclass of *kidney*.

3.6 MOR MeSH baseline

In 2009, in order to support internal research projects in the Medical Ontology Research (MOR) group at NLM, we created a fully automated process to transform the native XML representation of MeSH into RDF, based on EXtensible Stylesheet Language Transformations (XSLTs). Our goal with this representation was that it be close to the original XML representation and lossless. In other words, we made sure that all the information and only the information in the source XML had been captured during the transformation into RDF. We have updated this simple baseline representation regularly by applying our XSLTs to each new release of MeSH, but have kept both the XSLTs and RDF output internal to our research group.

The 3-level structure of MeSH with all relations between descriptors, concepts, and terms is preserved in this version. The descriptor hierarchy has been created for convenience, but kept in a separate graph, because it was not present in the native XML representation. The linkage of Supplementary Concept Records to descriptors and qualifiers is implemented through blank nodes rather than materialized descriptor-qualifier combinations, which is suboptimal as blank nodes lack shared semantics.

Table 1. Availability of existing RDF representations of MeSH

Name	Dissemination type	URL
Original MeSH-SKOS	Web site Download	http://thesauri.cs.vu.nl http://thesauri.cs.vu.nl/mesh/rdf/mesh1a.rdf
Science commons mesh-skos and qualified-headings	Web site Endpoint Download	http://neurocommons.org/page/Bundles/ http://beta.neurocommons.org http://neurocommons.org/page/Bundles/mesh/mesh-skos/ http://neurocommons.org/page/Bundles/mesh/qualified-headings/
Bio2RDF MeSH	Web site Endpoint Download	http://bio2rdf.org http://mesh.bio2rdf.org/sparql/ http://download.bio2rdf.org/release/2/mesh/
NCBO BioPortal UMLS MESH	Web site Endpoint Download	http://bioportal.bioontology.org http://sparql.bioontology.org http://bioportal.bioontology.org/ontologies/MESH/
NCBO BioPortal RH-MESH	Web site Endpoint Download	http://bioportal.bioontology.org http://sparql.bioontology.org http://bioportal.bioontology.org/ontologies/RH-MESH/
MOR MeSH baseline	Download/Endpoint	Not publicly available

4 Methods and Results

In this study we performed a review of existing representations of MeSH in RDF, established a list of desirable features for an authoritative representation, and implemented a prototype version of MeSH in RDF according to these criteria.

4.1 Analysis of the characteristics of existing RDF representations of MeSH

We conducted a manual analysis of the six existing representations of MeSH in RDF introduced in the Background section. We downloaded all representations that we reviewed and accessed them through their SPARQL endpoint whenever possible. However, we did not test any of the transformation scripts and did not create any local MeSH representations based on those.

We established a list of the characteristics of these resources, while focusing on the following features. We used the latest release date as an indication of the currency of the resource (the 2014 version of MeSH, available since September of 2013, was expected to be found). We categorized a representation as lossless only if the complete information provided in MeSH was exposed in the RDF representation. More specifically, we expected coverage of all three components of MeSH (descriptors, qualifiers and SCRs), as well as all important features (e.g., definitions). We also expected the semantics of MeSH relations to be preserved (e.g., hierarchical relations among descriptors represented as broader relations, not subclass relations). We noted the format(s) in which the resources were made available (e.g., RDF, OWL) and which specific terminological model or schema was used (e.g., SKOS). Some resources were developed as proof-of-concept and never intended to be maintained regularly, while others were in stable or beta version. We recorded this distinction. Whenever available, we added the information about the generation mechanism, as well as the original MeSH source used for creating the RDF representation (MeSH XML, MARC, or ASCII flat files, or UMLS Metathesaurus files). In terms of dissemination, we recorded whether the RDF resources were available for download or could be queried through a SPARQL endpoint, and whether the developers made the scripts used for the generation of RDF available to the community. Finally, we recorded how the idiosyncratic features of MeSH had been represented, especially hierarchical relations and descriptor-qualifier combinations.

Table 2. Characteristics of existing RDF representations of MeSH

Name	Latest Release Date	Lossless	Format	Status	Conversion	Dissemination	Features	
							Descriptor-Qualifer comb.	Hierarchical relations
Original MeSH-SKOS	2004	No	RDF (SKOS)	Proof-of-concept	Perl script, XSLT, XML	Download, Script	No	Yes (broader)
Science commons mesh-skos and qualified-headings	2008	No	RDF (SKOS)	Proof-of-concept	Using eswc06 Perl script, XSLT, XML	Endpoint, Script	Yes	Yes (broader)
Bio2RDF MeSH	2014	No	RDF	stable	PHP, ASCII	Endpoint, Download, Script	No* (as literals)	Yes (subclass)
NCBO Bioportal UMLS MESH	2014 (UMLS 2014AA)	UMLS view on MeSH	RDF	stable	UMLS Metathesaurus files	Endpoint, Download	No* (two separate relations and literals)	Yes (subclass)
NCBO Bioportal RH-MESH	2014	No	OWL	Beta	unspecified	Endpoint, Download	No	Yes (subclass)
MOR MeSH baseline	2014	Yes	RDF	Internal	XSLT	Used only internally	No* (through blank nodes)	Yes* (broader, in a separate graph)

The results of our analysis are summarized in Table 2. Four resources are up to date (i.e., reflect MeSH 2014 as of July 2014), but this was not the case at an earlier stage of our exploration a few months ago. The other two versions, developed for proof of concept, are not expected to be up to date. Most versions only capture a subset of the MeSH features, rather than all the details present in the original source. Two versions use OWL, one SKOS, and the other three use RDF with no specific terminological model. Except for one, the providers offer the resource for download, and most also provide a SPARQL endpoint. The transformation script is made available in three cases. Regarding the features of special interest to us, we found that only one resource materializes descriptor-qualifier combinations in a way that is suitable for linking to a MEDLINE dataset. While all resources investigated offer some kind of representation of hierarchical relations between MeSH descriptors, it is worth noting that the semantics of hierarchical relations had been reinterpreted by half of the providers as subclass relations, as opposed to broader relations.

4.2 Desirable features for an authoritative representation of MeSH in RDF

Based on our analysis of the characteristics of existing representations of MeSH in RDF, we compiled a list of desirable features for an authoritative representation of MeSH in RDF. Not mentioned in this list are the best practices for publishing linked data, such as guidelines for creating URIs, which are applicable to all RDF datasets, not only authoritative representations of vocabularies such as MeSH [27].

Completeness: Given the multiplicity of use cases for MeSH, it is likely that a representation of MeSH in RDF will be used in different ways by different users. Therefore, we believe it is best to provide a systematic representation in RDF of all features present in the XML version of MeSH. At a minimum, the authoritative representation of MeSH in RDF should represent those features exposed through the UMLS, as some sources do. However, representing only a subset of the MeSH entities (e.g., omitting the SCRs) would not be an option for most use cases.

Usability: As mentioned earlier, the structure of MeSH is too complex to be represented with the terminological model of SKOS. On the other hand, an RDF representation limited to the features of MeSH explicitly present in the XML version (such as our original baseline version), lacks the convenience of exposing important features, including hierarchical relations among descriptors, and materialized descriptor-qualifier combinations. Usability of the authoritative representation of MeSH in RDF should be analyzed in light of major use cases, which for MeSH include the role it plays in indexing and retrieval of the biomedical literature (MEDLINE). As illustrated by RH-MeSH, the authoritative representation could be extended by users in order to further facilitate traversal of the MeSH tree (at the expense of the original semantics of some MeSH relations).

Linkability: The authoritative representation of MeSH in RDF is meant to be linked to other resources in the Semantic Web. Although an authoritative version of MEDLINE in RDF has not been released yet, it would be a prime candidate for interoperating with MeSH in the Linked Open Data cloud. This requires coordinated development of resources within the institution developing these resources. Moreover, it requires harmonization of base URIs and predicates wherever possible. The representation provided by Science Commons, illustrated in Figure 2, prefigures what a MeSH-MEDLINE combined subset would look like.

```
@prefix c: <http://purl.org/science/owl/sciencecommons/> .
@prefix m: <http://purl.org/commons/record/mesh/> .
<http://purl.org/commons/record/pmid/11696761>
         c:has-as-minor-mesh m:D000368 ;
         c:has-as-minor-mesh m:D002292 ;
         c:has-as-minor-mesh m:D002292Q000150 ;
```

Figure 2. MeSH-MEDLINE combined subset as provided by Science Commons

Currency: A new version of MeSH becomes available each year and resources such as MEDLINE are synchronized with new versions of MeSH once a year. The authoritative representation of MeSH in RDF needs to be available in a timely fashion, and in coordination with related resources.

Availability: The authoritative representation of MeSH in RDF should be made available for download alongside the XML representation and other legacy representations. Users could load MeSH locally in an RDF database. Additionally, the resource should be available through a SPARQL endpoint so that local installation is not a requirement for use. (To some extent, this dual distribution mechanism is no different from the provision of datasets for download and web services, as is the case for the UMLS, for example.)

Transparency: Besides providing the RDF data for download, the transformation programs used to generate the RDF resource (e.g., scripts, XSLT, etc.) should be made available. This will give users insights into the transfor-

mation process and allow them to create RDF files locally, or create variants as required by their specific use cases. Exposing the transformation process might also help the community detect potential errors and suggest improvements.

4.3 Towards an authoritative representation of MeSH in RDF

The source version of MeSH we used in our prototype is the XML version (i.e., the 2014 MeSH XML files). The transformation rules from XML to RDF were coded using the XSLT (Extensible Stylesheet Language Transformations) language. The DTD file of each record type (Descriptor, Supplementary Concept Record, Qualifier) informed the creation of an XSLT for each type of MeSH record. The XSLT files were applied to the source XML files using the saxon XSLT processor. In addition to the features of MeSH explicitly represented in the XML file (and already present in our earlier baseline version), we chose to represent some other features for convenience purposes, i.e., in order to increase usability. We created a hierarchy among descriptors using the *skos:broader* relationship. We also elected to materialize descriptor-qualifier combinations for all the allowable qualifiers of each descriptor, in order to support upcoming representations of MEDLINE in RDF.

Our current prototype version of an authoritative representation of MeSH in RDF has the following features:

- Completeness: The losslessness of our transformation can easily be demonstrated by regenerating the original XML from the RDF using another set of XSLT files.
- Currency: The XSLT can easily be applied to any new version of MeSH. No changes to the XSLT are required unless changes are made to the XML DTD of the MeSH records. Because the transformation is completely automated and fast, it is conceivable to produce a nightly build reflecting the addition of supplementary concept records (e.g., for internal use by NLM for indexing purposes).
- Availability: The RDF file can easily be made available on the MeSH website, alongside the XML and legacy representations. Additionally, NLM intends to make it available through a SPARQL endpoint.
- Transparency: The XSLT files can be distributed on the MeSH website together with the XML and RDF versions.

It is difficult to comment on linkability at this prototype stage. However, the recently created NLM Linked Data Infrastructure Working Group will oversee the development of the final version of the authoritative representation of MeSH in RDF and of the other RDF datasets NLM intends to make available as Linked Data. The addition of descriptor-qualifier combinations to the prototype representation of MeSH in RDF prefigures the availability of an interoperable version of MEDLINE in RDF.

Usability is probably the most difficult criterion to fulfil and evaluate, due to the multiplicity of use cases for MeSH, beyond NLM's own use cases for indexing and retrieval of the biomedical literature. As already mentioned, while a complete, lossless version may best serve some complex use cases, other, more common use cases may be best served by simpler representations. Input from the community will help NLM determine which trade-offs to adopt for the final representation.

5 Discussion

In the following we provide several use cases illustrating the application of our prototype authoritative representation of MeSH in RDF. We also discuss current limitations and future directions for our work.

5.1 Use cases for an authoritative representation of MeSH in RDF

Currency: The Bibliographic Framework Initiative (BIBFRAME) [8, 10] initiated by the Library of Congress (LOC) provides a model for expressing and connecting bibliographic data on the web to replace the current standard for bibliographic exchange (MARC 21) [9]. BIBFRAME catalogues works and their instances, and associates them with authorities, which are resources that represent persons, organizations, topics, etc., (e.g., LOC subject headings can be accessed through the URI namespace <http://id.loc.gov/authorities/subjects/>). MeSH, as an authoritative resource, should be used for providing authorities for biomedical subjects (such as diseases, treatments, etc.). URIs for descriptors and SCRs should be added to catalog records where applicable. In BIBFRAME, topics can be assigned in combination with publication types. An extension of the representation of MeSH in RDF – possibly local to a given cataloging site – could include descriptor-publication type combinations (similar to the descriptor-qualifier combinations included in our prototype).

Quality assurance of MeSH: One of the key advantages of linked data is the possibility to link information across different data sources. But even for a single, locally available RDF graph, SPARQL queries can help gather infor-

mation from the data, which would be difficult to retrieve using other representations (e.g. flat files, XML). We helped the MeSH development team detect and remove cyclic relationships in the MeSH graph and assess that the 2014 version of MeSH is acyclic. Other applications include summarizing all supplementary concept records for a given pharmacologic action.

Linked data applications: In a recent collaboration with the FDA Center for Drug Evaluation and Research (CDER), we developed a novel analytic tool for quantitative drug-adverse event (ADE) safety signal detection based on mining the biomedical literature (MEDLINE). We leveraged the MeSH indexing terms to extract associations between co-occurring drug entities (in the context of adverse effects) and clinical manifestations (induced by chemicals). Information about ADEs is captured by different kinds of entities in MeSH (main headings, pharmacological actions, and supplementary concept records) and their inter-relations. In addition to the representation of MeSH in RDF, we created a prototype version of MEDLINE in RDF for the subset of articles under investigation in our study. The use of Semantic Web technologies enabled us to perform complex queries across the MeSH and MEDLINE datasets and greatly facilitated our work.

5.2 Limitations and future work

This prototype of an authoritative representation of MeSH in RDF is current and complete, but not final. Important improvements have been made already to our original baseline version, including the explicit representation of the hierarchical relations among descriptors (to improve usability) and the materialization of descriptor-qualifier combinations (to improve linkability with MEDLINE). However, additional editorial changes have to be made. For example, base URIs, namespaces and predicate names were chosen somewhat arbitrarily in the early development phase, where the focus was on demonstrating feasibility and scalability of the transformation method. However, these elements become important as this prototype is evolving into the authoritative representation of MeSH in RDF. Feedback from the user community will also inform future developments.

In order to accommodate some features of the XML version, our baseline version had introduced blank nodes to represent, for example, entry combinations and concept relations. While developing the current prototype, we critically reexamined earlier design choices and found that in most cases blank nodes were unnecessary or could be replaced by materialized combinations (e.g., for descriptors and qualifiers). All the blank nodes created initially were removed.

6 Conclusions

In the absence of an authoritative version of MeSH in RDF from the NLM, there have been several efforts over the past few years to make MeSH available for the Sematic Web. We identified six existing representations of MeSH in RDF and conducted a manual analysis of these representations. Based on the characteristics of these resources we compiled a list of desirable features for an authoritative representation of MeSH in RDF (completeness, usability, linkability, currency, availability and transparency). We implemented a prototype of an authoritative representation of MeSH in RDF that fulfills these criteria and illustrated its suitability on three use cases. Our prototype was influenced by our early baseline representation in RDF of all features present in the XML version of MeSH. However, we made substantial changes in order to improve usability and linkability. NLM intends to release an authoritative representation of MeSH in RDF in the Fall of 2014. We believe that the availability of such a resource will foster the adoption of MeSH in biomedical Semantic Web applications.

Acknowledgements

This work was supported by the Intramural Research Program of the NIH, National Library of Medicine (NLM). The authors would like to thank former NLM colleagues Ramez Gazzaoui and Genaro Hernandez, who contributed to an early prototype of MeSH in RDF in 2009, and Nancy Fallgren for sharing her insights on the BIBFRAME project. Thanks to the NLM Linked Data Infrastructure Working Group and to colleagues from the MeSH development team for their encouragement and support.

References

1. Berners-Lee T, Hendler J, Lassila O. The Semantic Web - A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Sci Am. 2001 May;284(5):34-+.
2. Ruttenberg A, Clark T, Bug W, Samwald M, Bodenreider O, Chen H, et al. Advancing translational research with the Semantic Web. BMC Bioinformatics. 2007;8 Suppl 3:S2.
3. Cyganiak R, Jentzsch A. LOD Cloud. Available from: http://lod-cloud.net/.
4. Redaschi N, Consortium U. UniProt in RDF: Tackling Data Integration and Distributed Annotation with the Semantic Web.: Nature Precedings; 2009; Available from: http://precedings.nature.com/documents/3193/version/1.
5. Biebl M, Hakaim AG, Hugl B, Oldenburg WA, Paz-Fumagalli R, McKinney JM, et al. Endovascular aortic aneurysm repair with the Zenith AAA Endovascular Graft: does gender affect procedural success, postoperative morbidity, or early survival? Am Surg. 2005 Dec;71(12):1001-8.
6. Hassanzadeh O, Kementsietsidis A, Lim L, Miller RJ, Wang M. LinkedCT: A Linked Data Space for Clinical Trials. CoRR. 2009;abs/0908.0567.
7. da Silva KR, Costa R, Crevelari ES, Lacerda MS, de Moraes Albertini CM, Filho MM, et al. Glocal clinical registries: pacemaker registry design and implementation for global and local integration--methodology and case study. PLoS One. 2013;8(7):e71090.
8. Pathak J, Kiefer RC, Bielinski SJ, Chute CG. Applying semantic web technologies for phenome-wide scan using an electronic health record linked Biobank. J Biomed Semantics. 2012;3(1):10.
9. MARC 21 format for bibliographic data 1999 Edition Update No. 17. Library of Congress; 2013 [cited 2014 March 10]; Available from: http://www.loc.gov/marc/bibliographic/.
10. Miller E, Ogbuji U, Mueller V, MacDougall K. Bibliographic Framework as a Web of Data: Linked Data Model and Supporting Services. Washington, DC: Library of Congress 2012 November 21.
11. Nelson SJ, D. JW, L. HB. Relationships in Medical Subject Headings (MeSH). In: Bean CA, Green R, editors. Relationships in the organization of knowledge. Dordrecht; Boston: Kluwer Academics Publishers; 2001. p. 171-84.
12. NLM. Unified Medical Language System (UMLS). 2013; Available from: https://uts.nlm.nih.gov/.
13. Tse LW, Steinmetz OK, Abraham CZ, Valenti DA, Mackenzie KS, Obrand DI, et al. Branched endovascular stent-graft for suprarenal aortic aneurysm: the future of aortic stent-grafting? Can J Surg. 2004 Aug;47(4):257-62.
14. W3C. URI. Available from: http://www.w3.org/TR/uri-clarification/.
15. W3C. RDF. Available from: http://www.w3.org/RDF/.
16. W3C. SPARQL. Available from: http://www.w3.org/TR/sparql11-overview/.
17. W3C. OWL. Available from: http://www.w3.org/TR/owl2-overview/.
18. W3C. SKOS. Available from: http://www.w3.org/2004/02/skos/.
19. NLM. MeSH Browser. 2014; Available from: https://www.nlm.nih.gov/mesh/MBrowser.html.
20. Ma HJ, Hu R, Jia CY, Yang Y, Song LJ. Case of drug-induced bullous pemphigoid by levofloxacin. J Dermatol. 2012 Dec;39(12):1086-7.
21. van Assem M, Menken MR, Schreiber G, Wielemaker J, Wielinga B, editors. A Method for Converting Thesauri to RDF/OWL. 3rd Int'l Semantic Web Conf (ISWC'04); 2004: Springer-Verlag.
22. Kanda J, Kaynar L, Kanda Y, Prasad VK, Parikh SH, Lan L, et al. Pre-engraftment syndrome after myeloablative dual umbilical cord blood transplantation: risk factors and response to treatment. Bone Marrow Transplant. 2013 Jan 21.
23. Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. J Biomed Inform. 2008 Oct;41(5):706-16.
24. Callahan A, Cruz-Toledo J, Dumontier M. Ontology-Based Querying with Bio2RDF's Linked Open Data. J Biomed Semantics. 2013 Apr 15;4 Suppl 1:S1.
25. Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W170-3.
26. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D267-70.
27. W3C. Best Practices for Publishing Linked Data. [updated 2014]; Available from: http://www.w3.org/TR/ld-bp/.

Comparing the Value of Mammographic Features and Genetic Variants in Breast Cancer Risk Prediction

Yirong Wu, PhD[1], Jie Liu, MS[1], David Page, PhD[1], Peggy Peissig, PhD[2], Catherine McCarty, PhD[3], Adedayo A. Onitilo MD, MSCR, FACP[2,4,5], and Elizabeth S. Burnside, MD, MPH, MS[1]

[1] University of Wisconsin, Madison, WI, USA;[2] Marshfield Clinic Research Foundation, Marshfield, WI, USA; [3] Essentia Institute of Rural Health, Duluth, MN, USA;[4] Department of Hematology/Oncology, Marshfield Clinic Weston Center, Weston, WI, USA; [5] School of Population Health, University of Queensland, Brisbane, Australia

Abstract

The goal of this study was to compare the value of mammographic features and genetic variants for breast cancer risk prediction with Bayesian reasoning and information theory. We conducted a retrospective case-control study, collecting mammographic findings and high-frequency/low-penetrance genetic variants from an existing personalized medicine data repository. We trained and tested Bayesian networks for mammographic findings and genetic variants respectively. We found that mammographic findings had a higher discriminative ability than genetic variants for improving breast cancer risk prediction in terms of the area under the ROC curve. We compared the value of each mammographic feature and genetic variant for breast risk prediction in terms of mutual information, with and without consideration of interactions of those risk factors. We also identified the interactions between mammographic features and genetic variants in an attempt to prioritize mammographic features and genetic variants to efficiently predict the risk of breast cancer.

Introduction

Technology advances in genome-wide association studies (GWAS) and successes with cost reduction in genome-sequencing have engendered optimism that we have entered a new age of precision medicine[1], in which the risk of a breast cancer can be predicted on the basis of a person's genetic variants. More recently, however, the optimism of these studies has been tempered by disappointment and caution[2, 3]. Now it is widely agreed that phenotypic data, in concert with genetic variants will likely be necessary for advancing personalized breast cancer risk prediction. The availability of imaging findings acquired from mammography provides the opportunity to combine mammographic features and genetic variants to improve risk prediction[4, 5]; however, the comparative value that mammographic features and genetic variants provide for advancing risk prediction is unknown. The potential to quantify the value and prioritize phenotypic data (mammographic features) and genotypic data (germline genetic variants) for clinical decision-making presents an exciting opportunity to explore feature ranking algorithms for optimizing breast cancer risk prediction.

Mutual information analysis has been widely used to rank variables by quantifying the information that each variable provides for estimating the outcomes of interest [6, 7]. Prior studies have explored mutual information for genome-wide data analysis to discover association between single nucleotide polymorphisms (SNPs) and disease; however, few have considered interactions between risky SNPs[8]. Many believe that epistatic interactions of SNPs are important in determining susceptibility to breast cancer as well as disease mechanism[2, 9] . Hence, recent studies propose to utilize mutual information analysis for joint analysis of multiple SNPs that are potentially associated with breast disease[10]. Coincident to those studies, mutual information analysis has also been used to identify diagnostically important mammographic features. Most of these studies selected only the top-ranked features without considering interactions among features[11]. Recently investigators have used multidimensional mutual

information analysis to rank mammographic features by considering interactions when ranking for feature selection[12]. Overall, mutual information analysis has been successfully utilized to rank either SNPs or mammographic features for breast cancer risk estimation. However, to our knowledge, few studies have attempted to select the most important risk factors from a combination of mammographic features and genetic variants, and fewer have investigated interactions between mammographic features and genetic variants.

In this study, we aim to compare the values of mammographic features and SNPs for selecting the most informative risk factors in the diagnosis of breast cancer. We use mutual information analysis and Bayesian reasoning by considering interactions among risk factors to select the most valuable mammographic features and SNPs in breast cancer risk estimation.

Materials and Methods

The Marshfield Clinic Institutional Review Board approved the use of Marshfield Clinic's Personalized Medicine Research Project (PMRP) cohort in the research.

Subjects

The PMRP cohort, details of which have been previously published[13], was used in this study. To summarize, Marshfield Clinic patients residing in one of 19 ZIP code areas surrounding Marshfield, WI and aged 18 years or older, were invited to participate in PMRP. Written informed consent was obtained from each participant along with a blood sample from which DNA, plasma and serum were extracted and stored. Permission was given by each participant to link their electronic health record information to biological samples for use in the research.

We used PMRP to select subjects with an available DNA sample, a mammogram, a breast biopsy within 12 months after the mammogram from the Marshfield Clinic Data Warehouse. We employed a retrospective case-control study design. Cases were defined as women having a confirmed diagnosis of breast cancer obtained from the cancer registry, which includes either invasive breast cancer (ductal and lobular) or ductal carcinoma in situ. Controls were determined through the electronic medical records (and absence from the cancer registry) as never having had a breast cancer diagnosis. To construct case and control cohorts that were similar in age distribution, we employed an age matching strategy to ensure that the age of the matched control was within 5 years of the case.

Mammography Features

The American College of Radiology developed the Breast Imaging Reporting and Data System (BI-RADS) lexicon to standardize mammographic findings and recommendations[14]. The BI-RADS lexicon consists of a number of mammographic features, including the characteristics of masses and micro-calcifications, breast composition and other associated findings, which can be organized in a hierarchy (Figure 1). In Marshfield Clinic's electronic health record, mammographic findings including breast composition were described in BI-RADS lexicon and embedded in free text clinical reports, from which we used a parser to extract 46 mammographic features[15].

Genetic Variants

Our study focused on high-frequency/low-penetrance genetic variants that affect breast cancer risk as opposed to low frequency genetic variants with high penetrance (BRCA1 and BRCA2) or intermediate penetrance (CHEK-2). In clinic, individuals with BRCA1 and BRCA2 mutations demonstrating a high risk of breast cancer are managed with more intensive screening and have options for chemoprevention. Our study was designed for normal risk individuals, for which recommendation of screening and chemoprevention options are less clear. We included 22 genetic variants which have been identified by recent large-scale genome-wide association studies[4] (Table 1). The SNPs used in Gail model[16, 17] and Wacholder et al study[18] were included in our study. When we built the models

with the genetic variants, we coded each genetic variant as whether the subject carries the minor allele, rather than the specific genotype the subject carries.

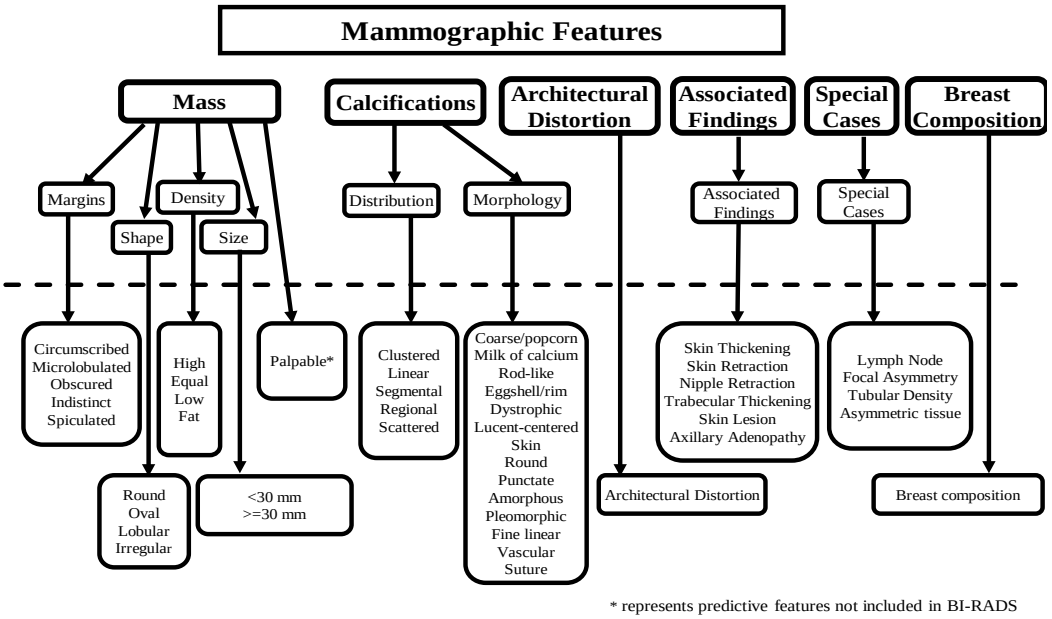


Figure 1. Mammographic features adopted from BI-RADS lexicon

Table 1. SNPs evaluated for breast cancer risk

SNP ID	Chromosome	Minor allele	SNP ID	Chromosome	Minor allele
rs11249433	1	C	rs2046210	6	T
rs4666451	2	A	rs13281615	8	G
rs13387042	2	G	rs2981582	10	T
rs1045485	2	C	rs3817198	11	C
rs17468277	2	T	rs2107425	11	T
rs4973768	3	T	rs6220	12	G
rs10941679	5	G	rs999737	14	T
rs981782	5	G	rs3803662	16	T
rs30099	5	T	rs8051542	16	T
rs889312	5	C	rs12443621	16	G
rs2180341	6	G	rs6504950	17	A

Mutual Information

Originating from Shannon's information theory[7], mutual information (MI) of a variable v_1 with respect to the other variable v_2 is defined as the amount by which the uncertainty of v_1 is decreased with the knowledge that v_2 provides. The initial uncertainty of v_1 is quantified by entropy $H(v_1)$. The average uncertainty of v_1 given knowledge of v_2 is conditional entropy $H(v_1|v_2)$. The difference between initial entropy and conditional entropy represents therefore MI of v_1 with respect to v_2. MI is defined as follows:

$$\mathrm{MI}(v_1; v_2) = H(v_1) - H(v_1|v_2) = \sum_{v_2}\sum_{v_1} p(v_1, v_2) log \frac{p(v_1, v_2)}{p(v_1)p(v_2)}$$

where $p(v_1)$ and $p(v_2)$ are the marginal probability of v_1 and v_2, and $p(v_1, v_2)$ is their joint probability.

In the following context, we use $\mathrm{MI}(x_1; x_2)$ to denote the information value that one risk factor x_1 (either a mammographic feature or a SNP) provides for estimating the other risk factor x_2 to quantify the interaction between them. We use single-dimensional mutual information $\mathrm{SMI}(x; y)$ to denote the information that one risk factor x provides for estimating the outcome y (breast cancer).

$$\mathrm{SMI}(x; y) = H(x) - H(x|y) = \sum_{y}\sum_{x} p(x, y) log \frac{p(x, y)}{p(x)p(y)}$$

where x is one of risk factors and y is the outcome. SMI does not take into account interaction effects among risk factors.

We use multidimensional mutual information $\mathrm{MMI}(x; y)$ to denote the information that one risk factor x provides for estimating the outcome y when interaction with other risk factors is considered. We assess MMI by an algorithm that **m**inimizes **R**edundancy among risk factors while **M**aximizing **R**elevance to the outcome (mRMR)[19-22]. "Redundancy" is related to MI of risk factors with each other, and "relevance" is defined as SMI of risk factors with the outcome. Specifically, in this study, to use mRMR algorithm to rank most important risk factors, we choose the risk factor with the highest SMI as the most important one. We select subsequent important risk factors sequentially, such that each risk factor simultaneously maximizes its SMI and minimizes MI between the risk factor of interest and already selected risk factors. Specifically, we choose the next most important risk factor x_i, $i = 2, 3, 4, \cdots$, that maximizes

$$\mathrm{SMI}(x_i;\ y) - \frac{1}{i-1}\sum_{j<i} \mathrm{MI}(x_i; x_j)$$

where risk factor x_j is one of important risk factor selected ahead of x_i and y is the outcome.

Study Design and Statistical Analysis

To quantify the predictive performance of mammographic features and SNPs, we trained and tested Bayesian networks (BN) using a tree augmented naïve Bayes algorithm on all 46 mammographic features and 22 SNPs respectively in Weka (Weka, version 3.6.4; University of Waikato, Hamilton, New Zealand)[23]. We chose a BN as our prediction method since it has a clear semantic interpretation of model parameters[24]. We employed 10-fold cross-validation to help confirm the validity of predictions. We constructed a receiver operating characteristic (ROC) curve based on estimated probabilities from the BN, and obtained the area under the ROC curve (AUC) as a measure of overall predictive performance. We compared AUCs for 46 mammographic features and 22 SNPs by using the DeLong method[25], implemented in MATLAB software (MathWorks, Natick, MA). We used a P-value of 0.05 as the threshold for statistical significance testing to determine the difference between two AUC values.

We calculated SMI of each risk factor (either a mammographic feature or a SNP) with respect to the outcome, and ranked all risk factors according to SMI values. To evaluate the performance of SMI ranking approach, we first constructed risk factor sets by sequentially selecting the most informative risk factors, one by one, in order of SMI values. Then, using 10-fold cross-validation, we trained and tested BNs on the sets of sequentially selected risk factors. We created ROC curves based on estimated probabilities from the BNs, and obtained AUCs.

We then calculated MMI of each risk factor by using mRMR algorithm and ranked all risk factors based on those MMI values. Using a similar procedure of evaluating SMI rankings, we assessed the performance of MMI ranking approach. We first created risk factor sets with sequentially selected risk factors, one by one, in order of

MMI values and then trained BNs with those risk factor sets. We created ROC curves based on estimated probabilities from the BN, and obtained AUCs.

We also used "parsimony" to describe the performance of two ranking approaches. We define parsimony here as the smallest number of the risk factors needed to reach a performance level such that there is no significant difference of AUC as compared with the maximum AUC. The difference of parsimony between SMI and MMI describes the interaction effects among risk factors.

Results

We succeeded in identifying 373 cases and 395 controls, all European American. The age range for these subjects was 29 to 90 years of age (mean = 62, standard deviation = 12.8). Specifically, 3.12% subjects were less than 40 years old, 16.28% subjects were in between 40 and 49 years old, 27.34% subjects were in between 50 and 59, 22.92% subjects were in between 60 and 69, and 30.34% subjects were older than 69 years old.

Both mammographic features and SNPs demonstrated substantial capability of improving breast cancer risk prediction (Figure 2). We found that AUC value from BN trained and tested with mammographic features was significantly higher than that without mammographic features (SNPs alone) (0.736 vs. 0.581, P-value < 0.001). The difference represents the value provided from mammographic features for breast cancer risk estimation. We also found that AUC value from BN trained and tested with SNPs improved significantly compared with that without SNPs (mammographic features alone) (0.736 vs. 0.704, P-value = 0.012). Furthermore, the risk estimation using only mammographic features was superior to using only SNPs (0.704 vs. 0.581, P-value < 0.001). Notably, one mammographic feature (spiculated mass margins) could achieve a similar predictive performance as all 22 SNPs in terms of AUC (0.590 vs. 0.581, P-value = 0.765).

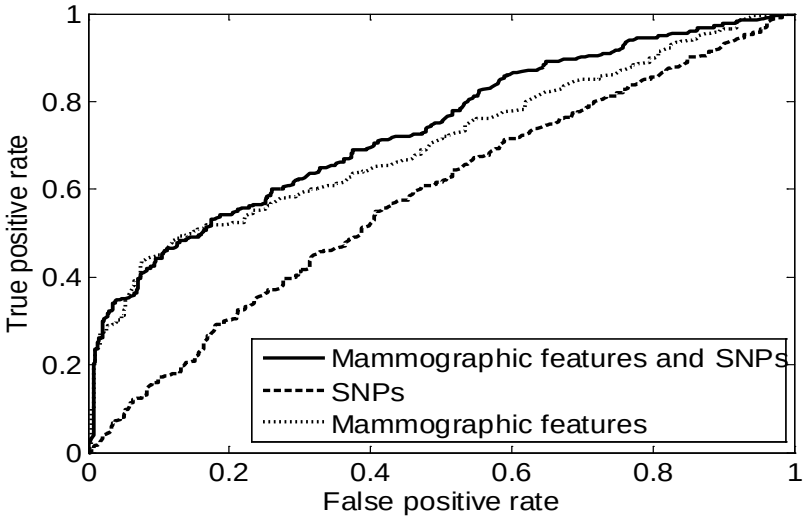


Figure 2. ROC curves for mammographic features, SNPs, and a combination of mammographic features and SNPs.

Both SMI and MMI approaches were able to prioritize breast cancer risk factors. However, in terms of parsimony, the MMI approach outperformed SMI. Specifically, when we calculated AUC values to evaluate SMI ranking results, we observed that AUC values increased as more features were included (solid curve with triangle data points in Figure 3). Parsimony for the SMI approach (the number of features needed to reach a non-significant difference between SMI AUC and maximum AUC—0.733 vs. 0.752, P-value = 0.053) was twelve features. For MMI ranking results, we observed that correspondent AUC values also increased as additional features were

included (dashed curve with cycle data points in Figure 3). Parsimony for the MMI approach (the number of features needed to reach a non-significant difference between MMI AUC and maximum AUC—0.733 vs. 0.752, P-value = 0.109) was eight. Both parsimony sets included mammographic features and SNPs (Table 2).

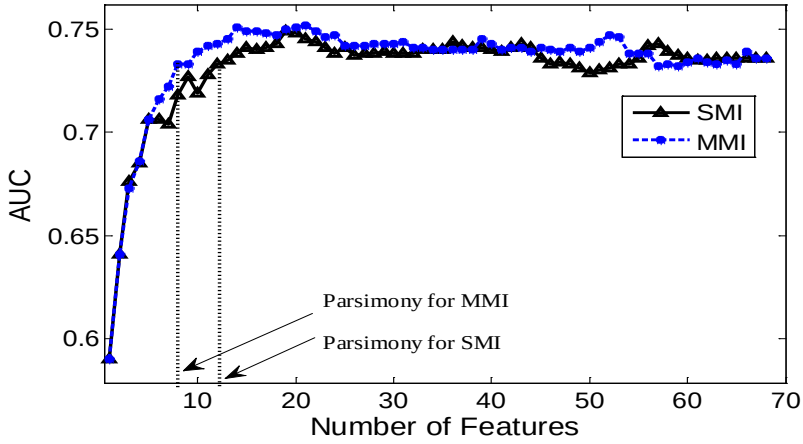


Figure 3. AUC changes with the number of selected features. Solid curve, SMI; Dashed curve, MMI.

Table 2. Ranking results based on SMI and MMI (partial)

Variables	SMI ranking	MMI ranking
Spiculated mass margins	1*	1*
Irregular mass shape	2*	2*
Mass size	3*	5*
Architectural distortion	4*	4*
rs13387042	5*	3*
Breast composition	6*	10
Clustered distribution	7*	25
rs8051542	8*	6*
Segmental distribution	9*	7*
Palpable	10*	15
rs10941679	11*	8*
Fine calcification	12*	9
Amorphous calcification	13	12
rs2107425	14	11
…….	……	……
* risk factors included in parsimony sets		

The rankings of mammographic features and SNPs were different for SMI and MMI approaches (Table 2). The difference demonstrated the effects of associations of mammographic feature-mammographic feature, SNP-

SNP, and mammographic feature-SNP. We found that the association of feature-SNP was relatively weak compared with the associations of feature-feature and SNP-SNP. For the three most informative mammographic features in breast cancer prediction, spiculated mass margin, irregular mass shape, and mass size, we observed that other mammographic features instead of SNPs were strongly associated with them when we used a 0.01 threshold in MI analysis (Table 3). Similarly, for the three most informative SNPs in breast cancer prediction, rs13387042, rs8051542, and rs10941679, we found that other SNPs instead of mammographic features were strongly associated them (Table 4). For all mammographic features and SNPs, pleomorphic classification-rs1045485/rs17468277 was the only mammographic feature-SNP pair demonstrating strong association.

Table 3. Strongly associated risk factors for the three most important mammographic features

Spiculated mass margins		**Irregular mass shape**		**Breast composition**	
MI	**Variables**	**MI**	**Variables**	**MI**	**Variables**
0.0203	Irregular mass shape	**0.0465**	Mass size	**0.0327**	Skin thickening
0.0184	Clustered distribution	**0.0321**	Clustered distribution	**0.0275**	Palpable
0.0179	Mass size	**0.0238**	Pleomorphic distribution	**0.0172**	Clustered distribution
0.0122	Breast composition	**0.0203**	Spiculated mass margins	**0.0131**	Dystrophic calcification
0.0118	Architectural distortion	**0.0121**	Focal asymmetry	**0.0129**	Nipple retraction
0.0098	Pleomorphic distribution	0.0085	Circumscribed margins	**0.0126**	Pleomorphic distribution
				0.0122	Spiculated margins
				0.0111	Coarse calcification
				0.0107	Obscured margins
				0.0104	Mass size
				0.0099	Focal asymmetry
MI values more than 0.01 are shown in bold.					

Table 4. Strongly associated risk factors for the three most important SNPs

rs13387042		**rs8051542**		**rs10941679**	
MI	**Variables**	**MI**	**Variables**	**MI**	**Variables**
0.0072	Breast composition	**0.0861**	rs3803662	**0.01654**	rs981782
		0.02104	rs12443621	0.0062	Breast composition
		0.0081	Linear calcification		
MI values more than 0.01 are shown in bold.					

Discussion

Both mammographic features and SNPs are important in breast cancer risk prediction. Our study shows that mammographic features can significantly improve breast cancer diagnosis over genetic variants. It is unlikely that germline genetic analysis alone will be sufficient to fulfill the promise of precision medicine in the arena of breast cancer risk prediction[26]. The results from this study advocate greater use of phenotypic data such as mammographic findings to predict breast cancer risk. Concurrently, our study shows that genetic variants can significantly improve breast cancer diagnosis over mammographic findings. Traditional oncologic risk estimation based on tissue analysis may be augmented by genetic variants. GWAS have identified a series of genetic variants underlying breast disease, which would greatly facilitate personalized cancer diagnosis. However, there are no standards for evidence-based genetic biomarker development and some genetics studies have yielded contradictory results[27]. There are many questions to be answered and barriers to be overcome before genetics can be most effectively used in the clinics to benefit patients. In summary, for maximal accuracy of next-generation breast cancer

risk prediction models, genetic variants and clinical data should be combined and tuned for better predictive performance.

Computational and experimental methods to detect interactions of risk factors are promising for the future of breast cancer screening and diagnosis. We find that multidimensional mutual information addresses the issue of interaction of risk factors and may have the potential to assist physicians in prioritizing predictive variables in order to select the most parsimonious set of mammographic features and genetic variants with the highest predictive ability. There are many risk factors potentially used to estimate breast cancer risk. SMI analysis demonstrated that twelve risk variables were needed to reach an AUC that wasn't significantly different from the maximum AUC. Performance analysis based on MMI reduced the number of important variables to eight. Assessing additional variables beyond these most important ones does not improve risk prediction accuracy. This suggestion of using a smaller set of important mammographic features and SNPs for breast cancer diagnosis may help improve breast cancer screening and diagnosis workflow in the future.

Mutual information approaches (SMI and MMI) have the capability of determining the most informative genetic variants for breast cancer diagnosis. Both SMI and MMI approaches demonstrated that rs13387042, rs8051542, and rs1094169 were the three most important risk-conferring factors in 22 SNPs for the development of breast cancer (Table 2). This result is in line with some of previous findings that these three SNPs are strongly associated with breast cancer[27-29]. In addition, both approaches showed that rs13387042 was the most important risk factor in 22 SNPs, which is consistent with our knowledge about its significance in breast cancer risk estimation. This SNP has been included in Gail model[16, 17] and Wacholder analysis[18] for breast cancer risk prediction. Overall, our results indicate that mutual information analysis may be an effective approach to evaluate the association between SNPs and breast cancer.

Recently the analysis of associations between imaging features and somatic mutations/molecular markers has attracted much interest in the imaging and biomarker community. This trend has led to widespread radiogenomic studies in the hopes of better diagnostic accuracy and more precise prognostication[30-32]. The results from our radiogenomic analysis reveal that some association between mammographic features and germline SNPs exists though their magnitude is less than mammographic feature-mammographic features associations or SNP-SNP associations. We found that pleomorphic classification-rs1045485/rs17468277 was the only feature-SNP pair demonstrating strong association (mutual information > 0.01), which is in concert with previous studies[4]. This sporadic association may be caused by several factors. First, GWAS has analyzed hundreds of thousands of SNPs to determine whether they are associated with breast cancer. Our study has used a very limited set of SNPs. Including a larger set of SNPs is crucial for the success of radiogenomic analysis. Second, our study considers association between one SNP and one mammographic feature only. More complex interactions of mammographic feature-mammographic feature and SNP-SNP will likely be a fruitful area of future research. Third, other imaging modality like MRI has proven to be more sensitive than mammography[31] or convey functional information (rather than anatomic data). It would be interesting to induce imaging features collected from MRI into our future radiogenomic studies.

Limitations and Future Work

There are several limitations to our study. First, the sample size is small compared with large-scale genome-wide association studies, due to the inherent difficulty of collecting a rich multi-modality dataset. Second, our study focused on discussion of predictive accuracy associated with risk factors but did not consider benefit and cost related to the decision. We plan to extend our study in this direction soon since cost-effectiveness analysis allows physicians and policymakers to compare the health gains that various decision of choosing the most important mammographic features and genetic variants can achieve. Third, our study used Bayesian networks to assess ranking results of mutual information analysis. A possible line of future research is to employ other prediction

algorithms such as logistic regression, artificial neural network, or support vector machine for validating our results. Finally, we used AUC to measure the performance of our Bayesian networks. The AUC evaluates the performance of a prediction method over the full range of possible threshold levels. In practice, however, only a limited range or a single optimal threshold level may be of interest clinically. We plan to extend our study by determining the optimal threshold level to measure the predictive performance of our Bayesian networks in terms of sensitivity and specificity.

Conclusion

Our study represents one of the first explorations of breast cancer risk prediction using genetic polymorphisms in combination with mammographic features. We find that genetic risk factors improve risk prediction to a statistically significant degree, which raises the possibility that stratification based on these risk factors may provide an opportunity to personalize care in clinical practice. This work confirms that disease prediction, which is narrowly focused on one data type (e.g. genomics) may miss opportunities for improved performance offered by incorporation of phenotypic data (e.g. mammographic features). We demonstrate that genetic risk factors can be combined and tuned with clinical imaging findings for better predictive performance. Moreover, considering interactions among risk factors, MMI outperforms SMI in determining the smallest set of informative risk factors. In applications where addition of risk factors incurs additional time or monetary cost, MMI may help reduce the cost of diagnostic testing. Encouraged by these promising results, we plan to further explore genotype/phenotype associations to shed light on disease processes that may, in the future, improve diagnosis and treatment.

Acknowledgements

The authors gratefully acknowledge the support of the Wisconsin Genomics Initiative, NCI grant R01CA127379-01 and its ARRA supplement 3R01CA127379-03S1, NIGMS grant R01GM097618-01, NLM grant R01LM011028-01, NIEHS grant 5R01ES017400-03, the UW Institute for Clinical and Translational Research (ICTR) and the UW Carbone Cancer Center.

References

1. Collins FS, Green ED, Guttmacher AE, Guyer MS. A vision for the future of genomics research. Nature. 2003;422.
2. Devilee P, Rookus MA. A tiny step closer to personalized risk prediction for breast cancer. N Engl J Med. 2010 Mar 18;362(11):1043-5.
3. Pharoah PD, Antoniou AC, Easton DF, Ponder BA. Polygenes, risk prediction, and targeted prevention of breast cancer. N Engl J Med. 2008 Jun 26;358(26):2796-803.
4. Liu J, Page D, Nassif H, Shavlik J, Peissig P, McCarty C, Onitilo AA, Burnside ES. Genetic variants improve breast cancer risk prediction on mammograms. Proceedings of the American Medical Informatics Association Symposium (AMIA); 2013; Washington, DC.
5. Liu J, Page D, Peissig P, McCarty C, Onitilo AA, Trentham Dietz A, Burnside ES. New genetic variants improve personalized breast cancer diagnosis. AMIA Summit on Translational Bioinformatics (AMIA-TBI); 2014; San Francisco, CA.
6. Benish W. Mutual information as an index of diagnostic test performance. Methods of Information in Medicine. 2003;42(3):260-4.
7. Shannon C, Weaver W. The Mathematical Theory of Communication. Urbana, IL: University of Illinois Press; 1949.
8. Yuan X, Zhang J, Wang Y. Mutual information and linkage disequilibrium based SNP association study by grouping case-control. Genes & Genomics. 2011;33:65-73.
9. Briollais L, Wang Y, Rajendram I, Onay V, Shi E, Knight J, Ozcelik H. Methodological issues in detecting gene-gene interactions in breast cancer susceptibility: a population-based study in Ontario. BMC Med. 2007;5:22.

10. Anunciacao O, Vinga S, Oliveira AL. Using information interaction to discover epistatic effects in complex disease. PLoS One. 2013;8(10):1-11.
11. Wu Y, Alagoz O, Ayvaci M, Munoz del Rio A, Vanness DV, Woods R, Burnside ES. A comprehensive methodology for determining the most informative mammographic features. J Digital Imaging. 2013;26(5):941-7.
12. Wu Y, Vanness DV, Burnside ES. Using multidimensional mutual information to prioritize mammographic features for breast cancer diagnosis. Proceedings of the American Medical Informatics Association Symposium (AMIA); 2013; Washington, DC.
13. McCarty CA, Wilke RA, Giampietro PF, Wesbrook SD, Caldwell MD. Marshfield Clinic Personalized Medicine Research Project (PMRP): design, methods and recruitment for a large population-based biobank. Personalized Med. 2005;2(1):49-79.
14. Breast Imaging Reporting And Data System (BI-RADS®). 4th ed. Reston VA: American College of Radiology; 2003.
15. Nassif H, Woods R, Burnside ES, Ayvaci M, Shavlik J, Page D. Information extraction for clinical data mining: a mammography case study. Proc IEEE Int Conf Data Min; 2009; Miami, Florida.
16. Gail MH. Discriminatory accuracy from single-nucleotide polymorphisms in models to predict breast cancer risk. J Natl Cancer Inst. 2008 Jul 16;100(14):1037-41.
17. Gail MH. Value of adding single-nucleotide polymorphism genotypes to a breast cancer risk model. J Natl Cancer Inst. 2009 Jul 1;101(13):959-63.
18. Wacholder S, Hartge P, Prentice R, Garcia-Closas M, Feigelson HS, Diver WR, Thun MJ, Cox DG, Hankinson SE, Kraft P, Rosner B, Berg CD, Brinton LA, Lissowska J, Sherman ME, Chlebowski R, Kooperberg C, Jackson RD, Buckman DW, Hui P, Pfeiffer R, Jacobs KB, Thomas GD, Hoover RN, Gail MH, Chanock SJ, Hunter DJ. Performance of common genetic variants in breast-cancer risk models. N Engl J Med. 2010 Mar 18;362(11):986-93.
19. Balagani K, Phoha V. On the feature selection criterion based on an approximation of multimensional mutual information. IEEE Trans Pattern Analysis and Machine Intelligence. 2010;32(7):1342-3.
20. Battiti R. Using mutual information for selecting features in supervised neural net learning. IEEE Trans Neural Networks. 1994;5(4):537-50.
21. Ding C, Peng H, editors. Minimum redundancy feature selection from microarray gene expression data. Proc Second IEEE Computational Systems Bioinformatics; 2003.
22. Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Analysis and Machine Intelligence. 2005;27(8):1226-38.
23. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten I. The Weka data mining software: an update. SIGKDD Explorations. 2009;11(1).
24. Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. Machine Learning. 1997;29:131-63.
25. DeLong E, DeLong D, Clarke-Pearson D. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44:837-45.
26. Mullin R. A shaky new age: researchers see a rocky path from genomics research to truly personalized medicines. Chemical & Engineering News. 2014:18-23.
27. Wang X, Zhang L, Chen Z, Ma Y, Zhao Y, Rewuti A, Zhang F, Fu D, Han Y. Association between 5p12 Genomic Markers and Breast Cancer Susceptibility: Evidence from 19 Case-Control Studies. PLoS ONE. 2013;8(9).
28. Gu C, Zhou L, Yu J. Quantitative assessment of 2q35-rs13387042 polymorphism and hormone receptor status with breast cancer risk. PloS One. 2013;8(7).
29. He X, Yao G, Li F, Li M, Yang X. Risk-association of five SNPs in TOX3/LOC643714 with breast cancer in southern China. Int J Molecular Sciences. 2014;15:2130-41.
30. Rutman AM, Kuo MD. Radiogenomics: creating a link between molecular diagnostics and diagnostic imaging. Eur J Radiology. 2009;70:232-41.
31. Yamamoto S, Maki DD, Korn RL, Kuo MD. Radiogenomic analysis of breast cancer using MRI: a preliminary study to define the landscape. AJR Am J Roentgenol. 2012;199:654-63.
32. Kuo MD, Jamshidi N. Behind the numbers: decoding molecular phenotypes with radiogenomics-guiding principles and technical considerations. Radiology. 2014;270:320-5.

Information is in the eye of the beholder: Seeking information on the MMR vaccine through an Internet search engine

Elad Yom-Tov, PhD[1], Luis Fernandez-Luque, PhD Candidate[2, 3]
[1]Microsoft Research, Herzeliya, Israel; [2]UiT - The Arctic University of Norway, Computer Science Department Tromso, Norway, [3] Norut, Tromso, Norway

Abstract

Vaccination campaigns are one of the most important and successful public health programs ever undertaken. People who want to learn about vaccines in order to make an informed decision on whether to vaccinate are faced with a wealth of information on the Internet, both for and against vaccinations. In this paper we develop an automated way to score Internet search queries and web pages as to the likelihood that a person making these queries or reading those pages would decide to vaccinate. We apply this method to data from a major Internet search engine, while people seek information about the Measles, Mumps and Rubella (MMR) vaccine. We show that our method is accurate, and use it to learn about the information acquisition process of people. Our results show that people who are pro-vaccination as well as people who are anti-vaccination seek similar information, but browsing this information has differing effect on their future browsing. These findings demonstrate the need for health authorities to tailor their information according to the current stance of users.

Introduction

Vaccination programs are one of the greatest public health successes in history. Scourges such as smallpox and poliomyelitis have been completely eradicated or are confined to relatively small areas of the globe. In order to achieve the maximal potential of vaccination campaigns, a large percentage of the population needs to be vaccinated, so as to achieve "herd immunity" which significantly reduces the likelihood of epidemic outbreaks. Sadly, in parallel to vaccination programs, there is a worldwide trend towards a hesitancy to vaccinate, catalyzed by the activism of anti-vaccination groups[1].

The presence of an anti-vaccination movement in a country has been found to correlate with lower vaccination rates[2]. Studies have tracked the presence and importance of those movements in the online sphere, especially social media[3]. This is especially pertinent as nowadays the debate about vaccination is prominent on the Internet[4]. Several studies examined online data sets for the tracking and characterizing the online anti-vaccination movement. For example, Salathe et al.[5] studied the sentiment of people on the social network Twitter with regards to influenza vaccination in the USA. The Vaccine Confidence project is tracking online media to index the sureness of people in vaccine across different countries[6]. Together, these studies show that anti-vaccination content is prevalent and may have a detrimental effect on vaccination rates.

People who wish to make an informed decision on whether or not to vaccinate themselves or their children can therefore find information both supportive of vaccination and opposed to it. Understanding the information seeking process of people as they learn about vaccinations is crucial in order to address their concerns and provide them with convincing information on the importance of vaccination. Many factors affect the information seeking process, and include available sources of information, prior beliefs and understanding, and current news. It has been found that the anti-vaccination misconceptions vary across countries and even across different vaccinations[7]. In the USA, a study found that parents refusing to vaccinate their children were doing so based on information obtained by word-of-mouth, for religious or philosophical reasons, because of low perceived risk of diseases and due to anti-government sentiment[8].

McRee et al.[9] found in a study in the USA that nearly 70% of the parents who had questions concerning vaccinating their child against the Human Papilloma Virus (HPV) searched for information on the Internet. Although that study found that trust of healthcare professionals was high, many people supplement information from these professionals with information sought online.

One of the most common activities online is using a search engine to find information[9]. Internet search engine queries have been shown to reflect both the activities of people in the virtual world[10], as well as those in the physical one. For example, Ofran et al.[11] found a high correlation between the number of searches for specific types of cancer and their

prevalence in the population. Similarly, Yom-Tov and Gabrilovich[12] showed a high correlation exists between the number of prescriptions sold for a drug and the number of people who search for it. Because of this, Internet search engines have been used for learning about medicine and health in a variety of areas. These including knowledge acquisition by cancer patients and their family members[11], obesity and its correlated behaviors[13], and the side effects of medical drugs[12].

Internet search data has also been used to learn about public health, sometimes by measuring effects that are difficult to measure in the physical world. As such, prevalence of influenza was estimated by several researchers[14-16]. More recently, the effect of media reporting of celebrities who are suspected to be suffering from eating disorders on the development of anorexia was studied in Yom-Tov and boyd[17].

In this study we focus on the information acquisition process vis-a-vis vaccines, as evident from the searches of people conducted on a major Internet search engine. To the best of our knowledge there are no studies which examined how people utilize Internet search engines to garner information on vaccination.

Previous beliefs predispose people to a pro- or anti-vaccination stance even before they begin their information acquisition process. This predisposition can lead people who seek information on vaccination to selective exposure, that is, the well-documented tendency of people to people seek information which affirms their viewpoint and avoid information which challenges it[18,19]. However, recent work has shown that, at least in the political domain, provided the information is chosen appropriately, it is possible to cause people to overcome selective exposure and read views opposing to theirs[20]. Our goal is to validate whether selective exposure occurs when learning about vaccination, and if so, to be able in future to use similar approaches to those developed for the political domain in order to better inform parents on the importance and necessity of vaccinations.

Our findings show that when people acquire information about the MMR vaccine they do so with preconceived notions that bias their search for information. Selective exposure occurs both in the information that people choose to seek and in the way they process it. This suggests that providers of information on vaccines need to tailor their content according to the predicted stance of the user.

Methods

The methodology in this work comprises of several stages. First, we devised a method for scoring search engine queries to the attitude of the users who made them towards vaccines. We then applied this score to anonymized data from the Bing search engine. These data were aggregated with vaccine uptake information from CDC and analyzed to show both the impact of individual queries and the temporal progression of users as evident from their queries. In addition, we performed a novel validation of our scoring method using an online advertisement campaign.

Search engine query log data

We extracted all queries to the Bing search engine made by users in the USA during 6 months starting March 2013 which included keywords related to the MMR vaccine. The list of keywords included the vaccine name (MMR or MMRV), as well as the trade names of the vaccine: Priorix, Tresivac, and Trimovax. This resulted in 252,526 queries from approximately 115,714 users. Of these, 9,985 users made five or more MMR-related queries. Data on each query comprised of an anonymized user identifier, time, query text, zip code of the user, the pages displayed to the user as a response to the query, and of these, the pages clicked by users. In order to maintain user privacy, data were anonymized before the investigators had access to them. They were then aggregated prior to analysis, and no individual-level user datum was examined. Thus, all the searches of each individual can be identified, but they cannot be attributed to a specific individual.

Our analysis throughout this paper is based on a scoring of queries made by users and the webpages displayed to them in response to these queries. This scoring is intended to reflect the likelihood that a person making these queries or reading these documents will vaccinate their child against MMR. Broadly, we label a person with a low probability to vaccinate as anti-vaccination, and a person with a high probability to vaccinate as pro-vaccination.

Vaccine attitude scores

We score search engine queries as to the likelihood that a person who made them would be receptive towards vaccination. We refer to these scores as the **Vaccine Attitude Score** (VAS), and in this section detail its calculation.

Our scoring is based on CDC data[1] on vaccination rates (percentage of children vaccinated) in US states and specific urban centers. Each query and page (URL) was scored according to the average of the vaccination rate in the areas from which users to whom the pages were displayed came from. Let z_i be the vaccination rate at the zip code from which the *i*-th query was issued. We score each query and each page d_j displayed in response to this query such that:

$$S_{q_i} = \frac{1}{N} \sum_{i=1}^{N} z_i$$

and

$$S_{d_j} = \frac{1}{N} \sum_{i=1}^{N} z_i$$

Clearly, $0 \leq s_{d_j} \leq 100$. Such a scoring method is akin to the methods used in Yom-Tov et al.[20] to elicit the political intent of queries, using the voting patterns in each zip code. We refer to this scoring method as the VAS. A high VAS implies a higher likelihood that a person issuing the query or visiting the web page will vaccinate. We validate VAS and its use for scoring vaccine-related queries and web pages in the Results, through two separate means.

Results

Assessing the accuracy of VAS for queries and pages

A described in the Methods Section, each query and URL were scored according to the average vaccination rates at the area where people who made these queries reside, so as to obtain an estimate of the likelihood that people reading these pages will vaccinate against MMR. This is similar to the method used in Yom-Tov et al.[20] to score political queries. In this section we validate whether this scoring method, which we refer to as VAS, is accurate in the context of vaccine-related queries scored according to the vaccination rates.

Here we employ two methods for validation. First, we use a novel method utilizing online advertising to capture whether queries are posted by people who are pro- or anti-vaccination. Second, we use human assessors to classify web pages according to whether they perceive the information in them as leading to a higher chance of vaccination.

Estimation of queries using online advertisements

We selected the 10 queries with the highest VAS values, computed according to the methods described in the Methods Section, as well as the 10 queries with the lowest VAS values. For each of these, we placed advertisements on the Bing search engine, such that when people queried for one of these 20 queries, one of two advertisements would be shown at the right hand side of the search results page. The advertisements were shown with equal probability, and the Bing ads system showed them when the price we were willing to pay (our bid) for placing these ads was greater than that of any other advertisement placed for these search terms. In some cases additional, competing ads were shown above or below the advertisements we placed.

Both advertisements were similar in their title ("MMR vaccine"), shape, colors, and the link to which they referred. One advertisement, shown in Figure 1, stated in its text "Do you want to learn about the importance of this vaccine?",

[1] http://www.cdc.gov/nchs/nis/data_files_teen.htm

while the other substituted the word "importance" with the word "dangers". In both cases, clicking the advertisement led people to the CDC page on the MMR vaccine.

We hypothesized that the first advertisement would appeal to people who were pro-vaccination, while the second advertisement would appeal more to people who were anti-vaccination. Thus, we expect that ads on the dangers of the vaccine would be clicked more when they appear next to results for queries made by people with an anti-vaccination stance, and that ads on the importance of the vaccine would be clicked more when they appear next to results for queries made by pro-vaccination people.

The advertisements were shown a total of 5476 times over a period of 15 days. Table 1 shows the clickthrough rate (CTR) for each query and advertisement combination. CTR is the percentage of advertisements which were clicked out of all the advertisements shown. As the table shows, advertisements mentioning the importance of the vaccine tended to be clicked 2.55 times more by people who made queries with high VAS, compared to those by people who made queries with low VAS. Similarly, advertisements which mentioned the dangers of the vaccine were 1.19 times more likely to be clicked by people who made queries with low VAS, compared to those by people who made queries with high VAS.

Thus, we conclude that our results support our hypothesis, that pro- (anti-) vaccination people are more likely to click pro- (anti-) vaccination advertisements, and, furthermore, that our method for scoring queries is accurate in that it assigns high VAS values to queries made more by pro-vaccination people, and vice versa.

Estimation of page VAS using human assessment

Three human assessors were tasked with classifying the 20 pages with the highest page VAS values and the 20 pages with the lowest pages VAS values. Each assessor was asked to label each pages according to the question "Does this page lead to a more pro-vaccination stance?". Possible answers were "yes", "no", or "undecided".

The free marginal Kappa[21] for the three annotators was 0.312, indicating a medium-level agreement. Limiting our analysis to 20 pages which had a standard deviation across users smaller than 2%, because of the need to focus on pages which are accessed by a relatively homogenous population, we find that for pages categorized by a majority of labelers as leading to a more pro-vaccination stance the average VAS was only slightly higher at 72.7, compared to 72.1 (not statistically significant). This minor difference can be for one of two main reasons: Either the page VAS given according to vaccination rates is unrepresentative, or else, the page VAS should be evaluated in the context of how it influenced the next page read.

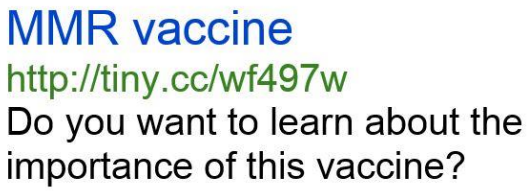


Figure 1. Pro-vaccination ad placed next to Bing searches

	Advertisement	
	Anti	Pro
Low VAS values	0.556	0.468
High VAS values	0.472	1.197

Table 1. Clickthrough rate for pro-vaccination and anti-vaccination advertisements, as a function of query VAS values.

To distinguish between the two, we scored each consecutive pair of pages read by a user, denoted by d_t and d_{t+1} according to the differences in VAS values between them, $s_t = d_{t+1} - d_t$. The average page VAS for those pages labeled as leading to a more pro-vaccination stance was 0.52, compared to a VAS of 0.05 for pages labeled as not leading to a more pro-vaccination stance. Thus, we hypothesize that the latter assumption, that page VAS values should be evaluated according to how they modify a user's reading habits, is the correct one, and that our labeling thus evaluated is a useful way to score pages as to the propensity of users to vaccinate against MMR.

Queries and domains with the highest and lowest VAS

Table 2 shows the list of queries with the highest and lowest VAS in our data. Several interesting observations can be seen from this table. First, both types of queries discuss similar issues, but do so from different viewpoints. For example, the (discredited) link between MMR and autism is addressed as a given in queries with the lowest VAS ("MRR vaccine linked to autism"), but as a possibility in queries with the highest VAS ("link to autism"). Similarly, an Italian court case that suggested that MMR caused autism is searched for in a factual manner in queries with high VAS, whereas people making the queries with a low VAS search for information validating the link ("courts confirm mmr vaccine causes autism"). This suggests that at least some of the people who search for information on vaccines have a preconceived notion of whether or not they intend to vaccinate their children, and are only seeking affirmation of their position. We address this in more detail below.

Among the pages with the lowest VAS, two pages are from CDC, one from Wikipedia, and the remaining are from social media and medical information websites. Pages with the highest VAS include 3 pages from CDC, one from the World Health Organization (WHO), and the remaining are from social media and medical information websites.

In order to obtain a more robust estimation of information sources, we repeated the analysis at the website level, using all websites that had at least 10 scored pages. The score for a website was the average VAS of pages on it. Here, the lowest scored websites included authoritative sources (BMJ, Mayo Clinic, *vaccineinformation.org,* as well as several anti-vaccination websites (*vaccinetruth.org*, *vaccineinjuryhelpcenter.com*), parenting websites (*momtastic.com*), and pharma websites (*merckvaccines.com*). The highest scored websites included parent websites (*mamapedia.com*, *netmums.com*), anti-vaccination websites (*mercola.com*, *vran.org*), government sites (*ny.gov*), and medical websites (*pediatriconcall.com*). Thus, both pro-vaccination and anti-vaccination people read information from similar sources (including government organizations), but, apparently, interpret them differently.

Lowest VAS queries	Highest VAS queries
mmr vaccine linked to autism	mmr vis
side effects of mmr vaccine in adults	when do kids get mmr vaccine
ingredients in mmr vaccine	rash from mmr vaccine
mmr vaccine lot numbers	fever after mmr vaccine
cdc mmr	mmr vaccine administration for adults
mmr vaccine court case	mmr vaccine link to autism
adult mmr vaccine schedule	how to give mmr vaccine
courts confirm mmr vaccine causes autism	mmr vaccine italian court
rash after mmr vaccine photos	mmr vaccine and ppd
mmr vaccination side effects of mmr	vaccine for toddlers

Table 2. Queries with the highest and lowest VAS, according to the vaccination rates in askers geographies.

Modeling the information acquisition

Our goal is to understand what drives a user to read more pro- (or anti-) vaccination information. Therefore, we estimate the contribution of a specific page on the likelihood that a user will read pages with a higher or lower VAS, by modeling the transition of users between pages as a chain of transitions. Let the state of a user at time t_i (i=1,2,...,M) be represented by the VAS of the current page, s_{d_i}, the VAS of vaccine-related pages they read so far, $\sum_{j=1}^{i-1} s_{d_j}$, and the a-priory score of the user, according to the vaccination rate at their locale. The effect of the current page is measured by predicting difference between the VAS of the next page that the user will read and the current page VAS.

The results of a rank regression model, using 14,031 transitions, are shown in Table 3. The table shows that, taken individually, the current page VAS and the average page VAS of past reading have the highest influence on the VAS of the next page. Interestingly, both have a negative effect, that is, the higher the current VAS, the more likely it is to lead to pages with a lower VAS than the current one. This is, possibly, because of interactions between the scores up to time t_i and the score at time t_i. Therefore, model 4 uses all three attributes. Here we find that the R^2 of this model is only slightly greater than that of the model which uses only the current page. However, past reading and user scores contribute to a higher VAS of the next page. The implication of the negative coefficient for page VAS is that many pages are correlated with users reading information associated with a lower vaccination rate. This is an alarming finding, since it suggests that much of the online content on vaccination leads people to read more harmful content.

Since the current page VAS is most strongly correlated with the difference to the next page VAS, we attempt to estimate what in the language of the current page leads to a higher or lower VAS. To do so, we first remove the effect of past reading and user score, by building a rank regression model between these two factors and the difference in next page and current page VAS. Then, we represented each document through its vector space model[22] of words and lexical affinities[23], keeping words that appeared in at least 200 documents, but in no more than a quarter of the documents. We modeled the relationship between the words in the documents and the residual difference using a linear model, constructed using a linear Support Vector Machine (SVM) classifier (with the LibSVM implementation[24], using default settings).

Model number	Variable	Regression coefficient	Model R^2	p-value
1	Page VAS	-0.532	0.283	$<10^{-5}$
2	Past read	-0.387	0.150	$<10^{-5}$
3	User score	0.013	10^{-4}	0.12
4	Page VAS	-0.59	0.297	$<10^{-5}$
	Past read	0.06		$<10^{-5}$
	User score	0.11		$<10^{-5}$

Table 3. Rank regression models of user transitions. Page VAS is the score of the current page read. Past read is the average VAS of all page VASs up to the current time. User score is the vaccination rate at the users' locale.

The 20 words most likely to lead to a higher page VAS are (categorized by the authors, in parenthesis are explanations of the word):

1. Health authorities: guidelines, provider
2. Adverse effect: respiratory, experience (as in "may experience a rash"), avoid (women should avoid becoming pregnant within a month of vaccination), develop
3. Benefits of the vaccination: diphtheria (protection against)
4. Others: answer, lead, review (literature review), personal, scientific, USA, between vaccine, prior, viruses, syndrome, past, started, increased

The 20 words most likely to lead to a lower page VAS are (in descending order of importance, in parenthesis are explanations of the word):

1. Health authorities: control prevention, department, professional, FDA
2. Autism-related: diagnosis (autism diagnosis)
3. Adverse effect: meningitis, swelling, separate (vaccine given in separate parts, rather than all 3 components)
4. Benefits of the vaccination: benefits, pertussis (protection against)
5. Others: issue, activities, problem, means, December, recently, effect, Facebook, put, individual

Thus, both kinds of pages deal with similar topics, but pages leading to lower VAS emphasize adverse effects (and rumored effects) over the more mild symptoms usually experienced.

Differing understanding of similar information

The previous sections suggested that information on the current page and past user reading were most predictive of future reading, and that people who are opposed to vaccination, as well as those who were pro-vaccination read similar information. In this section, we attempt to directly quantify this finding. The null hypothesis is that information on a given page should cause a similar influence on people with differing views, since the information displayed on a page is identical, regardless of a persons' view.

We analyzed the sequence of clicks of each user on pages related to the MMR vaccine, and specifically concentrated on pages that were two or more pages after the beginning of a users' sequence, and at least two pages before it ended. This was done so that the average browsing VAS of a user could be computed both before and after the current click. However, this limited our analysis to 2941 page views, by 979 people who clicked on at least 5 pages.

Users were divided according to whether the average VASs of the pages they clicked on until the current page were above or below the median page VASs. Only 33% of pages were read by people of both groups. This means that people are only 33% likely to read opposing information, and though it is a low percentage, it is higher than that observed for political opinions, which is approximately 20%[20].

Focusing on 108 pages which were read by at least two people from both groups, we measured the fraction of people from each group that read pages with a higher VAS after reading the current page. The Spearman correlation between the fraction of people from each group that read pages with a higher VAS is ρ=-0.004, and not statistically significant. Thus, pages have dissimilar effects on people in the two groups. This means that our null hypothesis is refuted, in that if a page contains information that should lead people to a more positive view of vaccination, the likelihood of reading more positive information differs greatly between people who read pages with lower VASs, compared to those who read higher VASs.

In 78% of pages, the effect was more positive on people who had an average VAS below the median, than on people who had a VAS above it. The association between the page VAS itself and whether it had a bigger effect on one group

over the other is not statistically significant (ranksum test). This implies that, surprisingly, the majority of pages have a more positive effect on people who were previously under the median VAS.

Transitions between VAS quantiles

Our final analysis is concerned with the likelihood of transitioning between page VASs. All page VASs were divided into five quantiles, and the probability of the quantile of the next clicked page given the quantile of the current page was computed. The results are shown in Figure 2. First, as this figure shows, self-loops, indicating that the next page is within the same quantile, are the most common. The average probability of a self-loop is 0.701 (s.d. 0.075). Some of this is due to the coarse division into only five quantiles. However, even when dividing the pages into 20 quantiles, the average probability of a self-loop is 0.607 (s.d. 0.079). Thus, most transitions are to pages with similar VASs.

Figure 2 also shows that transitions are more likely towards quantile 3 than towards quantiles 1 and 5. Indeed, a random walk with restarts[25] (using a random restart probability of λ=0.15) finds that the most likely stable quantile is quantile number 3, with a probability of 0.404, states 2 and 4 have a stationary probability of 0.181 and 0.192, respectively, and stages 1 and 5 have a stationary probability of 0.108 and 0.114, respectively. Therefore, those people who do transition between quantiles are likely to end their search process in pages with VASs in the middle range, indicating that they did not accept extreme views, either towards vaccination or against it.

Discussion

Vaccinations are most effective when the overwhelming majority of the population receives them. Since few countries force citizens to vaccinate, it is important to provide accurate and convincing information to people so as to encourage vaccination. This is a pressing issue especially in the face of anti-vaccination information prevalent on the Internet, which has been linked to a decline in vaccination rates[2].

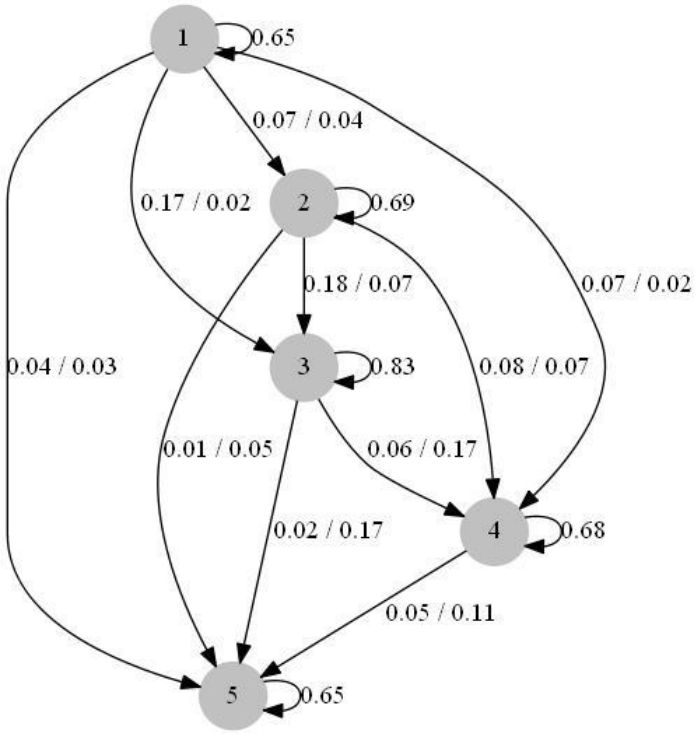


Figure 2. Probability of transition between page VAS quantiles. The left number of each edge represents the probability in the direction of the arrow and the right number the probability in the reverse. Lower quantiles imply lower probability to vaccinate.

In this paper we first developed and validated a way to score queries made to an Internet search engine, and the pages clicked on by users, as to the likelihood that they will be read by people with a pro- or anti-vaccination stance. We used a novel method based on online advertisements to validate our scoring. Our results indicate that our scoring method is useful in identifying such queries, in that people who are pro-vaccination, according to the query VAS value, are significantly more likely to click on advertisements which purport to provide information on the advantages of vaccines, and vice versa. This means that the proposed scoring methods can be used to learn about the information acquisition process, through the observation of sequential page VASs.

When people search for information on the MMR vaccine, their queries reveal their current bias towards vaccines. This represents a problem for pro-vaccination information providers, because their information may not be returned as a result to anti-vaccination queries, and thus pro-vaccination information will not be shown to those people. One possible solution to this problem is to create several types of information, each geared towards people with different attitudes towards vaccination.

Interestingly, though past reading and the information people receive from a web page have statistically significant correlations with future reading, depending on their past reading, people may interpret the same pages differently. Thus, mainstream medical sites feature prominently in the reading of people with an anti-vaccination stance, and the same pages have unequal effects on people of differing stances towards vaccination.

The findings our study have important implications for health authorities, since the Internet is gradually becoming a prominent channel for health education. Firstly, battling the online health misinformation by creating more content is very likely to be suboptimal. The problem it is not the lack of online informative resources about vaccination, but how to design informative resources that are found by those seeking information, whether pro- or anti-vaccination, and providing the most appropriate information for each. Our study provides insights on how to design online information about vaccination. Simple strategies can help, including the usage, within the websites (and their metadata) a vocabulary commonly used by those with anti-vaccination stances. However, further research is required to better understand the reasons for anti-vaccination decisions.

Another approach could be the use of query data for personalizing the health websites. For example, our scores about vaccination stance from search queries can be used to adapt a website about vaccination to provide information which depends on the stance of the visitor. Thus, instead of serving the same content to every person, health providers could offer several versions of their content, each matched to the language and bias of individual users.

Finally, we note that the findings of this study can be further applied to tackle misinformation which has been described by the World Economic Forum as one of the three major threats of the modern hyper-connected society[26].

Acknowledgment

The authors thank Per Egil Kummervold for helpful discussions and assistance in labeling of data.

References

1. Maurice JM, Davey S. State of the world's vaccines and immunization. World Health Organization. 2009.
2. Gangarosa EJ, Galazka A, Wolfe C, Phillips L, Gangarosa R, Miller E, Chen R. Impact of anti-vaccine movements on pertussis control: the untold story. Lancet, 1998;351(9099):356–361.
3. Briones R, Nan X, Madden K, Waks L. When vaccines go viral: an analysis of HPV vaccine coverage on youtube. Health Commun, 2012;27(5):478–485.
4. Betsch C, Sachse K. Dr. Jekyll or Mr. Hyde? (how) the internet influences vaccination decisions: Recent evidence and tentative guidelines for online vaccine communication. Vaccine, 2012;30(25):3723–3726.
5. Salathe M, Khandelwal S. Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. PLoS Comput Biol, 2011;7(10):e1002199.
6. McRee AL, Reiter PL, Brewer NT. Parents internet use for information about HPV vaccine," Vaccine, 2012;30(25):3757–3762.
7. Larson HJ, Smith D, Paterson P, Cumming M, Eckersberger E, Freifeld CC, Ghinai I, Jarrett C, Paushter L, Brownstein JS, et al. Measuring vaccine confidence: analysis of data obtained by a media surveillance system used to analyse public concerns about vaccines. Lancet infect dis, 2013.
8. Fredrickson D, Davis T, Arnould C, Kennen E, Hurniston S, Cross J, Bocchini Jr J. Childhood immunization refusal: provider and parent perceptions. Fam med, 2004;36(6):431.
9. Purcell K. Search engine use. Pew Internet and American Life Project, Tech. Rep., 2012.
10. Goel S, Broder A, Gabrilovich E, Pang B. Anatomy of the long tail: Ordinary people with extraordinary tastes. ACM WSDM, 2010:201–210.
11. Ofran Y, Paltiel O, Pelleg D, Rowe JM, Yom-Tov E. Patterns of information-seeking for cancer on the internet: An analysis of real world data. PloS One, 2012:7(9):e45921.
12. Yom-Tov E, Gabrilovich E. Postmarket drug surveillance without trial costs: Discovery of adverse drug reactions through large-scale analysis of web search queries. JMIR, 2013;16(5): e124.
13. Kuebler M, Yom-Tov E, Pelleg D, Puhl RM, Muennig P. When overweight is the normal weight: An examination of obesity using a social media internet database. PloS One, 2013;8(9):e73479.
14. Eysenbach G. Infodemiology: Tracking Flu-Related Searches on the Web for Syndromic Surveillance. AMIA, 2006: 244–248.
15. Ginsberg J, Mohebbi M, Patel R, Brammer L, Smolinski M, Brilliant L. Detecting influenza epidemics using search engine query data. Nature, 2009; 457(7232): 1012–1014.
16. Polgreen PM, Chen Y, Pennock DM, Nelson FD. Using Internet Searches for Influenza Surveillance. Clin Infect Dis, 2008; 47(14438).
17. Yom-Tov E, boyd d. On the link between media coverage of anorexia and pro-anorexic practices on the web. Int J Eat Disorder, 2014; 47(2): 196-202.
18. Frey D. Recent research on selective exposure to information. Adv exp soc psychol, 1986;19:41–80.
19. Mutz DC, Martin PS. Facilitating communication across lines of political difference: The role of mass media. Am Polit Sci Rev, 2001;95(1):97–114.
20. Yom-Tov E, Dumais S, Guo Q. Promoting civil discourse through search engine diversity. Soc Sci Comput Rev, 2013.
21. Randolph JJ, Thanks A, Bednarik R, Myller N. Freemarginal multirater kappa (multirater free): An alternative to Fleiss fixed-marginal multirater Kappa. Joensuu learning and instruction symposium, 2005.
22. van Rijsbergen C. Information Retrieval (2nd ed.). Butterworth, 1979.
23. Carmel D, Farchi E, Petruschka Y, Soffer A. Automatic query refinement using lexical affinities with maximal information gain. ACM SIGIR 2002:283–290.
24. Chang CC, Lin CJ. LIBSVM: A library for support vector machines. ACM TIST, 2011;2(27):1-27.
25. Meyn SSP, Tweedie RL. Markov chains and stochastic stability. Cambridge University Press, 2009.
26. Howell L. Global Risks 2013, Eighth Edition. World Economic Forum, 2013. Accessed July 2014 at: http://reports.weforum.org/global-risks-2013/

MEDCIS: Multi-Modality Epilepsy Data Capture and Integration System

Guo-Qiang Zhang[1,2], PhD, Licong Cui[1], PhD,
Samden Lhatoo[3], MD, Stephan U. Schuele[4], MD, Satya S. Sahoo[2], PhD
[1]Department of EECS, Case Western Reserve University, Cleveland, OH
[2]Division of Medical Informatics, Case Western Reserve University, Cleveland, OH
[3]Department of Neurology, Case Western Reserve University, Cleveland, OH
[4]Department of Neurology, Northwestern Memorial Hospital, Chicago, IL

Abstract

Sudden Unexpected Death in Epilepsy (SUDEP) is the leading mode of epilepsy-related death and is most common in patients with intractable, frequent, and continuing seizures. A statistically significant cohort of patients for SUDEP study requires meticulous, prospective follow up of a large population that is at an elevated risk, best represented by the Epilepsy Monitoring Unit (EMU) patient population. Multiple EMUs need to collaborate, share data for building a larger cohort of potential SUDEP patient using a state-of-the-art informatics infrastructure. To address the challenges of data integration and data access from multiple EMUs, we developed the Multi-Modality Epilepsy Data Capture and Integration System (MEDCIS) that combines retrospective clinical free text processing using NLP, prospective structured data capture using an ontology-driven interface, interfaces for cohort search and signal visualization, all in a single integrated environment. A dedicated Epilepsy and Seizure Ontology (EpSO) has been used to streamline the user interfaces, enhance its usability, and enable mappings across distributed databases so that federated queries can be executed. MEDCIS contained 936 patient data sets from the EMUs of University Hospitals Case Medical Center (UH CMC) in Cleveland and Northwestern Memorial Hospital (NMH) in Chicago. Patients from UH CMC and NMH were stored in different databases and then federated through MEDCIS using EpSO and our mapping module. More than 77GB of multi-modal signal data were processed using the Cloudwave pipeline and made available for rendering through the web-interface. About 74% of the 40 open clinical questions of interest were answerable accurately using the EpSO-driven VISual AGregagator and Explorer (VISAGE) interface. Questions not directly answerable were either due to their inherent computational complexity, the unavailability of primary information, or the scope of concept that has been formulated in the existing EpSO terminology system.

Introduction

Epilepsy is the most common serious neurological disorder, affecting 65 million persons worldwide; 150,000 new cases of epilepsy are diagnosed in the United States each year [1]. A third of epilepsy patients fail medical treatment and continue to have seizures [2, 3]. Sudden Unexpected Death in Epilepsy (SUDEP) is the leading mode of epilepsy-related death and is most common in patients with intractable, frequent, and continuing seizures [4]. SUDEP is characterized as "sudden, unexpected, witnessed or unwitnessed, non-traumatic and non-drowning death in an individual with epilepsy, with or without evidence for a seizure and excluding documented status epilepticus where postmortem examination does not reveal a cause for death" [5, 6]. Despite an increasing focus on SUDEP research and its inclusion as an NINDS Area III Epilepsy Research Benchmark priority [4, 7], limited progress has been made in characterizing SUDEP risk factors and mechanisms that lead to death. Effective prevention or treatment approaches are unavailable at present [8, 9].

The 2010 Institute of Medicine (IOM) report, "Elements of an Integrated National Strategy to Accelerate Research and Product Development for Rare Diseases," recommends a national strategy that "shares research resources and infrastructure to make good and efficient use of scarce funding, expertise, data, and biological specimens." This recommendation is especially relevant to SUDEP research due to its low rate of reported incidences. For example, the incidence of SUDEP in community-based studies has varied from 0.09 to 0.35 per 100 person-years [10, 11, 12, 13]. A statistically significant cohort of patients for SUDEP study requires meticulous, prospective follow up of a large population that is at an elevated risk, best represented by the Epilepsy Monitoring Unit (EMU) patient population. Hence, multiple EMUs need to collaborate, share data for building a larger cohort of potential SUDEP patient using state-of-the-art informatics and data analytics infrastructure.

In this paper we present the architecture and initial deployment results of MEDCIS, a Multi-Modality Epilepsy Data Capture and Integration System for data integration across multiple EMUs with both retrospective and prospective patient information. MEDICS offers the following collection of main functionalities, each of which has been tested and validated independently:

1. A standardized data entry platform for patient information at different points of care [14];
2. An epilepsy-focused natural language processing (NLP) tool to extract patient information from clinical free text in existing patient records [15, 16];
3. An integrated signal processing application that will allow clinicians to seamlessly interface between signal data and patient information [17, 18, 19]; and
4. A query environment to identify patient cohorts using data integrated from multiple sources based on a shared ontology [20, 21, 22].

1 Background

MEDCIS has been developed as a part of the NINDS-funded Prevention and Risk Identification of SUDEP Mortality (PRISM) project. PRISM has been led by Case Western Reserve University (CWRU) and involves participating EMUs at the University of California Los Angeles (UCLA) Ronald Reagan Medical Center, Northwestern Memorial Hospital at Northwestern University, and the National Hospital for Neurology and Neurosurgery at University College London (UCL). As part of the PRISM project, we have made significant progress in epilepsy informatics, specifically in area of scalable computing for electrophysiological data using cloud-computing tools [17, 18], paradigm-changing applications of terminological systems for epilepsy research including data capture and data visualization [19, 21], and ontology-driven federated approach to large-scale data integration across multiple centers [20, 22]. This section provides a brief overview of the components that have been developed as a part of an overall integrated environment.

1.1 Epilepsy and Seizure Ontology (EpSO)

EpSO [22] models the necessary domain concepts to describe epilepsy phenotype data at significant level of detail by following an established four-dimensional classification framework in epilepsy [23]. EpSO covers concepts of seizures, location of seizures, etiology and related medical conditions according to the four-dimensional scheme. In addition, it models EEG patterns and comprehensive drug information (anti-epileptic, neuroleptic, and anti-depressants) by using the U.S. National Library of Medicine RxNorm standard [24]. EpSO concepts are mapped to the NINDS Common Data Elements (CDE), which represents nine categories of terms describing imaging, neurological exam, neuropsychology, seizures, and syndromes.

1.2 The Ontology-driven Patient Information Capture (OPIC) system

OPIC uses EpSO to implement a flexible web-based interface to capture data describing demography, patient history, details of paroxysmal events, medication, results of prior electrophysiological evaluations, and patient diagnosis. OPIC leverages EpSO to automatically generate multi-level drop menus that are populated with only relevant terms based on previous user selection (skip patterns) and branching logic to model combinations of user selections. Results of patient evaluation in form of EEG, ECG, and other image files can be directly uploaded and attached with clinical reports in OPIC. The OPIC forms are primarily composed the of structured data entry widgets that reduce user-generated errors, support automated consistency checking, and ensure data completeness, using EpSO as the reference terminology system.

1.3 Epilepsy Data Extraction and Annotation (EpiDEA)

EpiDEA [15, 16] is an ontology-driven clinical free text processing system that extends the clinical Text Analysis and Knowledge Extraction System (cTAKES) [25] for analyzing epilepsy-specific clinical reports. EpiDEA processes two types of textual content in clinical notes: the semi-structured sections with attribute-values pairs and the unstructured sections with sentence-based text. An EpSO-driven epilepsy named entity recognition module and a negation detection module processes the output of these modules. EpSO is used in EpiDEA to support three functionalities: term disambiguation, term normalization, and query expansion using subsumption reasoning. For evaluation, EpiDEA has been used to create a database of 500 patients retrospectively with high precision and recall.

1.4 Cloudwave

Electrophysiological signal data, such as EEG, are often used as gold standard in the diagnosis and treatment of epilepsy. However, signal information generated during a patient's admission in an EMU results in very large size multi-modal datasets that cannot be managed using traditional standalone signal processing applications. This is especially important in case of multi-center collaborative clinical studies that require researchers to share and interact with signal data in real time. To address this challenge, we introduced the Cloudwave platform [19] that features a Web-based intuitive signal analysis interface integrated with a Hadoop-based data processing module implemented on clinical data stored in a "private cloud." Cloudwave provides real-time rendering of multi-modal signals with "montages" for signal feature characterization of multi-modal patient data. Cloudwave also supports signal processing [17] with several magnitudes of speed increase over traditional computing environment.

1.5 VISual Aggregator and Explorer (VISAGE)

VISAGE (Visual Aggregator and Explorer) is a query interface for clinical research cohort search [26]. It was developed for Physio-MIMI (Multi-Modality, Multi-Resource Environment for Physiological and Clinical Research), a multi-CTSA-site collaborative project. Physio-MIMI provides an ontology-driven framework for a federated approach to data integration. The interface design features of VISAGE include auto-generated slider bar, selection boxes, and built-in charting. VISAGE also includes administrative and query lifecycle management functionalities, such as role-based access control, auditing, query builder, query manager, and query explorer. VISAGE is an interface framework, which can not only ingest ontologies in the OWL format, but also unify ontology navigation activities with query widget generation [27, 28].

2 Methods

In this section we describe the MEDCIS architecture that integrates the components described in the Background section together into a robust and comprehensive system. The robustness of MEDCIS rests on the centralized usage of the EpSO as the common reference terminology across all of the components (Fig. 1).

The consistent use of EpSO in all MEDCIS components is an important step in achieving the architecture integrity. In the following sections we describe in more detail the method and steps of importing EpSO to VISAGE, and the semi-automated mappings that results in unique and beneficial interface features.

2.1 Importing EpSO

We use the Apache Jena Ontology API for parsing the EpSO OWL file to extract all the classes while preserving the class hierarchy that will be used in the query execution module to support subsumption reasoning. The extracted classes and their hierarchical structure are saved in a comma separated value (CSV) file format and imported into VISAGE using a Ruby language script. The imported classes can then be easily searched and navigated by clinicians, physicians, nurses, trainees, biostatisticians, epidemiologists, oncologists, and investigators from non-clinical disciplines in VISAGE query builder to compose query for cohort identification.

2.2 Data mapping and query

The federated data integration approach in Physio-MIMI requires the use of explicit mappings between the source database and the domain ontology, such as EpSO, for query translation and query execution. For MEDCIS, all the relational database components are mapped to appropriate EpSO ontology concepts manually through consultation between the informaticians and epileptologists to ensure quality of mappings. For example, in the table storing the patients interictal EEG pattern information, the column for the pattern content is mapped to EpSO concept "InterictalPattern," and the column for the pattern location is mapped to the concepts data type property "hasLocation." The mapping also supports the use of ontology concepts for query construction in the VISAGE query builder module.

We use a federated data integration approach in MEDCIS to allow cross-cohort queries over data from multiple centers. Federated data integration is a flexible alternative approach to traditional centralized data warehouse approach that requires periodic Extract Transform Load (ETL) processes to keep data updated in the data warehouse. The federated approach is ideally suited for multi-center studies, allowing tracking and control of data by individual centers while

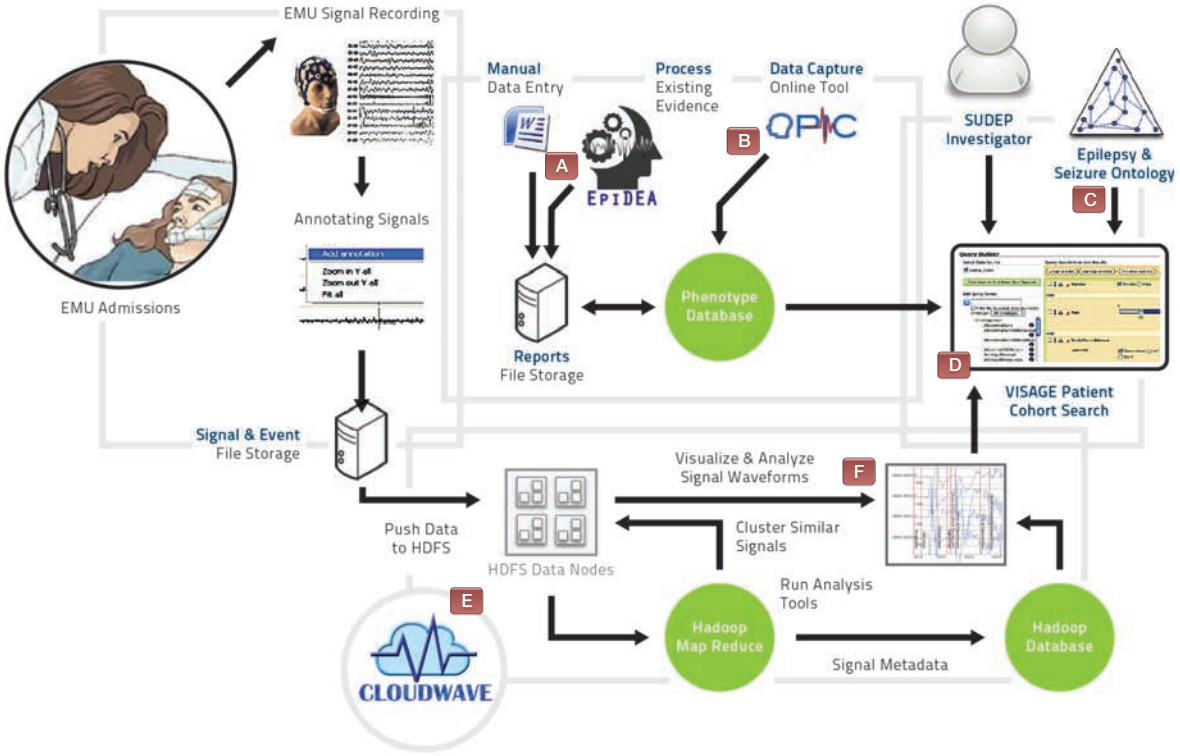


Figure 1: Architecture and data workflow of MEDCIS. A. EpiDEA is used for retrospective information extraction from clinical free-text; B. OPIC is used for prospective structured data capture; C. EpSO is used for mostly other components behind the scenes, but is most directly visible in the VISAGE query interface, which incorporates a built-in ontology browser; D. Once data from EpiDEA and OPIC are ingested into a common database, possibly from multiple sources, the VISAGE interface can be directly accessed by investigators to perform cohort search; E. Multi-modal signal data is processed and annotated using distributed a cloud-computing approach; F. The signal data can be visualized based on the cohort returned from VISAGE, which provides a direct link to the signal data and other clinical documents.

allowing collaborators with appropriate credentials to query the data. Using EpSO as the common terminology schema, MEDCIS allows integrated querying of data from multiple EMUs. The data from each EMU is maintained in separate databases and mapped to EpSO. After mapping, MEDCIS provides seamless access for comprehensive comparative studies of SUDEP and near-SUDEP cases vs. cohort survivors with subjects from participating EMUs.

The key objectives of the VISAGE query interface is to allow clinicians to compose the queries using terms that they are familiar with, combine the terms in different ways (e.g. AND, OR), and explicitly specify negation of specific conditions (e.g. patient not prescribed Keppra). We have already shown in the Physio-MIMI project that VISAGE has multiple advantages over traditional query composition interfaces that often closely reflected the SQL query structure, which is not an intuitive structure for clinical researchers. MEDCIS instantiates the design template of VISAGE using EpSO, thus achieving an interface specifically for the multi-center SUDEP study.

2.2.1 Query widget composition

Each concept in EpSO can generate a visual "query widget" that reflects the specific type of the concept and includes the sub-classes of the concept. For example, a researcher searching for patients with Aura can just select the concept Aura from the "ontology search and browsing" section (Fig. 3), and the appropriate query widget is automatically generated (Fig. 2). The top part of the query widget is the "Aura" class hierarchy as defined in EpSO, and by default includes all its subclasses. In addition to its subclasses, Fig. 2 also shows the ontology property associated with a given class, for example an Aura is associated with the laterality in a patient. The query widget automatically identifies the appropriate values for the laterality information associated with an Aura and displays on the interface for selection by the user.

To explore subclasses of a concept such as "AuditoryAura," a user can click "AuditoryAura" and a new widget is rendered in the query composition interface with specific subclasses of the concept (lower part of Fig. 2). This seamlessly provides the ontological browsing of the concept hierarchy. The user can also easily remove specific query widgets that are no longer needed. This allows users to flexibly modify and update their query structure as they explore the ontology classes and are interested in exploring new hypothesis.

Figure 2: Query widget for the concept "Aura," which automatically generates the display boxes for its subclasses, including "AuditoryAura" (upper). Clicking "AuditoryAura" automatically generates the display boxes for its subclasses as well (lower).

2.2.2 Query generation

The query widgets are translated into the SQL query statement that is executed over the relational database storing the patient data. For each query widget, the EpSO instantiated VISAGE interface records the identifier concept, query concept, data type property of the query concept as well as their selections. It relies on the "database to ontology" mapping to generate the appropriate query. The VISAGE query module does not require each ontology concept to be mapped to a database component. Instead, it leverages the ontology class structure to search for the closest mapped ancestor of the concept that forms the query widget to generate the relational database query. For example, the concept "Aura" is not mapped to any of the data source column, although VISAGE automatically find its closest ancestor "Seizure" which is mapped to the data source column capturing the seizure semiology information.

In the next step, the query generation module identifies the relational database column labeled as "seizure_semiology" found in the table "seizure_semiologies." In addition to the class, the original de-identified discharge summary for the selected patient is also retrieved using the DocumentID column in the "seizure semiologies" table. The property of the class, such as "Laterality" of "AuditoryAura," is mapped to the column laterality in the table "seizure_semiologies." The following sample SQL query statement is created by the query generation module for a query to identify a patient cohort with "Auditory Aura" and having "Right" laterality:

```
SELECT count(DISTINCT seizure_semiologies.`doc`)
FROM seizure_semiologies
   WHERE (CAST(seizure_semiologies.`seizure_semiology` AS CHAR(255))
   IN ('AuditoryAura', 'ComplexAudi-toryAura', 'ElementaryAuditoryAura) and
   CAST(seizure_semiologies.`laterality` AS CHAR(255)) IN ('Right'))
```

An important feature of the query generation module is the ability to support subsumption reasoning based on the EpSO class hierarchy. For example, selecting "Aura" not only includes the eight subtypes as indicated in Fig. 3, but also includes all their descendants. Multiple query widgets can be combined by AND or OR, depending on how the user groups the widgets (Fig. 3).

3 Results

Our prototype implementation involved patients from University Hospitals Case Medical Center (UH CMC) EMU and the EMU at Northwestern Memorial Hospital (NMH) of Northwestern University (under an appropriate IRBs and Data Use Agreement). Multi-modal data of a total of 936 epilepsy patients are currently in MEDCIS. Of the 936 patients, 504 were processed through EpiDEA retrospectively and 432 were captured prospectively using OPIC. All patients processed through EpiDEA were from UH CMC EMU, and 57 captured by OPIC were from NMH. Patients from UH CMC EMU and NMH were stored in different databases and then federated through MEDCIS using EpSO and our mapping module.

For the 504 patients EpiDEA processed a total of 100,836 sentences, 603,605 word tokens and 250,387 noun phrases. 212,246 noun phrases were mapped to appropriate EpSO classes. EpiDEA achieved an overall precision of 93.59%, a recall of 84.01% and an F-measure of 88.53%, as reported in [15]. The MEDCIS database schema consisted of seven tables capturing information about etiology, epileptogenic zone, seizure semiology, lateralizing sign, interictal EEG pattern, ictal EEG pattern, and medication, in addition to demographics.

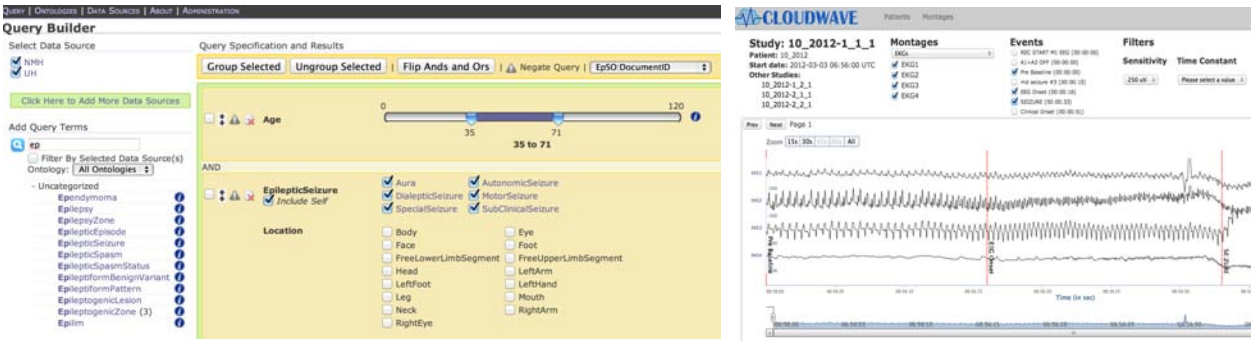


Figure 3: Left: Screenshot of the VISAGE cohort search interface guided by the EpSO ontology. Data from multiple sites are mapped to EpSO, allowing VISAGE to query across projects. The query interface is "driven" by EpSO in that the available seizure types and locations are automatically generated as check boxes for user selection. Right: Screenshot of the Cloudwave web-based signal visualization interface featuring montage composition, events overlay, and a dashboard displaying positioning information which allows the signal to be rendered at multiple desirable resolution.

Multi-modal signal data of a subset of the patients from the UH CMC EMU were linked to Cloudwave through the VISAGE interface. More than 77GB of signal data were processed using the Cloudwave pipeline and made available for rendering through the web-interface.

3.1 Query result rendering

Query results are rendered in format as shown in Fig. 4, listing each patient's gender, age, epileptogenic zone, and EEG pattern as well as location that are clinically relevant.

3.2 Linking to multi-modal data

VISAGE provided a hyperlink (the first column of the table in Fig. 4) to the original discharge summary report and electrophysiological signal data (Fig. 5) of the patient that can be reviewed further by the clinical researcher.

3.3 Evaluation

Each of the MEDCIS components already have their respective evaluations performed and reported as published results [14, 15, 17]. The basic functionality of the VISAGE query environment was evaluated in terms of the usability by clinical investigators as part of the Phsyio-MIMI project [26].

For MEDCIS, our evaluation focused on two integrative aspects. One is expressiveness (Section 3.3.1): the ability

EDF/Discharge Summary	Doc number	Gender	Age	Epileptogenic Zone	EEG Pattern
view	2012_32	Female	26	**LobarEpileptogenicZone**: LeftTemporalLobe, RightTemporalLobe;	**Spike**: RightTemporalLobe, LeftTemporalLobe; **IntermittentSlowActivity**: RightTemporalLobe, LeftTemporalLobe; **SeizurePattern**: LeftTemporalLobe;
	De141	Female	57	**LobarEpileptogenicZone**: RightAnteriorTemporalLobeSegment, TemporalLobe;	**Spike**: RightMesialTemporalLobeSegment, TemporalLobe; **SeizurePattern**: TemporalLobe, RightMesialTemporalLobeSegment;

Figure 4: Sample screenshot of the query result. The result is displayed in a tabular format with data for key fields shown.

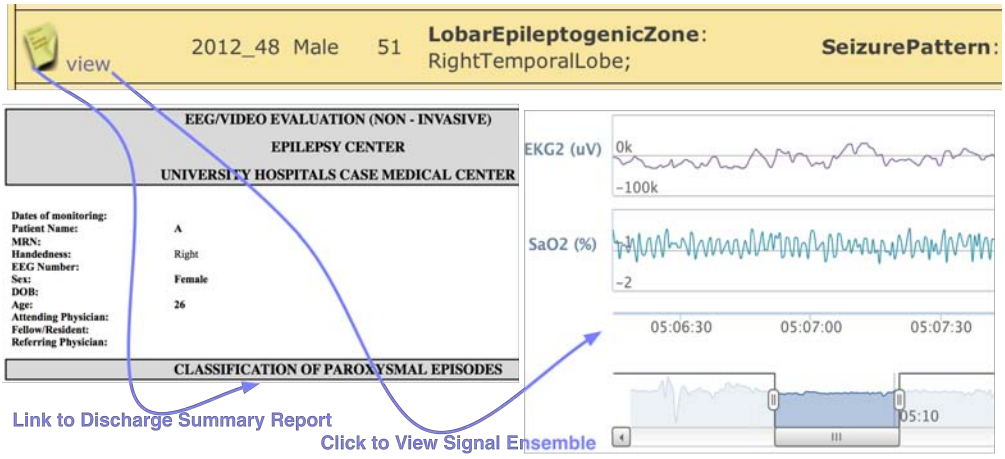


Figure 5: Sample screenshot of the links from query results to discharge summary reports and Cloudwave viewer.

of the query interface to support the type of questions a clinical investigator would like to ask. The other is validity (Section 3.3.2): for those questions that can be translated into appropriate queries, the degree of agreement that the data reported in the VISAGE resulting reports with the information contained in the original patient discharge summaries.

3.3.1 Expressiveness of the EpSO-driven query interface

We focused on the utility of the EpSO domain-specific ontology in supporting the intuitive construction of queries of clinical interest in the PRISM project. 40 questions of clinical interest were formulated by clinical investigators in the UH CMC EMU. Three such questions that were captured using VISAGE are displayed below. Figure 6 (top of next page) shows the screen capture of the query corresponding to the first question.

- Find all patients ages between 20-65 with generalized tonic clonic seizure exhibiting lateralizing sign.
- Find all female patients who have taken Depakote in the past, are currently on Keppra, with either abdominal aura or autonomic aura.
- Find all patients with right visual aura and slow spike EEG pattern from left occipital lobe.

Of the 40 questions, about 26% did not have a direct translation to a query in VISAGE. The non-translatable questions can be classified into three types.

The first type of questions unable to be captured in VISAGE is due to complexity. For example, the question "Find all patients who had been on only two anti-epileptic medications" did not have a simple translation. There are 144 distinct epileptic related drugs. Take a combination of exactly two from this list will result in 10296 choices. This clearly requires a program to execute, and is not an issue that can be easily addressed by a query interface such as VISAGE directly, as far as we know.

The second type of questions unable to be captured in VISAGE is due to limited information capture in the data source. For example, the query "All patients with history of illegal drug use, now or in the past (psychosocial history) and having epileptic paroxysmal episodes (diagnosis)" is not supported because no data on illegal drug use were captured in the EMU.

The third type of questions unable to be captured in VISAGE is due to the incompleteness of the EpSO terminology system. For example, we were not able to capture "All patients with diagnosis of epileptic paroxysmal episodes and periodic limb movement disorder/nocturnal myoclonus" because EpSO has not covered diseases such as "limb movement disorder." This points to the need to continue expanding and evolving EpSO concurrently with the need and scientific advances in the field.

In summary, of the 17 questions not captured by VISAGE, 9 were due to complexity of translation, 5 due to limited source information, and 3 due to terminology incompleteness.

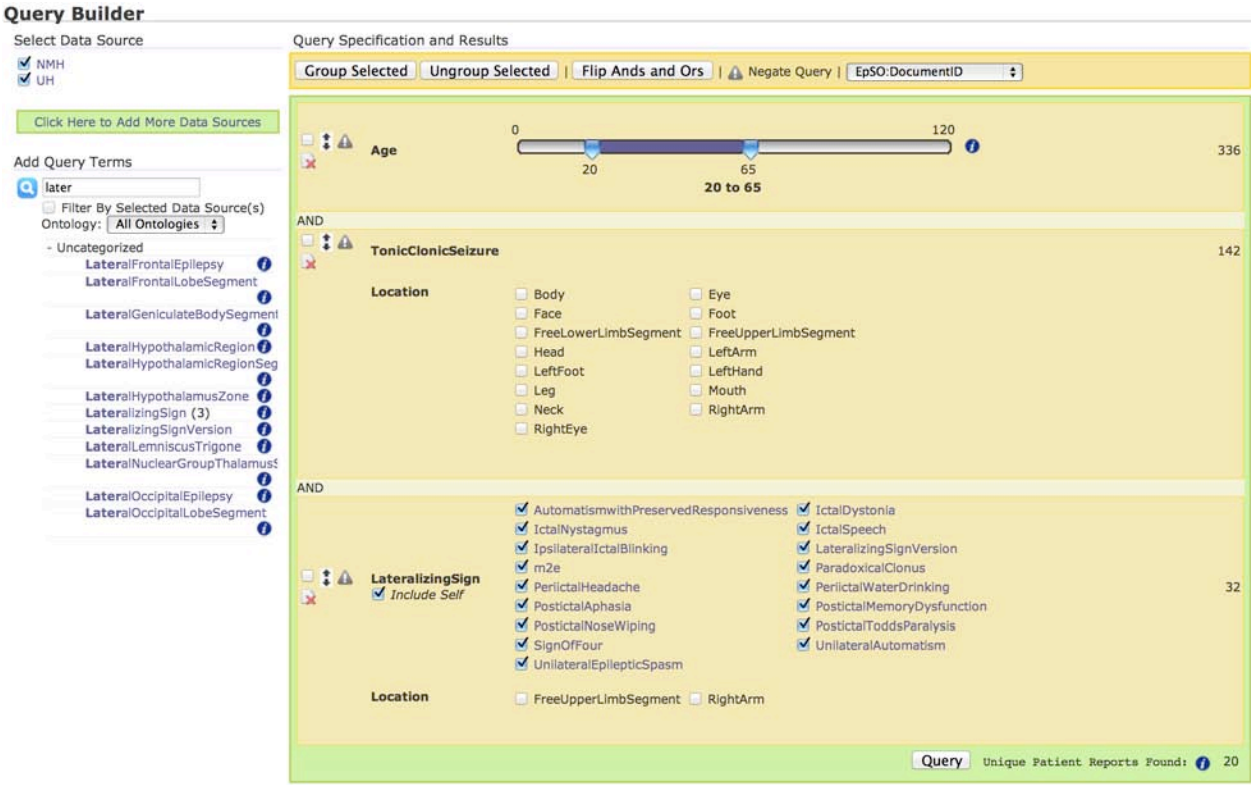


Figure 6: Screenshot of VISAGE query interface corresponding to the criteria "all patients ages between 20-65 with generalized tonic clonic seizure exhibiting lateralizing sign."

3.3.2 Validity of query results against the ground truth

For the 74% of the remaining questions captured using VISAGE queries, we manually validated their correctness by using the linked discharge summery reports as primary source, and inspected the results to be correct. However, since EpiDEA, as any other NLP tool, cannot possibly reach a 100% precision and a 100% recall, its actual precision of 93.59% and recall of 84.01% will impact the query results compared to ground truth (i.e., information found in discharge summaries). We expect OPIC captured patient information to achieve near 100% precision and near 100% recall, barring human data entry errors (although we have not independently evaluated this aspect of data entry quality).

4 Discussions

The functionality of MEDCIS, in terms of query composition, is only limited by the number of concepts modeled in EpSO. At present EpSO has not covered all existing epilepsy subspecialties, such as pediatric epilepsy. As the PRISM project continues, we are engaging the epilepsy community to expand and enhance the concepts modeled in EpSO to address this limitation. We also propose to conduct a comprehensive user survey, spanning six months, with the new batch of medical residents in the UH CMC EMU for further analysis of MEDCIS features. This will help us to maintain and update MEDCIS with new features according to changing user requirements in the PRISM project.

A larger EMU consortium has been formed, involving additional EMUs. Multi-modal physiological, biochemical, phenotypic, genetic and imaging data are to be collected for a targeted number of 2500 epilepsy patients for the next five years. MEDCIS has been adopted as the informatics and data infrastructure hosting this unique and largest nationally shared SUDEP research resource.

The reported SUDEP incidence in adult EMUs is about 5 per 1000 per year [29]. Since SUDEP is a rare event, we had no SUDEP incidence at UH CMC EMU while the patients were under active mornitoring during the PRISM project. Two SUDEP deaths among the UH CMC EMU patients occurred outside the EMU. In the next five years, we expect 10 to 20 SUDEP cases in UH CMC EMU, accounting for patient population growth. Prospectively capturing

all available EMU patient information, and sharing of such information from multiple EMUs, is critical for advancing the understanding the mechanism of SUDEP.

5 Conclusion

This paper presented key architecture and interface features of the MEDCIS platform to address the challenges of integrating structured and semi-structured information from multiple EMUs for SUDEP research. Our experience shows that the ontology-driven architecture for MEDCIS coupling the EpiDEA clinical free text processing tool, the OPIC structured data capturing interface, and the VISAGE query interface, provides a scalable solution to the data extraction and querying requirements in multi-center clinical research projects exemplified by the PRISM study.

Acknowledgement. This research was supported by the PRISM (Prevention and Risk Identification of SUDEP Mortality) Project (1-P20-NS076965-01) and in part by the Case Western Reserve University CTSA Grant NIH/NCATS UL1TR000439.

References

[1] Epilepsy Foundation. Available from: http://www.epilepsyfoundation.org/aboutepilepsy/whatisepilepsy/statistics.cfm. Accessed August 2nd, 2014.

[2] Boon P, Vonck K, De Herdt V, Van Dycke A, Goethals M, Goossens L, Van Zandijcke M, De Smedt T, Dewaele I, Achten R, Wadman W, Dewaele F, Caemaert J, Van Roost D. Deep brain stimulation in patients with refractory temporal lobe epilepsy. Epilepsia. 2007;48(8):1551-60.

[3] Fisher RS. Emerging antiepileptic drugs. Neurology. 1993;43(suppl):12-20.

[4] Tomson T, Nashef L, Ryvlin P. Sudden unexpected death in epilepsy: current knowledge and future directions. Lancet neurology. 2008;7(11):1021-31.

[5] Nashef L, So EL, Ryvlin P, Tomson T. Unifying the definitions of sudden unexpected death in epilepsy. Epilepsia. 2012;53(2):227-33.

[6] Nashef L. Sudden unexpected death in epilepsy: terminology and definitions. Epilepsia. 1997;38(Suppl 11):S6-8.

[7] Kelley MS, Jacobs MP, Lowenstein DH. The NINDS epilepsy research benchmarks. Epilepsia. 2009;50(3):579-82.

[8] Tomson T, Walczak T, Sillanpaa M, Sander JW. Sudden unexpected death in epilepsy: a review of incidence and risk factors. Epilepsia. 2005;46 (Suppl 11):54-61.

[9] So EL. What is known about the mechanisms underlying SUDEP? Epilepsia. 2008;49 (Suppl 9):93-8.

[10] Tomson T, Nashef L, Ryvlin P. Sudden unexpected death in epilepsy: current knowledge and future directions. Lancet neurology. 2008;7(11):1021-31. Epub 2008/09/23.

[11] Tomson T, Walczak T, Sillanpaa M, Sander JW. Sudden unexpected death in epilepsy: a review of incidence and risk factors. Epilepsia. 2005;46 (Suppl 11):54-61.

[12] Lhatoo SD, Sander JWAS. The epidemiology of epilepsy and learning disability. Epilepsia. 2001;42(s1):6-9.

[13] Ficker DM, So EL, Shen WK, Annegers JF, O'Brien PC, Cascino GD, Belau PG. Population-based study of the incidence of sudden unexplained death in epilepsy. Neurology. 1998;51(5):1270-4.

[14] Sahoo SS, Zhao M, Luo L, Bozorgi A, Gupta A, Lhatoo SD, Zhang GQ. OPIC: Ontology-driven patient information capturing system for epilepsy. The American Medical Informatics Association (AMIA) Annual Symposium, 2012;2012:799-808.

[15] Cui L, Bozorgi A, Lhatoo SD, Zhang GQ, Sahoo SS. EpiDEA: Extracting structured epilepsy and seizure information from patient discharge summaries for cohort identification. The American Medical Informatics Association (AMIA) Annual Symposium, 2012;2012:1191-1200.

[16] Cui L, Sahoo SS, Lhatoo SD, Garg G, Rai P, Bozorgi A, Zhang GQ. Complex epilepsy phenotype extraction from narrative clinical discharge summaries. Journal of Biomedical Informatics, Published Online: June 25, 2014, doi: http://dx.doi.org/10.1016/j.jbi.2014.06.006.

[17] Sahoo S, Jayapandian C, Garg G, Kaffashi F, Chung S, Bozorgi A, Chen CH, Loparo K, Lhatoo SD, and Zhang GQ. Heart beats in the cloud: distributed analysis of electrophysiological "big data" using cloud computing for epilepsy clinical research. J Am Med Inform Assoc. 2014 Mar 1;21(2):263-71.

[18] Jayapandian CP, Chen CH, Bozorgi A, Lhatoo SD, Zhang GQ, Sahoo SS. Cloudwave: distributed processing of "Big Data" from electrophysiological recordings for epilepsy clinical research using Hadoop. The American Medical Informatics Association (AMIA) Annual Symposium, 2013 Nov 16;2013:691-700.

[19] Jayapandian CP, Chen CH, Bozorgi A, Lhatoo SD, Zhang GQ, Sahoo SS. Electrophysiological signal analysis and visualization using Cloudwave for epilepsy clinical research. Stud Health Technol Inform. 2013;192:817-21.

[20] Zhang GQ, Sahoo SS, Lhatoo SD. From classification to epilepsy ontology and informatics. Epilepsia 2012: 53 (2 Suppl): 28-32.

[21] Sahoo SS, Zhang GQ, Lhatoo SD. Epilepsy informatics and an ontology-driven infrastructure for large database research and patient care in epilepsy. Epilepsia. 2013;54(8):1335-41.

[22] Sahoo SS, Lhatoo SD, Gupta DK, Cui L, Zhao M, Jayapandian C, Bozorgi A, Zhang GQ. Epilepsy and seizure ontology: towards an epilepsy informatics infrastructure for clinical research and patient care. Journal of American Medical Association. 2014;21(1):82-9.

[23] Lüders HO, Amina, S., et al. Modern technology calls for a modern approach to classification of epileptic seizures and the epilepsies. Epilepsia. 2012;53(3):405-11.

[24] Bodenreider O, Peters L, Nguyen T. RxNav: Browser and application programming interfaces for drug information sources. AMIA Annual Symposium, 2011:2129 (demo).

[25] Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17(5):507-513.

[26] Zhang GQ, Siegler T, Saxman P, Sandberg N, Mueller R, Johnson N, Hunscher D, and Arabandi S. VISAGE: a query interface for clinical research. AMIA Jt Summits Transl Sci Proc. 2010 Mar 1;2010:76-80.

[27] Zhang GQ, Cui L, Teagno J, Kaebler D, Koroukian S, Xu R. Merging ontology navigation with query construction for web-based medicare data exploration. AMIA Jt Summits Transl Sci Proc. 2013 Mar 18;2013:285-9.

[28] Cui L, Mueller R, Sahoo SS, Zhang GQ. Querying complex federated clinical data using ontological mapping and subsumption reasoning. IEEE International Conference on Healthcare Informatics 2013 (ICHI 2013), pp. 351-360.

[29] Ryvlin P, Nashef L, Lhatoo SD, Bateman LM, Bird J, Bleasel A, et al. Incidence and mechanisms of cardiorespiratory arrests in epilepsy monitoring units (MORTEMUS): a retrospective study. The Lancet Neurology, 2013;12(10):966-977.

Towards Drug Repositioning: A Unified Computational Framework for Integrating Multiple Aspects of Drug Similarity and Disease Similarity

Ping Zhang, PhD, Fei Wang, PhD, Jianying Hu, PhD
Healthcare Analytics Research, IBM T.J. Watson Research Center, New York, USA

Abstract

In response to the high cost and high risk associated with traditional de novo drug discovery, investigation of potential additional uses for existing drugs, also known as drug repositioning, has attracted increasing attention from both the pharmaceutical industry and the research community. In this paper, we propose a unified computational framework, called DDR, to predict novel drug-disease associations. DDR formulates the task of hypothesis generation for drug repositioning as a constrained nonlinear optimization problem. It utilizes multiple drug similarity networks, multiple disease similarity networks, and known drug-disease associations to explore potential new associations among drugs and diseases with no known links. A large-scale study was conducted using 799 drugs against 719 diseases. Experimental results demonstrated the effectiveness of the approach. In addition, DDR ranked drug and disease information sources based on their contributions to the prediction, thus paving the way for prioritizing multiple data sources and building more reliable drug repositioning models. Particularly, some of our novel predictions of drug-disease associations were supported by clinical trials databases, showing that DDR could serve as a useful tool in drug discovery to efficiently identify potential novel uses for existing drugs.

Introduction

The inefficiency of pharmaceutical drug development with high expenditure but low productivity has been widely discussed[1, 2]. Drug repositioning, the process of finding additional indications (i.e., diseases) for existing drugs, presents a promising avenue for identifying better and safer treatments without the full cost or time required for *de novo* drug development. Candidates for repositioning are usually either market drugs or drugs that have been discontinued in clinical trials for reasons other than safety concerns. Because the safety profiles of these drugs are known, clinical trials for alternative indications are cheaper, potentially faster and carry less risk than *de novo* drug development. Any newly identified indications can be quickly evaluated from phase II clinical trials. Drug repositioning can reduce drug discovery and development time from 10-17 years to potentially 3-12 years[3]. Therefore, it is not surprising that in recent years, new indications, new formulations, and new combinations of previously marketed products accounted for more than 30% of the new medicines that reach their first markets[4]. Drug repositioning has drawn widespread attention from the pharmaceutical industry, government agencies, and academic institutes. However, current successes in drug repositioning have primarily been the result of serendipitous events based on *ad hoc* clinical observation, unfocused screening, and "happy accidents". Comprehensive and rational approaches are urgently needed to explore repositioning opportunities.

A reasonable systematic method for drug repositioning is the application of phenotypic screens by testing compounds with biomedical and cellular assays. However, this method also requires additional wet bench work of developing appropriate screening assays for each disease being investigated, and it thus remains challenging in terms of cost and efficiency. Big data analytics for both drugs and diseases provide an unprecedented opportunity to uncover novel statistical associations between drugs and diseases in a scalable manner. Many computational methods have been developed in this direction, including: (1) matching drug indications by their disease-specific response profiles based on the Connectivity Map (CMap) data[5, 6]; (2) predicting novel associations between drugs and diseases by the "Guilt by Association" (GBA) approach[7]; (3) utilizing structural features of compounds/proteins to predict new targets or indications, such as molecular docking[8, 9], and quantitative structure-activity relationship (QSAR) modelling[10]; (4) identifying associations between drugs and diseases in genetic activities, such as genome-wide association study (GWAS)[11], pathway profiles[12], and transcriptional responses[13]; (5) constructing drug network and using network neighbors to infer novel drug uses based on phenotypic profiles, such as side effects[14-16], and gene expression[17, 18]. All of these methods only focus on different aspects of drug/disease activities and therefore result in biases in their predictions. Also, these methods suffer from the noise in the given information source. Recently, several integrative methods which combine chemical, genetic, or phenotypic features were proposed to predict drug indications, for example, PREDICT[19], SLAMS[20], PreDR[21], Li and Lu[22], Huang *et al*[23], and Napolitano *et al*[24].

In this paper, we propose a unified computational framework for drug repositioning hypothesis generation, by integrating multiple **D**rug information sources and multiple **D**isease information sources to facilitate drug **R**epositioning tasks (DDR). DDR utilizes drug similarity network, disease similarity network, and known drug-disease associations to explore the potential associations among other unlinked drugs and diseases. In the experiment, we investigate three types of drug information (i.e., chemical structure, target protein, and side effect) and three types of disease information (i.e., phenotype, ontology, and disease gene). The proposed framework is also extensible, and thus DDR can incorporate additional types of drug/disease information sources.

Compared to prior integrative drug repositioning methods, it is worthwhile to highlight the following novel aspects that DDR can achieve simultaneously: (1) DDR can predict additional drug-disease associations by considering both drug information and disease information. With the exception of PREDICT[19], which integrates drug similarity scores and disease similarity scores using unweighted geometric mean, other integrative methods only consider either some drug information sources or some disease information sources. (2) DDR can determine interpretable importance of different information sources during the prediction. To our knowledge ours is the first study to do so. (3) As by-products, DDR can also discover the drug and disease groups, such that the drugs or diseases within the same group are highly correlated with each other, thus providing additional insights for targeted downstream investigations including clinical trials.

Construction of Drug Similarity and Disease Similarity Measures

In this section we introduce drug/disease similarities to quantify the degree of sharing common characteristics between pairs of drugs/diseases. A drug/disease similarity provided for a pair of drugs/diseases is a score that ranges from 0 to 1, with 0 representing the lowest similarity and 1 standing for the highest similarity. For each drug pair, we calculated three types of similarities based on chemical structures, target proteins, and side effects. For each disease pair, we calculated three types of similarities based on disease phenotypes, disease ontology, and disease genes.

Drug Similarity of Chemical Structures $\mathbf{D^{chem}}$. It is generally believed that drugs with similar chemical structures would carry out common therapeutic function, thus likely treat common diseases. We calculated the first drug pairwise similarity based on a chemical structure fingerprint corresponding to the 881 chemical substructures[25] defined in PubChem database[26]. Each drug *d* was represented by an 881-dimensional binary profile *h(d)* whose elements encode for the presence or absence of each PubChem substructure by 1 or 0, respectively. Then the pairwise chemical similarity between two drugs *d* and *d'* is computed as the Tanimoto coefficient of their chemical fingerprints:

$$D^{chem}_{d,d'} = \frac{h(d) \bullet h(d')}{|h(d)| + |h(d')| - h(d) \bullet h(d')} \quad (1)$$

where $|h(d)|$ and $|h(d')|$ are the counts of substructure fragments in drugs *d* and *d'* respectively. The dot product $h(d) \bullet h(d')$ represents the number of substructure fragments shared by two drugs.

Drug Similarity of Target Proteins $\mathbf{D^{target}}$. A drug target is the protein in the human body whose activity is modified by a drug resulting in a desirable therapeutic effect. Drugs sharing common targets often possess similar therapeutic function. We collected all target proteins for each drug from DrugBank[27]. Then we calculated the pairwise drug target similarity between drugs *d* and *d'* based on the average of sequence similarities of their target protein sets:

$$D^{target}_{d,d'} = \frac{1}{|P(d)||P(d')|} \sum_{i=1}^{|P(d)|} \sum_{j=1}^{|P(d')|} SW(P_i(d), P_j(d')) \quad (2)$$

where given a drug *d*, we presented its target protein set as *P(d)*; then $|P(d)|$ is the size of the target protein set of drug *d*. The sequence similarity function of two proteins *SW* was calculated as a Smith-Waterman sequence alignment score[28].

Drug Similarity of Side Effects $\mathbf{D^{se}}$. Drug side effects, or adverse drug reactions, indicate the malfunction by off-targets. Thus side effects are useful to infer whether two drugs share similar target proteins and treat similar diseases. We obtained side effect keywords from SIDER[29], an online database containing drug side effect information extracted from package inserts using text mining methods. Each drug *d* was represented by 4192-dimensional binary side effect profile *e(d)* whose elements encode for the presence or absence of each of the side

effect key words by 1 or 0 respectively. Then the pairwise side effect similarity between two drugs *d* and *d'* is computed as the Tanimoto coefficient of their side effect profiles:

$$D^{se}_{d,d'} = \frac{e(d) \bullet e(d')}{|e(d)| + |e(d')| - e(d) \bullet e(d')} \tag{3}$$

where $|e(d)|$ and $|e(d')|$ are the counts of side effect keywords for drugs *d* and *d'* respectively. The dot product $e(d) \bullet e(d')$ represents the number of side effects shared by two drugs.

Disease Similarity of Phenotypes S^{pheno}. Disease phenotypes indicate phenotypic abnormalities encountered in human diseases. We used the phenotypic similarity constructed by van Driel *et al*[30]. The disease phenotypic similarity was constructed by identifying similarity between the MeSH terms[31] appearing in the medical description ("full text" and "clinical synopsis" fields) of diseases from OMIM database[32]. To be specific, each disease *s* in OMIM was represented by *K*-dimensional (*K* is the number of the MeSH terms) MeSH term feature vector *m(s)*: each entry in the feature vector represents an MeSH term, and the counts of the term found for disease *s* are the corresponding feature value. Then the pairwise disease phenotype similarity between two diseases *s* and *s'* is computed as the cosine of the angle between their feature vectors:

$$S^{pheno}_{ss'} = \frac{\sum_{i=1}^{K} m(s)_i m(s')_i}{\sqrt{\sum_{i=1}^{K} m^2(s)_i} \sqrt{\sum_{i=1}^{K} m^2(s')_i}} \tag{4}$$

where $m(s)_i$ denotes the *i*-th entry of the feature vector *m(s)*.

Disease Similarity of Disease Ontology S^{do}. The Disease Ontology (DO)[33] is an open source ontological description of human disease, organized from a clinical perspective of disease etiology and location. The terms in DO are disease names or disease-related concepts and are organized in a directed acyclic graph (DAG). Two linked diseases in DO are in an "is-a" relationship, which means one disease is a subtype of the other linked disease. And the lower a disease is in the DO hierarchy, the more specific the disease term is. We calculated the semantic similarity between any pair of the diseases using the tool DOSim[34]. For a disease term *s* in DO, the probability that the term is used in disease annotations is estimated as p_s, which is the number of disease term *s* or its descendants in DO divided by the total number of disease terms in DO. Then the semantic similarity of two diseases *s* and *s'* is defined as the information content of their lowest common ancestor by:

$$S^{do}_{ss'} = -\log \min_{x \in C(s,s')} p_x \tag{5}$$

where *C(s,s')* is the set of all common ancestors of diseases *s* and *s'*.

Disease Similarity of Disease Genes S^{gene}. Disease-causing aberrations in the normal function of a gene define that gene as a disease gene. We collected all disease genes for each disease from "phenotype-gene relationships" field from OMIM database. Then we calculated the pairwise disease similarity between diseases *s* and *s'* based on the average of sequence similarities of their disease gene sets:

$$S^{gene}_{ss'} = \frac{1}{|G(s)||G(s')|} \sum_{i=1}^{|G(s)|} \sum_{j=1}^{|G(s')|} SW(G_i(s), G_j(s')) \tag{6}$$

where given a disease *s*, we presented its disease gene set as *G(s)*; then $|G(s)|$ is the size of the disease gene set of disease *s*. The sequence similarity function of two disease genes *SW* was calculated as a Smith-Waterman sequence alignment score.

Methodology

In this section we present the details of the proposed DDR approach. Suppose we have *n* information sources to measure drug similarity and *m* information sources to measure disease similarity. Let $\boldsymbol{D}_k \in \mathbb{R}^{n \times n}$ be the drug similarity matrix measured on the *k*-th information source, and suppose there are in total K_d information sources to measure the drug similarities. Similarly, let $\boldsymbol{S}_l \in \mathbb{R}^{m \times m}$ be the disease similarity matrix measured on the *l*-th information source and suppose there are K_s sources to measure the disease similarities. Let $\boldsymbol{U} \in \mathbb{R}^{n \times C_D}$ be the latent drug grouping matrix with C_D the number of drug groups, and $\boldsymbol{U}_{ij}$ indicates the possibility that the *i*-th drug belonging to the *j*-th drug cluster. $\boldsymbol{V} \in \mathbb{R}^{m \times C_S}$ be the latent disease grouping matrix with C_S the number of disease groups, and $\boldsymbol{V}_{ij}$

indicating the possibility that the *i*-th disease belonging to the *j*-th disease cluster. $\boldsymbol{R} \in \mathbb{R}^{n \times m}$ be the observed (i.e., known) drug-disease association matrix with $\boldsymbol{R}_{ij}=1$ if the association between the *i*-th drug and *j*-th disease is observed, and $\boldsymbol{R}_{ij}=0$ otherwise. Then we aim to analyze the drug-disease network by minimizing the following objective:

$$J = J_0 + \lambda_1 J_1 + \lambda_2 J_2 \tag{7}$$

where the three parts in the objective are:

- The reconstruction loss of observed drug-disease associations:

$$J_0 = \| \Theta - U \Lambda V^T \|_F^2 \tag{8}$$

  Here $\boldsymbol{\Theta} \in \mathbb{R}^{n \times m}$ is the estimated dense version of $\boldsymbol{R}$, and $\boldsymbol{\Lambda} \in \mathbb{R}^{C_D \times C_S}$ encodes the relationship between drug clusters and disease clusters.

- The reconstruction loss of drug similarities:

$$J_1 = \sum_{k=1}^{K_d} \omega_k \| D_k - UU^T \|_F^2 + \delta_1 \| \omega \|_2^2 \tag{9}$$

  Here the estimated drug similarity matrix is $\boldsymbol{UU}^T$, and $\omega \in \mathbb{R}^{K_d \times 1}$ is the nonnegative weight vector when aggregating the reconstruction loss on different drug information sources. The L_2 norm regularization is added to avoid trivial solution[35] and $\delta_1 \geq 0$ is the tradeoff parameter.

- The reconstruction loss of disease similarities:

$$J_2 = \sum_{l=1}^{K_s} \pi_l \| S_l - VV^T \|_F^2 + \delta_2 \| \pi \|_2^2 \tag{10}$$

  Here the estimated disease similarity matrix is $\boldsymbol{VV}^T$, and $\pi \in \mathbb{R}^{K_s \times 1}$ is the nonnegative weight vector when aggregating the reconstruction loss on different disease information sources. The L_2 norm regularization is added for the same reasons in equation (9).

Putting everything together, we obtained the optimization problem to be resolved:

$$\min_{\mathbf{U},\mathbf{V},\boldsymbol{\Lambda},\boldsymbol{\Theta},\omega,\pi} J \tag{11}$$

$$\text{subject to } \boldsymbol{U} \geq 0,\ \boldsymbol{V} \geq 0,\ \boldsymbol{\Lambda} \geq 0,\ \omega \geq 0,\ \omega^T \mathbf{1}=1,\ \pi^T \mathbf{1}=1,\ P_\Omega(\boldsymbol{\Theta}) = P_\Omega(\boldsymbol{R})$$

where Ω is the set of indices of the observed associations, and P_Ω is the projection operator on obtaining the entries of a matrix indexed by the indices in Ω. Thus the constraint $P_\Omega(\boldsymbol{\Theta}) = P_\Omega(\boldsymbol{R})$ restricts the estimated drug-disease associations should include the ones that are already observed. Note that to enhance the interpretability of the learned model, we require $\boldsymbol{U}$, $\boldsymbol{V}$, and $\boldsymbol{\Lambda}$ to be nonnegative, ω and π to be in simplexes. As there are lots of symbols and notations involved in problem (11), we summarize them in Table 1. To further help understanding those symbols as well as their roles in problem (11), we also provide a graphical illustration of the main idea of DDR in Figure 1.

Table 1. Notations and symbols of the methodology

Notation	Size	Meaning
$\boldsymbol{D}_k$	n×n	The *k*-th drug similarity matrix
$\boldsymbol{S}_l$	m×m	The *l*-th disease similarity matrix
$\boldsymbol{U}$	$n \times C_D$	Drug cluster assignment matrix
$\boldsymbol{V}$	$m \times C_S$	Disease cluster assignment matrix
$\boldsymbol{\Lambda}$	$C_D \times C_S$	Drug-disease cluster relationship matrix
$\boldsymbol{R}$	n×m	Observed drug-disease association matrix
$\boldsymbol{\Theta}$	n×m	Densified estimation of $\boldsymbol{R}$
ω	$K_d \times 1$	Drug similarity weight vector
π	$K_s \times 1$	Disease similarity weight vector

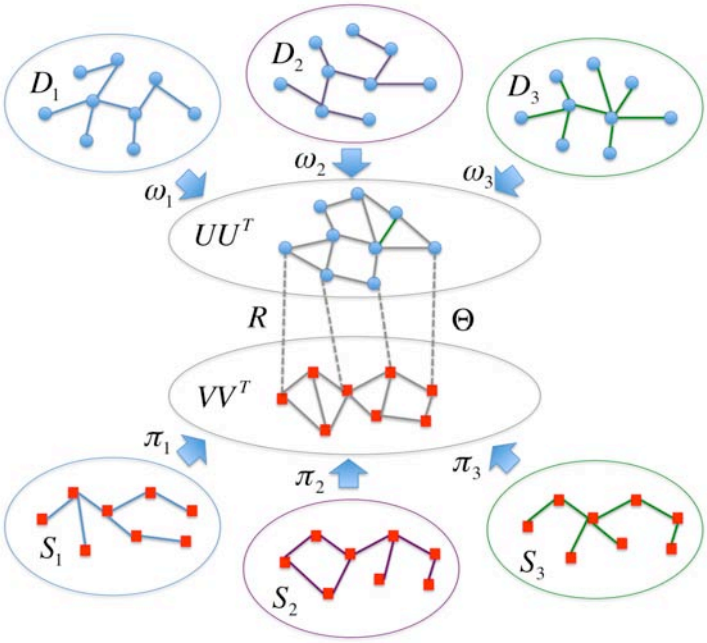


Figure 1. A graphical illustration of the main idea of DDR. There are multiple information sources that we can utilize to construct drug/disease similarities, and the constructed drug/disease similarity matrices are denoted by $\{\boldsymbol{D}_k\}_{k=1}^{K_d}$ or $\{\boldsymbol{S}_l\}_{l=1}^{K_s}$. There is also an observed drug-disease association matrix $\boldsymbol{R}$. Then DDR can learn the drug/disease grouping matrix $\boldsymbol{U}$ or $\boldsymbol{V}$, the estimated drug-disease association matrix $\boldsymbol{\Theta}$ and the importance of different drug/disease information sources ω or π.

Our proposed DDR method integrates multiple drug similarities, multiple disease similarities, and known drug-disease associations to achieve a global estimation on the entire drug-disease network including the intrinsic drug similarity, intrinsic disease similarity, as well as drug-disease associations. DDR formulates such a network estimation problem as a constrained nonlinear optimization problem. Since there are multiple groups of variables involved in the optimization problem (11), we adopt an efficient solution based on the Block Coordinate Descent (BCD) strategy[36]. The BCD approach works by solving the different groups of variables alternatively until convergence. At each iteration, it solves the optimization problem with respect to one group of variables with all other groups of variables fixed. Due to lack of space, details of the BCD solution procedure and its complexity analysis is provided at http://astro.temple.edu/~tua87106/ddr_bcd.pdf.

Results and Discussion

In this section we present experimental evaluation results of the proposed DDR algorithm on a drug repositioning task.

Data Description. The benchmark dataset, which is used to test the performance of DDR using a community standard, was extracted from NDF-RT[37] by Li and Lu[22]. It spans 3,250 treatment associations between 799 drugs and 719 diseases. We considered drug information from three data sources: chemical structure, target protein, and side effect. Thus, three 799×799 matrices were used to represent drug similarities between 799 drugs from different perspectives. Similarly, we considered disease information from three data sources: disease phenotype, disease ontology, and disease gene. Thus, three 719×719 matrices were used to represent disease similarities between 719 human diseases from different perspectives. The presence or absence of known associations between drug and disease was denoted by 1 or 0 respectively. Thus, a 799×719 matrix $\boldsymbol{R}$ used to represent the known drug-disease associations. We plotted the statistic of the known drug-disease associations in Figure 2. In that dataset, most of drugs (75%) treat <5 diseases; 18% of drugs treat 5 to 10 diseases; only 7% of drugs treat >10 diseases (Figure 2(a)). Although the disease hypertension has 78 related drugs, 80% of diseases have only <5 drugs; 10% of diseases have 5-10 drugs; and remaining 10% of diseases have >10 drugs.

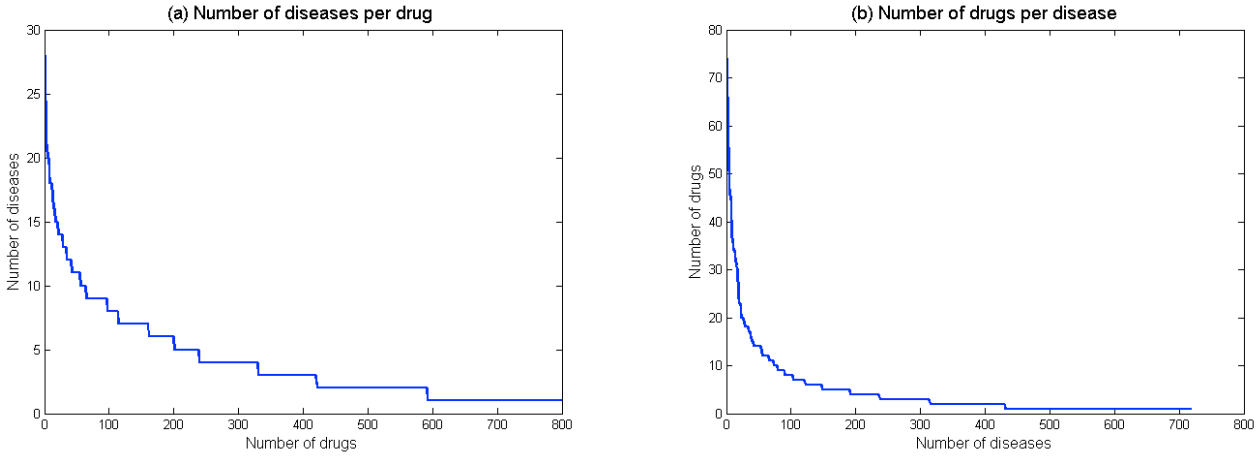


Figure 2. Statistics of the known drug-disease association dataset. (a) The number of indicated diseases per drug. (b) The number of drugs per disease.

Method Comparison. We used a 10-fold cross-validation scheme to evaluate drug repositioning approaches. To ensure the validity of the test cases, we held out all the associations involved with 10% of the drugs in each fold, rather than holding out associations directly. To obtain robust results, we performed 50 independent cross-validation runs, in each of which a different random partition of dataset to 10 parts was used. In our comparisons, we considered five drug repositioning methods: (1) **DDR using Simple Average**. The method only considers reconstruction loss of observed drug-disease associations (i.e., J_0 of objective formula (7) in the methodology section), and assumes each drug/disease source is equally informative. Thus the method uses the average of drug/disease similarity matrices as the integrated drug/disease similarity. (2) **DDR with Weighted Drug Similarity.** The method considers reconstruction losses of observed drug-disease associations and drug similarities (i.e., J_0 and J_1 in objective formula (7)). The method uses the average of disease similarity matrices as integrated disease similarity, and automatically learns drug similarity weight vector (ω) based on the contributions of drug information sources to the prediction. (3) **DDR with Weighted Disease Similarity.** The method considers reconstruction losses of observed drug-disease associations and disease similarities (i.e., J_0 and J_2 in objective formula (7)). The method uses the average of drug similarity matrices as integrated drug similarity, and automatically learns disease similarity weight vector (π) based on the contributions of disease information sources to the prediction. (4) **DDR with Weighted Drug and Disease Similarities.** The method considers all reconstruction losses proposed in the paper (i.e., formula (7) as a whole). The method automatically learns drug similarity weight vector (ω) and disease similarity weight vector (π) together based on the contributions of drug and disease information sources to the prediction. (5) **PREDICT with All Drug and Disease Similarities.** To our knowledge, PREDICT[19] is the only other method could consider both drug and disease information sources. PREDICT uses unweighted geometric mean of pairs of drug-drug and disease-disease similarity measures to construct classification features and subsequently learns a logistic regression classifier that distinguishes between true and false drug-disease associations. PREDICT could not provide weight for each drug/disease information source. Figure 3 shows the averaged ROC curves of 50 runs of the cross-validation for different methods based on the experiment.

Figure 3 shows that our proposed DDR framework is effective for drug repositioning tasks. Without considering reconstruction loss of any similarity measure, DDR using Simple Average obtains an averaged AUC score of 0.7985. When considering weighted drug similarity (i.e., reconstruction loss of drug similarities) or weighted disease similarity (i.e., reconstruction loss of disease similarities), DDRs obtain averaged AUC scores of 0.8508 or 0.8366 respectively. In the experiment, drug-based optimization (i.e., DDR with Weighted Drug Similarity) obtains a higher AUC score than disease-based optimization (i.e., DDR with Weighted Disease Similarity). This could be partially explained with the following reason. The 799 drugs we studied are marketed medications, which usually have rich and precise pharmacological data; thus drug-based optimization might be preferred in this case. For novel drugs or clinical candidates, disease-based optimization might be preferred to overcome missing knowledge in the pharmacology of a drug[38] (e.g., additional targets, unknown side effect). When considering weighted drug similarity and weighted disease similarity together, DDR obtain the highest averaged AUC score (0.8700). The observation

indicates that drug-based optimization and disease-based optimization could be complementary, and computational drug repositioning tasks should optimize both drug similarity and disease similarity. Another observation is PREDICT with All Drug and Disease Similarities obtains an averaged AUC score of 0.8301. Although PREDICT considers drug/disease similarity and utilizes a logistic regression to weigh classification features, the result indicates that their strategy of feature construction (assembles all possible combinations of drug/disease similarity measures together as classification features) is less accurate than DDR.

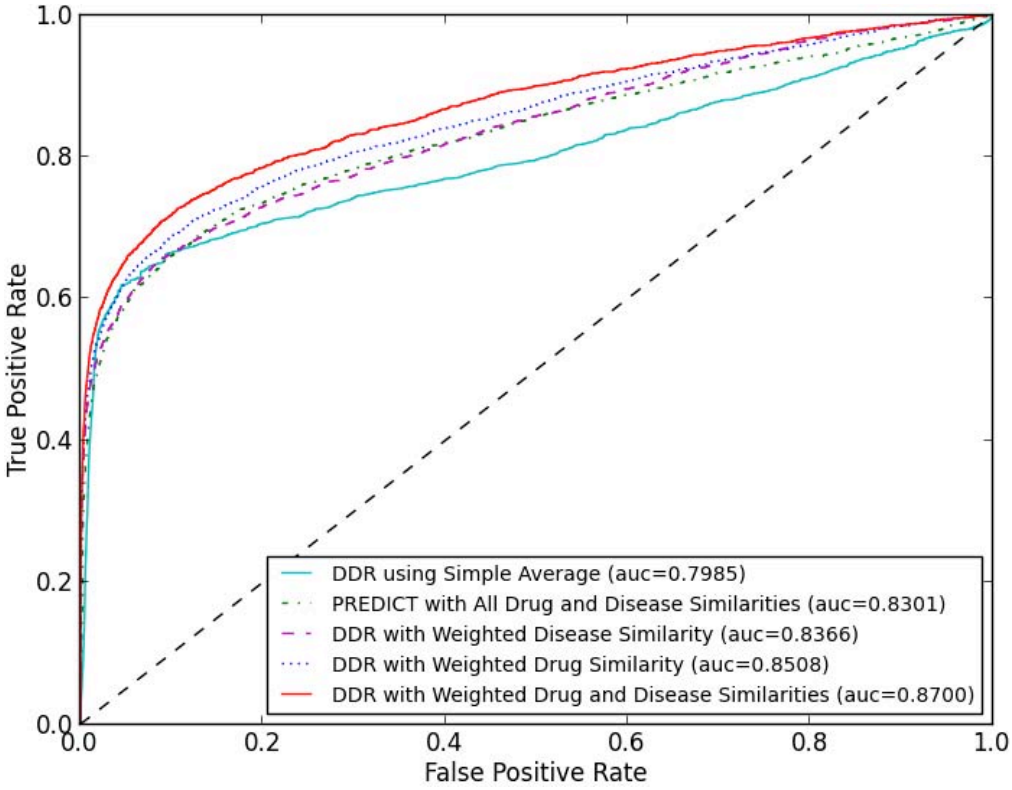


Figure 3. The averaged ROC comparison of five drug repositioning approaches generated from 50 runs of 10-fold cross-validation. Methods are sorted in legend of the figure according to their AUC score.

One "bonus" characteristic of DDR is it provides interpretable importance of different information sources based on their contributions to the prediction. The *i*-th element of drug/disease weight vector ω/π corresponds to the *i*-th drug/disease data sources. Since we constrained ω/π to be in a simplex in problem formula (11), the sum of all elements of ω/π is 1. Obtained from DDR with Weighted Drug and Disease Similarities, the averaged DDR weights of each data source and their standard deviations during the cross-validation experiments are plotted in Figure 4. For drug data sources, chemical structure obtains averaged weight of 0.2744, target protein obtains averaged weight of 0.2295, and side effect obtains a much higher averaged weight of 0.4961 (Figure 4(a)). This could be partially explained with the following reasons. Chemical structure and target protein sources focus on drug's molecular mechanism of action (MOA) from a genotypic perspective. However, the pre-clinical outcomes based on MOA often do not correlate well with therapeutic efficacy in drug development. It is estimated that of all compounds effective in cell assays, only 30% of them could work in animals. Even worse, only 5% of them could work in humans[39]. Side effects are generated when drugs bind to off-targets, which perturb unexpected metabolic or signaling pathways. For marketed drugs, which have relatively complete side effect profiles, side effect information from clinical patients may be seen as valuable read-outs of drug effects directly on human bodies[40] (i.e., with less translational problems). Thus, side effects could server as a promising perspective for drug repositioning. For disease data sources, phenotype obtains averaged weight of 0.4248, disease ontology obtains averaged weight of 0.3958, and disease gene obtains a lower averaged weight of 0.1794 (Figure 4(b)). The lower weight of disease gene data source may be due to the fact that the gap between phenotype (human disease) and genotype (human gene) is too large[41], and the known associations between diseases and genes (obtained from OMIM) are incomplete.

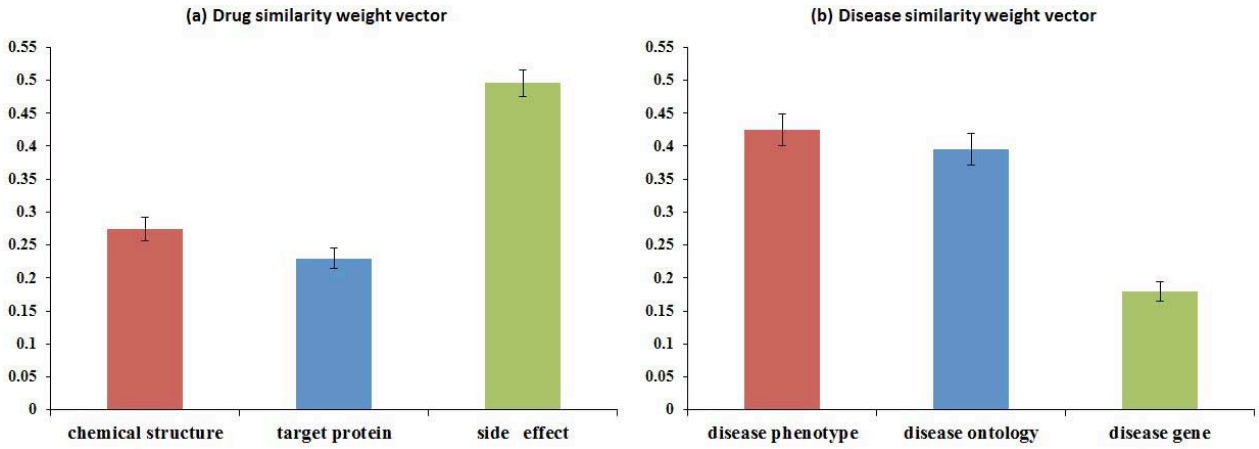


Figure 4. Distribution of averaged weights and standard deviations of the similarity weight vectors obtained by DDR. (a) Drug similarity weight vector ω contains weights of chemical structure, target protein, and side effect data sources. (b) Disease similarity weight vector π contains weights of disease phenotype, disease ontology, and disease gene sources.

Novel Predictions and Case Studies. We performed an additional leave-disease-out experiment to demonstrate the capability of DDR on uncovering drug-disease associations and predicting novel drug candidates for each disease. To ensure the validity of the test cases, we held out all the known drug-disease associations with the tested disease. The validation setting mimics a real-world setting: once rare/unknown diseases without any treatment information arise, a computational drug repositioning method should provide potential drugs based on characteristics (e.g. phenotypes, related genes) of the new diseases and the existing drug/disease similarities. In the experiment, we alternatively leave each disease *i* out and ran DDR (considered weighted drug and disease similarities). More specifically, we set all elements in *i*-th column of matrix $\boldsymbol{R}$ to 0, and used this $\boldsymbol{R}$ along with drug/disease similarity matrices as inputs of DDR. Then we used *i*-th column of the densified estimated matrix $\boldsymbol{\Theta}$ as the drug prediction scores for the disease *i*. In this way, we got prediction scores for all possible associations between the 799 drugs and 719 diseases.

As an example, treatment predictions for Alzheimer's disease (AD) were analyzed. For the six drugs which are known to treat AD, DDR assigned scores of 0.7091 to Selegiline, 0.6745 to Valproic Acid, 0.6348 to Galantamine, 0.5675 to Donepezil, 0.5571 to Tacrine, and 0.5233 to Rivastigmine, which are significantly larger than those of the other 793 drugs (mean and standard deviation are 0.1565±0.1628). Table 2(a) shows the top 10 drugs predicted for AD by our DDR approach. Of the 10 drugs, only three (Selegiline, Valproic Acid, and Galantamine) appear in our known drug-disease association list. The remaining 7 predicted drugs (along with other high-ranked ones in the leave-disease-out experiment) could be considered as drug repositioning candidates for AD. Some predictions are explainable and supported by clinical evidence from ClinicalTrials.gov (i.e., pharmaceutical investigators have been aware of the associations, which are still in the experimental stages). Metformin, a drug commonly used to treat type II diabetes, can help trigger the pathway used to instruct stem cells in the brain to become neural cells[42]. Clinical trial NCT01965756 is under way to evaluate Metformin as a potential therapy for AD. Bexarotene, a skin cancer drug (for cutaneous T-cell lymphoma) that rapidly removed the damaging protein implicated in the progression of the illness from the brains of mice[43], has been tested to treat AD in recent clinical trials (NCT01782742 and NCT02061878). Nilvadipine, a calcium channel blocker (CCB) for treatment of hypertension, also blocks the production of amyloid proteins linked to AD[44]. Nilvadipine has been tested in a clinical trial as a possible treatment for AD in Ireland (NCT02017340).

Another example we analyzed is treatment predictions for Systemic Lupus Erythematosus (SLE). For the three drugs which are known to treat SLE, DDR assigned scores of 0.7269 to Azathioprine, 0.6862 to Triamcinolone, and 0.6374 to Hydroxychloroquine, which are significantly larger than those of the other 796 drugs (mean and standard deviation are 0.1707±0.1617). Table 2(b) shows the top 10 drugs predicted for SLE by our DDR approach. The top 10 predictions include all the three known treatments to SLE, which shows the effectiveness of our method. The remaining 7 predicted drugs (along with other high-ranked ones in the leave-disease-out experiment) could be considered as drug repositioning candidates for SLE. Some predictions are explainable and supported by clinical

evidence from ClinicalTrials.gov. Leflunomide, a pyrimidine synthesis inhibitor, is used to treat moderate to severe rheumatoid arthritis and psoriatic arthritis. A genetic link study shows rheumatoid arthritis and SLE sufferers share a variant of the same STAT4 gene, and therapies developed to treat one disease may possibly be able to treat the other[45]. Therefore, it is not surprising to see Leflunomide is tested as a treatment for SLE in a clinical trial (NCT00637819). Nelfinavir, one of the protease inhibitors, has been approved for use in the treatment of human immunodeficiency virus (HIV). Protease inhibitors have been shown to interfere with binding of anti-double stranded DNA antibodies to their targets (some bindings may lead to organ damage) and may decrease inflammation in SLE[46]. Recently, a clinical trial (NCT02066311) has been proposed to evaluate Nelfinavir as a potential therapy for SLE.

Table 2. Top 10 drugs for diseases Alzheimer's Disease (AD) and Systemic Lupus Erythematosus (SLE) based on DDR predictions

(a) Top 10 drugs predicted for AD

Drug	Prediction Score	Clinical Evidence?
Selegiline*	0.7091	—
Carbidopa	0.6924	No
Amantadine	0.6897	No
Procyclidine	0.6826	No
Valproic Acid*	0.6745	—
Metformin	0.6543	Yes
Bexarotene	0.6426	Yes
Neostigmine	0.6385	No
Galantamine*	0.6348	—
Nilvadipine	0.6159	Yes

(b) Top 10 drugs predicted for SLE

Drug	Prediction Score	Clinical Evidence?
Desoximetasone	0.7409	No
Azathioprine*	0.7269	—
Leflunomide	0.7078	Yes
Fluorometholone	0.7054	No
Triamcinolone*	0.6862	—
Beclomethasone	0.6522	No
Etodolac	0.6445	No
Hydroxychloroquine*	0.6374	—
Nelfinavir	0.6371	Yes
Mercaptopurine	0.6150	No

* denotes the drug is known and approved to treat the disease

Conclusion

We have proposed a general computational framework, called DDR, to explore drug-disease association for drug repurposing hypothesis generation. Our method takes into consideration multiple drug similarities, multiple disease similarities, and known drug-disease associations, to uncover the potential additional associations among other unlinked drugs and diseases. Experimental results demonstrate the effectiveness of the proposed method, and suggest that our method could help identify drug repositioning opportunities, which will benefit patients by offering more effective and safer treatments.

References

1. Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, Schacht AL. How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nat Rev Drug Discov 2010; 9(3):203-214.
2. Berggren R, Moller M, Moss R, Poda P, Smietana K. Outlook for the next 5 years in drug innovation. Nat Rev Drug Discov 2012; 11(6):435-436.
3. Hurle MR, Yang L, Xie Q, Rajpal DK, Sanseau P, Agarwal P. Computational drug repositioning: from data to therapeutics. Clin Pharmacol Ther 2013; 93(4):335-341.
4. Sardana D, Zhu C, Zhang M, Gudivada RC, Yang L, Jegga AG. Drug repositioning for orphan diseases. Brief Bioinform 2011; 12(4):346-356.
5. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, Reich M, Hieronymus H, Wei G, Armstrong SA, Haggarty SJ, Clemons PA, Wei R, Carr SA, Lander ES, Golub TR. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science 2006; 313(5795):1929-1935.
6. Hu G, Agarwal P. Human Disease-Drug Network Based on Genomic Expression Profiles. PLoS ONE 2009; 4(8):e6536.
7. Chiang AP, Butte AJ. Systematic evaluation of drug-disease relationships to identify leads for novel drug uses. Clin Pharmacol Ther 2009; 86(5):507-510.
8. Luo H, Chen J, Shi L, Mikailov M, Zhu H, Wang K, He L, Yang L. DRAR-CPI: a server for identifying drug repositioning potential and adverse drug reactions via the chemical-protein interactome. Nucleic Acids Res 2011; 39(Web Server issue):W492-W498.

9. Dakshanamurthy S, Issa NT, Assefnia S, Seshasayee A, Peters OJ, Madhavan S, Uren A, Brown ML, Byers SW. Predicting new indications for approved drugs using a proteochemometric method. J Med Chem 2012; 55(15):6832-6848.
10. Cheng F, Zhou Y, Li J, Li W, Liu G, Tang Y. Prediction of chemical-protein interactions: multitarget-QSAR versus computational chemogenomic methods. Mol Biosyst 2012; 8(9):2373-2384.
11. Sanseau P, Agarwal P, Barnes MR, Pastinen T, Richards JB, Cardon LR, Mooser V. Use of genome-wide association studies for drug repositioning. Nat Biotechnol 2012; 30(4):317-320.
12. Li J, Lu Z. Pathway-based drug repositioning using causal inference. BMC Bioinformatics 2013; 14(Suppl 16):S3.
13. Iorio F, Bosotti R, Scacheri E, Belcastro V, Mithbaokar P, Ferriero R, Murino L, Tagliaferri R, Brunetti-Pierri N, Isacchi A, di Bernardo D. Discovery of drug mode of action and drug repositioning from transcriptional responses. Proc Natl Acad Sci 2010;107(33):14621-14626.
14. Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P. Drug target identification using side-effect similarity. Science 2008;321(5886):263–266.
15. Yang L, Agarwal P. Systematic Drug Repositioning Based on Clinical Side-Effects. PLoS ONE 2011;6(12):e28025.
16. Ye H, Liu Q, Wei J. Construction of drug network based on side effects and its application for drug repositioning. PLoS One 2014; 9(2):e87864.
17. Sirota M, Dudley JT, Kim J, Chiang AP, Morgan AA, Sweet-Cordero A, Sage J, Butte AJ. Discovery and preclinical validation of drug indications using compendia of public gene expression data. Sci Transl Med 2011; 3(96):96ra77.
18. Wang K, Sun J, Zhou S, Wan C, Qin S, Li C, He L, Yang L. Prediction of drug-target interactions for drug repositioning only based on genomic expression similarity. PLoS Comput Biol 2013; 9(11):e1003315.
19. Gottlieb A, Stein GY, Ruppin E, Sharan R. PREDICT: a method for inferring novel drug indications with application to personalized medicine. Mol Syst Biol 2011; 7:496.
20. Zhang P, Agarwal P, Obradovic Z. Computational Drug Repositioning by Ranking and Integrating Multiple Data Sources. Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases 2013; Part III: 579-594.
21. Wang Y, Chen S, Deng N, Wang Y. Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data. PLoS One 2013; 8(11):e78518.
22. Li J, Lu Z. A New Method for Computational Drug Repositioning Using Drug Pairwise Similarity. Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine 2012.
23. Huang YF, Yeh HY, Soo VW. Inferring drug-disease associations from integration of chemical, genomic and phenotype data using network propagation. BMC Med Genomics 2013; 6(Suppl 3):S4.
24. Napolitano F, Zhao Y, Moreira VM, Tagliaferri R, Kere J, D'Amato M, Greco D. Drug repositioning: a machine-learning approach through data integration. J Cheminform 2013; 5(1):30.
25. PubChem substructure description [ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.pdf]
26. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH. PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Res 2009; 37(Web Server Issue):W623-W633.
27. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, Maciejewski A, Arndt D, Wilson M, Neveu V, Tang A, Gabriel G, Ly C, Adamjee S, Dame ZT, Han B, Zhou Y, Wishart DS. DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res 2014; 42(Database Issue): D1091-1097.
28. Smith TF, Waterman MS, Burks C. The statistical distribution of nucleic acid similarities. Nucleic Acids Res 1985; 13(2):645-665.
29. Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P. A side effect resource to capture phenotypic effects of drugs. Mol Syst Biol 2010; 6:343.
30. van Driel MA, Bruggeman J, Vriend G, Brunner HG, Leunissen JA. A text-mining analysis of the human phenome. Eur J Hum Genet 2006; 14(5):535-542.
31. Lipscomb CE. Medical Subject Headings (MeSH). Bull Med Libr Assoc 2000; 88(3):265–266.
32. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 2005; 33(Database issue):D514-D517.
33. Schriml LM, Arze C, Nadendla S, Chang YW, Mazaitis M, Felix V, Feng G, Kibbe WA. Disease Ontology: a backbone for disease semantic integration. Nucleic Acids Res 2012; 40(Database issue):D940-D946.
34. Li J, Gong B, Chen X, Liu T, Wu C, Zhang F, Li C, Li X, Rao S, Li X. DOSim: an R package for similarity between diseases based on Disease Ontology. BMC Bioinformatics 2011; 12:266.
35. Wang F, Wang X, Li T. Generalized cluster aggregation. Proceedings of the International Joint Conference on Artificial Intelligence 2009.
36. Bertsekas DP. Block coordinate descent methods. in Nonlinear Programming 2nd Edition 1999.
37. Carter JS, Brown SH, Bauer BA, Elkin PL, Erlbaum MS, Froehling DA, Lincoln MJ, Rosenbloom ST, Wahner-Roedler DL, Tuttle MS. Categorical information in pharmaceutical terminologies. AMIA Annu Symp Proc 2006:116-120.
38. Dudley JT, Deshpande T, Butte AJ. Exploiting drug-disease relationships for computational drug repositioning. Brief Bioinform 2011; 12(4):303-311.
39. Pammolli F, Magazzini L, Riccaboni M. The productivity crisis in pharmaceutical R&D. Nat Rev Drug Discov 2011; 10(6):428-438.
40. Duran-Frigola M, Aloy P. Recycling side-effects into clinical markers for drug repositioning. Genome Med 2012; 4(1):3.
41. Chen Y, Wu X, Jiang R. Integrating human omics data to prioritize candidate genes. BMC Med Genomics 2013; 6:57.
42. Wang J, Gallagher D, DeVito LM, Cancino GI, Tsui D, He L, Keller GM, Frankland PW, Kaplan DR, Miller FD. Metformin activates an atypical PKC-CBP pathway to promote neurogenesis and enhance spatial memory formation. Cell Stem Cell 2012; 11(1):23-35.
43. Cramer PE, Cirrito JR, Wesson DW, Lee CY, Karlo JC, Zinn AE, Casali BT, Restivo JL, Goebel WD, James MJ, Brunden KR, Wilson DA, Landreth GE. ApoE-directed therapeutics rapidly clear β-amyloid and reverse deficits in AD mouse models. Science 2012; 335(6075):1503-1506.
44. Paris D, Quadros A, Humphrey J, Patel N, Crescentini R, Crawford F, Mullan M. Nilvadipine antagonizes both Abeta vasoactivity in isolated arteries, and the reduced cerebral blood flow in APPsw transgenic mice. Brain Res 2004; 999(1):53-61.
45. Remmers EF, Plenge RM, Lee AT, Graham RR, Hom G, Behrens TW, de Bakker PI, Le JM, Lee HS, Batliwalla F, Li W, Masters SL, Booty MG, Carulli JP, Padyukov L, Alfredsson L, Klareskog L, Chen WV, Amos CI, Criswell LA, Seldin MF, Kastner DL, Gregersen PK. STAT4 and the risk of rheumatoid arthritis and systemic lupus erythematosus. N Engl J Med 2007; 357(10):977-986.
46. Bloom O, Cheng KF, He M, Papatheodorou A, Volpe BT, Diamond B, Al-Abed Y. Generation of a unique small molecule peptidomimetic that neutralizes lupus autoantibody activity. Proc Natl Acad Sci 2011; 108(25):10255-10259.

Using Language Models to Identify Relevant New Information in Inpatient Clinical Notes

Rui Zhang, PhD[1,2], Serguei V. Pakhomov, PhD[1,3], Janet T. Lee, MD, MS[2], Genevieve B. Melton, MD, MA[1,2]

[1]Institute for Health Informatics, [2]Department of Surgery, and [3]College of Pharmacy, University of Minnesota, Minneapolis, MN, USA

Abstract

Redundant information in clinical notes within electronic health record (EHR) systems is ubiquitous and may negatively impact the use of these notes by clinicians, and, potentially, the efficiency of patient care delivery. Automated methods to identify redundant versus relevant new information may provide a valuable tool for clinicians to better synthesize patient information and navigate to clinically important details. In this study, we investigated the use of language models for identification of new information in inpatient notes, and evaluated our methods using expert-derived reference standards. The best method achieved precision of 0.743, recall of 0.832 and F1-measure of 0.784. The average proportion of redundant information was similar between inpatient and outpatient progress notes (76.6% (SD=17.3%) and 76.7% (SD=14.0%), respectively). Advanced practice providers tended to have higher rates of redundancy in their notes compared to physicians. Future investigation includes the addition of semantic components and visualization of new information.

Introduction

Clinical note documentation in Electronic Health Record (EHR) systems provides clinicians with the ability to store and share detailed contextual health information about patients for the primary purposes of communication, documentation, and billing. Most EHR systems allow functionality of "copy-and-pasting" of texts from a previous note to the current clinical note, which shortens the time clinicians spend on documenting encounters. However, an unintended consequence of copy-and-paste practices is creation of large amounts of replicated patient information within the EHR, especially in patients with complicated care or long hospital stays, thus making notes longer and less readable [1-3].

Notes with significant amounts of redundant information, combined with a large numbers of notes, increases the cognitive burden of clinicians [3-8]. In a time-constrained clinical practice environment, clinicians are limited in their review and synthesis of patient notes. Redundant information in clinical notes creates noise that masks new and clinically relevant information within notes. Moreover, redundant information in clinical notes can also contain a mixture of outdated information or errors in the copied information, making it difficult for clinicians to interpret the data in these notes effectively [4].

Several studies have reported the effect of copy-and-paste documentation behavior in clinical practice [9-11]. Redundant information can also create an integrity problem in clinical notes and create an impression that a note containing significant amounts of copied information is from an author who may not have read or independently constructed the note [9]. For example, in one report, a nurse observed that a historical event that occurred four years prior was subsequently repeated in many clinical notes afterwards [9]. This problem may also result in decreased use of and reliance on the information within clinical notes [6]. Other studies have demonstrated that the combination of redundant information and increased note length results in information overload and difficulties in finding information within notes, thus making narrative communication via notes less effective and efficient for patient care [10, 11].

Previous studies have found large amounts of redundant information in both inpatient and outpatient notes with automated methods [12, 13]. Wrenn et al. used global alignment techniques to quantify redundancy in inpatient clinical notes [12]. They found an average of 78% and 54% information duplicated from previous documents in signout and progress notes, respectively. Zhang et al. modified the Needleman-Wunsch algorithm to quantify redundancy and investigate the redundancy patterns in outpatient clinical notes [13].

This work demonstrated that redundancy scores appeared to have a cyclic pattern for each individual patient but also that the overall volume of redundant information increased over time.

In this study, we investigated the use of statistical language models to identify clinically relevant new information in progress notes during patient hospital stays, applied a number of discounting models to potentially improve performance of the applied language models, and sought to compare the quantity of redundant information in clinical notes between inpatient and outpatient clinical settings.

Background

N-gram model

Statistical language modeling (SLM) is widely used for many NLP tasks, such as part-of-speech tagging, parsing, information retrieval, and machine translation [14, 15]. SLM assigns a probability to a set of *n* words based on a probability distribution from a specific corpus. An *n*-gram model is a typical language model (LM), which estimates the probability of an *i*-th word in the context of *n* previous words. To simplify the calculation of the probability of the word, the *Markov assumption* states that the probability of the word is only based on the prior few words instead of all previous words. One commonly used statistical estimate called Maximum Likelihood Estimates (MLE) is unsuitable for statistical inference in NLP due to the sparseness of the data. MLE assigns zero to unseen events, and the zeros will propagate since the probability of a long string is computed by multiplying probabilities of subparts.

Discounting methods

Discounting is the process of replacing the original counts with modified counts based on the mathematic formula to redistribute the probability in order to avoid assigning zero probability to unseen events due to the sparseness of the training sample texts [14, 15].

The simplest one is Laplace smoothing (also called add-one method):

$$P_{Lap}(w_1 \cdots w_n) = \frac{C(w_1 \cdots w_n)+1}{N+B}$$

where $C(w_1 \ldots w_n)$ is the count of the *n*-gram $w_1 \ldots w_n$, N is the number of training instances, and B is the vocabulary size.

The Good-Turing (GT)[a] estimator is an improved method for determining the probability or frequency of *n*-grams*:*

$$if\ 1 \le C(w_1 \cdots w_n) \le gt\max,\ P_{GT}(w_1 \cdots w_n) = \frac{C(w_1 \cdots w_n)}{C(w_1 \cdots w_{n-1})} \frac{C'(w_1 \cdots w_n)/C(w_1 \cdots w_n) - A}{(1-A)}$$

$$where\ C'(w_1 \cdots w_n) = (C(w_1 \cdots w_n)+1)\frac{n[C(w_1 \cdots w_n)+1]}{n[C(w_1 \cdots w_{n-1})]},\ A = (gt\max+1)\frac{n[gt\max+1]}{n[1]}$$

where $n[a]$ is the notion for the number of unique n-grams that occurred *a* times. This smoothing method substitutes low frequency *n*-grams and is quite accurate. It is also suitable for large numbers of observations of data and assumes that the distribution is binomial. The GT estimator works well for *n*-grams, despite the fact that words and *n*-grams do not follow a binomial distribution.

Ney and Essen proposed a linear discounting model for estimating frequencies of *n*-grams:

$$if\ C(w_1 \ldots w_n) = r,\ P(w_1 \ldots w_n) = \begin{cases} (1-\alpha)r/N & if\ r > 0 \\ \alpha / N_0 & otherwise \end{cases}$$

where α is a constant slightly less than one.

[a] http://www-speech.sri.com/projects/srilm/manpages/ngram-discount.7.html

These estimates make the probability of unseen events a small number instead of zero and rescale the other probabilities to ensure that the probability mass is equal to 1.0. However, the Ney and Essen linear discounting method does not work as well for higher frequency *n*-grams. In this study, we directly tested the use of these three discounting methods for building language models as part of our method evaluation.

Methods

Data Collection

EHR notes were retrieved from University of Minnesota Medical Center affiliated Fairview Health Services. For this study, we randomly selected patients in the inpatient clinical setting. These notes were extracted in text format from the Epic™ EHR system[b] during a one-year period (05/2011 to 05/2012). For simplicity, we limited the notes to the progress notes authored by the primary team providers including physicians, residents, and advanced practice providers (including physician assistants (PA) and nurse practitioners (NP)). All notes were arranged chronologically for a given individual patients. Institutional review board approval was obtained and informed consent was waived for this minimal risk study.

Manually reviewed annotation as gold standard

Each series of in-patient notes starts with the patient's history and physical (H&P) note, followed by a set of progress notes, and ending with a discharge summary. Starting from the progress notes in a series of patient notes, two 4th-year medical students were asked to identify new and clinically relevant information based on all preceding documents chronologically within the same hospital stay using their clinical judgment. Each medical expert annotated progress notes from ten patients with one patient's set of notes overlapping with both. Annotation of new information in clinical notes was implemented through the publically available software General Architecture for Text Engineering (GATE)[c]. GATE allows for annotation of text and XML outputs through a graphical user interface, with a customized annotation schema.

We first asked the two medical students to annotate one sample note and then to compare and discuss the annotations with each other to reach a consensus on the annotation categories and standards for new information (definitions and examples shown in Table 1). Each medical student later manually annotated another one set of patient notes based on the same historical notes to measure inter-rater agreement. Cohen's Kappa statistic and percent agreement were used to assess inter-rater reliability at a sentence or statement level.

Overall, longitudinal inpatient clinical notes from 20 sets of patient notes were annotated for this study. Each medical student annotated 50 notes, with a total of 100 annotated progress notes in this study. Fifty of these notes were used for training and system development and another fifty for evaluation. We also asked a 6th-year surgical resident (JL) with clinical practice experience to proofread the annotated notes, particularly around the addition of diagnostic studies (such as radiology reports) and make modifications to ensure the quality of the gold standard. We refereed annotations before and after JL's modification as initial annotation and revised annotation, respectively.

Results of automated methods were compared to the reference standard (initial annotation and revised annotation) and performance reported including accuracy, precision, recall, and F1-measure at a sentence or statement level. Methods were also evaluated comparing redundancy by different author role. For example, measuring redundancy of notes written by physician and residents as well as notes by physician assistants and nurse practitioners. Performance of the best method on identification of new information for different note sections was also tested.

[b] http://www.epic.com
[c] http://gate.ac.uk

Table 1. Classification, definition and examples of new clinical information.

Category	Code	Definition	Example
Clinical	Additional history	New information on patient's medical history occurring prior to hospitalization including prior diagnoses, surgeries, labwork/imaging, immunizations, medications, allergies/reactions, family history, and social history (e.g. sexual history, drug use, intake, work history, hobbies, marital status)	Recently admitted elsewhere for a pneumonia.
	Assessment	Changes in the medical diagnosis or differential diagnosis regarding patient's hospitalizing condition as assessed by the note's author	I am thinking this was all related to RF and CHF exacerbation.
	Changes in symptoms	Changes in patient reported symptoms including new complaints or improvement or worsening of existing concerns.	He says he is breathing better.
	Medications	Changes in medication, dosage, or route of administration including returns to pre-hospitalization usage. Also includes changes in fluids/electrolytes/nutrition administration and immediate post-administration clinical status.	Metoprolol 100 mg po bid changed to carvedilol 25 mg po bid
	New imaging and diagnostic studies	Results of imaging previously unmentioned during hospitalization such as X-rays, computerized tomography (CT scans), and solography.	TTE shows grade III diastolic dysfunction with elevated RSVP indicating pulmonary hypertension
	New labwork	Results of labwork previously unmentioned during hospitalization such as the basic metabolic panel, liver function tests, and blood glucose levels. Also includes fluid intake and output.	Wound cx: Gram + cocci UA shows a protein level of 300
	New plan	Changes in clinical care plan for patient care	Send sample for C. Diff studies
	Patient status	Changes in the patient's condition as reported by any healthcare provider	Perhaps a little improved today
	Physical exam	Daily physical examination as performed by clinician	Constitutional: Awake, alert, cooperative, no apparent distress
	Procedure	Procedures performed during hospitalization and related post-operative concerns	C2-4 laminectomy
	Vitals	Daily vitals including temperature, blood pressure, heart rate, and respiratory rate.	Temp: [98.1 F (36.7 C)-99F (37.2 C)] 98.2 F (36.8 C)
Non-clinical	Author	Name and degree of the note's author if previously unmentioned	LastName, FirstName, MD
	Change in service	Patient transfer to other hospital services or clinical sites	Patient will go to 6A for close Neuro-surgical monitoring.
	Date and time	Date and time of note signing by the original author	08/03/12 1148
	Social context	Changes in social history and situation relevant to the patient's condition and care including patient's preference to management	Dialysis discussed with patient, he prefers to wait until July.

Automated methods

We used different *n*-gram models with and without discounting algorithms. We only focused on bigram models since our prior studies[16] have showed the bigram models outperformed than other *n*-gram models. The methods include six steps: 1) text preprocessing, 2) removal of classic stopwords[d] and term frequency - inverse document frequency (TF-IDF) stopwords, 3) lexical normalization, 4) baseline modeling, 5) modification with discounting algorithms, and 6) application of heuristic rules to classify clinical relevance. The details of these steps are as follows:

Step 1: All progress notes were ordered by time for individual patients and were separated into sentences or sections. We used regular expressions for sentence splitter and word tokenization.

Step 2: Remove both classic stopwords and stopwords defined by optimal threshold of TFIDF distribution based on the entire note corpus. This step deemphasizes these less important words for building the language models.

Step 3: Use lexical variant generation (LVG)[17] to normalize lexically different forms of the same term as equivalent when building the language models.

Step 4: Bigrams were counted in all previous notes for each individual patient. Probability of the bigrams without discounting algorithms was calculated and an optimal threshold value was used to identify new in the target progress notes.

Step 5: Discounting algorithms such as Laplace, Good-Turning, and Ney-Essen (details in Background) were used to calculate the probability of bigrams. An optimal threshold probability value was used to identify new versus redundancy information.

Step 6: Develop heuristic rules to judge the clinically relevance on section content, clinical note headers, signatures. For example, vitals in all notes are judged as relevant new information. All note headers and footer are non-relevant information.

Comparison of information redundancy between inpatient and outpatient notes

To investigate the difference of redundancy in outpatient and inpatient clinical notes, we calculated the information redundancy based on the reference standards using the below equation:

$$\text{Redundancy percentage} = 100 \times \left(1 - \frac{\#\text{ sentences with relevant new information}}{\#\text{ sentences}}\right)$$

Note that the medical experts and residents only annotated the new and clinically relevant information. Thus non-clinical new information such as header of the notes, signatures *etc*. was excluded. Averages, standard deviations, interquartile ranges of redundant and irrelevant information percentages for clinical notes in different clinical settings and note types were calculated.

Results

Annotation evaluation and method performance

Two medical students showed a good agreement on initial set of annotations for identifying new information in the overlapping annotations (Cohen's Kappa coefficient of 0.83 and percentage agreement of 92%). On subsequent review, JL found additional new information (average 3.80 sentences per note) and incorrectly annotated information (average 0.07 sentences per note). This resulted in two reference standards – the initial and the revised one. The performance characteristics of various algorithms on both reference standards are listed in Table 2. Generally, all discounting methods performed better than the baseline, although three discounting algorithms did not change significantly. Compared with revised annotation, the methods' precision increased significantly with small drop in recall. Recall of methods in notes written by advanced practice providers were higher than notes by physician or residents, while the precision was much lower. The bigram model with the Ney-Essen algorithm performed the best among these methods, achieving a recall of 0.832, a precision of 0.743, and F1-measure of 0.784 for all notes after JL annotation.

[d] http://www.textfixer.com/resources/common-english-words.txt

Table 2. Performance of algorithms on identification of clinically relevant new information. Precision = TP/(TP+FP), Recall = TP/(TP+FN), F1-Measure = 2×Precision×Recall/(Precision+Recall).

Author type: All

Algorithms	Initial Annotation			Revised Annotation		
	Recall	Precision	F1-measure	Recall	Precision	F1-measure
Baseline	0.812	0.572	0.671	0.807	0.645	0.717
LapLace	0.827	0.654	0.730	0.826	0.728	0.774
Good-Turning	0.834	0.669	0.742	0.829	0.735	0.779
Ney-Essen	0.841	0.680	0.752	0.832	0.743	0.784

Author type: Physician & Resident

Algorithms	Initial Annotation			Revised Annotation		
	Recall	Precision	F1-measure	Recall	Precision	F1-measure
Baseline	0.800	0.587	0.677	0.800	0.667	0.733
LapLace	0.817	0.670	0.707	0.812	0.746	0.762
Good-Turning	0.824	0.681	0.746	0.820	0.758	0.788
Ney-Essen	0.830	0.692	0.755	0.825	0.767	0.795

Author type: Physician Assistant & Nurse Practitioner (Advanced Practice Providers)

Algorithms	Initial Annotation			Revised Annotation		
	Recall	Precision	F1-measure	Recall	Precision	F1-measure
Baseline	0.861	0.506	0.637	0.857	0.553	0.651
LapLace	0.918	0.517	0.662	0.917	0.576	0.707
Good-Turning	0.923	0.522	0.667	0.920	0.584	0.714
Ney-Essen	0.931	0.531	0.677	0.927	0.589	0.720

Performance on identification of new information in different sections

Percentages of new information identified based on the revised annotations and separated by section, are also shown in Table 3. Top three sections with most new information are Physical Exam (33%), Assessment & Plan (27%), and Medication (14%). Performance of the best method (Ney-Essen algorithm) was also evaluated in different sections. Recall of this discounting method in the Medication section was the highest, and precision and F1-measure in the Vitals section were the best.

Table 3. Identification of new information in different sections on revised annotation.

Sections	Percentages of new information in notes	Recall	Precision	F1-measure
Physical Exam	33%	0.860	0.820	0.839
Assessment & Plan	27%	0.910	0.612	0.732
Medication	14%	0.982	0.764	0.859
Vitals	10%	0.939	0.957	0.948
Imaging	5%	0.723	0.933	0.815

Redundant and irrelevant information in inpatient versus outpatient clinical notes

Statistical descriptions of information redundancy for clinical notes are listed in Table 4. It is surprising that outpatient clinical notes contain redundant information at the same level (about 76%) as those in inpatient clinical notes. Notes written by advanced practice providers contain more redundant information than those by physicians and residents. Notes written by physicians had a smaller arithmetic mean and a larger standard deviation of redundancy than the notes by residents, fellows, and advanced practice providers.

Table 4. Redundant and irrelevant information for different clinical settings. PA, physician assistant; NP, nurse practitioner. Redundancy includes the non-clinical but new information.

Clinical Setting	Author Type	# Notes	Redundancy/irrelevancy mean (standard deviation)	Redundancy/irrelevancy (interquartile range)
Inpatient	All	100	76.6% (17.3%)	(70.6%, 87.9%)
	Physician	57	73.3% (19.1%)	(63.0%, 86.8%)
	Resident or Fellow	15	84.4% (10.0%)	(66.3%, 87.3%)
	PA or NP	27	84.5% (7.3%)	(82.6%, 89.4%)
Outpatient	Physician	90	76.7% (14.0%)	(72.4%, 86.2%)

Discussion

Our investigation of patterns of relevant new information in the inpatient clinical practice highlights the issue of redundancy of clinical information in EHR documentation, which has been increasingly gaining the interest of clinicians and informaticians in recent years. Automated methods to accurately identify and visualize relevant new information represent a potential way to improve the clinicians' reviewing process. Although researchers have developed some preliminary methods to deal with redundancy, most previous evaluations do not include the clinicians' own judgments based on clinical experience as a gold standard and thus fall short in evaluating these methods. However, it is vital to include clinicians' views on redundant information for the development of the methods since they will be potential end users of any future system. In this study, we focused on the development of automated methods to identify relevant new information in inpatient clinical notes as well as evaluation for the methods by comparing with the reference standards annotated by the end users - clinicians.

Constructing a quality reference standard is an important but challenging task to support the development of robust automatic methods. We followed the same process as our previous study [18] for reference standard: smaller sample annotation, discussion to reach a consensus, and then larger sample annotation. Although the annotation code book (Table 1) was meant to help with consistency of our medical student coders, we still found that they missed new information and our modified standard improved the consistency of these annotations. For instance, in the original set of annotations, coders sometimes ignored the changes of current medication list, including addition of a new drug or discontinuation of a current medication ("DISCONTD: sodium chloride 0.9 % flush 10 mL" as an example), and the possible reason is that it is difficult to find small changes in a long and tightly laid out section, such as a medication section. Other disagreements were from the different identification of new information boundary. For example, one annotator included the section title (e.g, objective, exam, assessment and plan) as new information if there was new information within the section; another annotator chose not to identify the title of section as new information.

After comparing the results produced by automated methods with reference standards, we found that discounting algorithms help to improve the performance of the methods. All methods did not perform very well on precision as our methods were developed on the lexical level of the texts and did not consider the semantic meaning of the sentences. For example, the sentence "continue to hold all nephrotoxic meds" in the target note was not identified as redundant by comparison with the sentence "hold lasix, lisinopril and spironolactone given acute kidney injury" in a previous note. Due to the limitations of the methods, they cannot recognize that specific drugs such as lasix, lisinopril and spironolactone are nephrotoxic medications.

The text formats in different sections result in variability in performance on different sections. "Assessment/Plan" (A&P) is one of the longest sections in the notes, where clinicians input their thoughts and tend to rephrase sentences even for the same meaning from the previous notes. For example, in the above example, the physician typed "nephrotoxic medications" instead of specific medication names "Lasix, lisinopril and spironolactone". Thus, this probably is the main reason why the precision in the A&P section was the lowest. As for the Vitals section, the format is unique as "Temp: 97.2 °F (36.2 °C) | BP: 100/75 | Resp: 20…" in most notes, allowing the methods to easily recognize the pattern, resulting in a higher precision and recall compared with other sections. Similarly, the "Physical Exam" section usually contains the short statement for each part such as "General: Lying nearly flat in bed, comfortable, NAD, Interactive".

Analysis indicated a high level of redundant and irrelevant information in inpatient progress notes (average 76.6% for 100 notes), although we included the irrelevant note format or noise in this calculation. This number may be different if we were to consider a larger dataset. To our knowledge, there is no prior study reporting the percentage of redundant information in outpatient notes. Surprisingly, the redundancy in outpatient notes contains the same amount of redundant information. One reason for this is that we only selected chronically diseased patients, allowing the larger sets of longitudinal clinical notes for our previous study. Another reason is that all those notes for calculating redundancy in outpatient clinical setting were the last three notes from each set of individual patients. In our previous study [13], we found that the information redundancy of longitudinal outpatient notes was increasing over time. In other words, the last three notes tend to contain relatively higher redundancy than the earlier notes averagely. Therefore, the actual percentages of information redundancy for the entire set of outpatient notes could be lower than the reported number (76.7%) here in Table 3. In addition, we observed that advanced practice providers had higher levels of redundancy and less variability in this (lower standard deviation) than physicians. While it is unclear why this was the case, we speculate that physician providers have more diagnostic and case-based reasoning in notes with significant clinical events, and advanced practice providers are more prescriptive in their narrative. Future studies are needed to both confirm these findings on a larger corpus, as well as perform an analysis of why these differences may exist.

Our methods have certain limitations. All methods focused only on the lexical level. Semantic level issues were out of the scope of this paper, such as co-reference (e.g., "it", "this") and experiencer detection (e.g., "patient", "sister"). For example, "Pt has diabetes" and "His mother has diabetes" shared most of the words, but they are semantically different as the experiencers are changed. Acronym and symbol disambiguation were also not included in the study. Moreover, relevant new information was only limited to the addition of information in the newer notes. The deletion of relevant new information in the more recent clinical notes was not considered in this study. Due to the asymmetric nature of the new information identification process, deletion of relevant information can only be obtained by comparing the object notes and target note in reverse and warrants additional investigation. Future research will add more semantic components to make the system more accurate and comprehensive, and design the ways to visualize the relevant new information by incorporating within existing EHR systems. This implementation will ultimately enhance the efficiency of reviewing and using clinical documentation, and improve the satisfaction of clinicians with EHR systems.

Conclusion

We developed language models with discounting algorithms to identify relevant new information in inpatient progress notes, and evaluated the performance by building up and comparing with a medical expert-derived reference standard. Inpatient clinical notes have approximately the same amount (76%) of redundant or irrelevant information as outpatient clinical notes. Further investigation is needed to improve the performance of the system and visualize the information in EHR systems to enhance the efficiency of using clinical documentation.

Acknowledgments

This research was supported by the Agency for Healthcare Research & Quality grant (#1R01HS022085-01) (GM), and University of Minnesota Clinical and Translational Science Award (#8UL1TR000114-02) (Blazer). The authors thank Fairview Health Services for support of this research.

References

1. Markel A. Copy and paste of electronic health records: a modern medical illness. *Am J Med*. 2010 May;123(5):e9.
2. Hirschtick RE. A piece of my mind. Copy-and-paste. *JAMA*. 2006 May 24;295(20):2335-6.
3. Yackel TR, Embi PJ. Copy-and-paste-and-paste. *JAMA*. 2006 Nov 15;296(19):2315; author reply -6.
4. Hammond KW, Helbig ST, Benson CC, Brathwaite-Sketoe BM. Are electronic medical records trustworthy? Observations on copying, pasting and duplication. *AMIA Annu Symp Proc*. 2003:269-73.
5. Weir CR, Hurdle JF, Felgar MA, Hoffman JM, Roth B, Nebeker JR. Direct text entry in electronic progress notes. An evaluation of input errors. *Methods Inf Med*. 2003;42(1):61-7.
6. Hripcsak G, Vawdrey DK, Fred MR, Bostwick SB. Use of electronic clinical documentation: time spent and team interactions. *J Am Med Inform Assn*. 2011 Mar;18(2):112-7.
7. Patel VL, Kaufman DR, Arocha JF. Emerging paradigms of cognition in medical decision-making. *J Biomed Inform*. 2002 Feb;35(1):52-75.
8. Reichert D, Kaufman D, Bloxham B, Chase H, Elhadad N. Cognitive analysis of the summarization of longitudinal patient records. *AMIA Annu Symp Proc*. 2010;2010:667-71.
9. Embi PJ, Weir CR, Efthimiadis EN, Thielke S, Hadeen A, Hammond K. Computerized provider documentation: findings and implications of a multisite study of clinicians and administrators. *J Am Med Inform Assoc*. 2013 Jul-Aug;20(4):718-26.
10. Embi PJ, Yackel TR, Logan JR, Bowen JL, Cooney TG, Gorman PN. Impacts of computerized physician documentation in a teaching hospital: perceptions of faculty and resident physicians. *J Am Med Inform Assoc*. 2004 Jul-Aug;11(4):300-9.
11. Weir CR, Nebeker JR. Critical issues in an electronic documentation system. *AMIA Annu Symp Proc*. 2007:786-90.
12. Wrenn JO, Stein DM, Bakken S, Stetson PD. Quantifying clinical narrative redundancy in an electronic health record. *J Am Med Inform Assoc*. 2010 Jan-Feb;17(1):49-53.
13. Zhang R, Pakhomov S, MaInnes BT, Melton GB. Evaluating Measures of Redundancy in Clinical Texts. *AMIA Annu Symp Proc*. 2011:1612-20.
14. Manning CD, SchÜtze H. Foundations of Statistical Natural Language Processing. Cambridge, Massachusetts: The MIT Press; 2003.
15. Jurafsky D, Martin JH. Speech and Language Processing. Upper Saddle River, NJ: Prentice Hall; 2009.
16. Zhang R, Pakhomov S, Melton GB. Automated identification of relevant new information in clinical narrative. *2nd ACM SIGHIT Inter Health Inform (IHI) Symp Proc*. 2012: 837-41.
17. McCray AT, Srinivasan S, Browne AC. Lexical methods for managing variation in biomedical terminologies. *Proc Annu Symp Compt Appl Med Care* 1994:5.
18. Zhang R, Pakhomov S, Lee JT, Melton GB. Navigating longitudinal clinical notes with an automated method for detecting new information. *Studies in health technology and informatics*. 2013;192:754-8.

Developing Analytical Inspection Criteria for Health IT Personnel with Minimum Training in Cognitive Ergonomics: A Practical Solution to EHR Improving EHR Usability

Zhen Zhang, MD, MPH, MSc, Amy Franklin, PhD, Muhammad Walji, PhD, Jiajie Zhang, PhD, Yang Gong, MD, PhD
The University of Texas School of Biomedical Informatics, Houston, TX
National Center for Cognitive Informatics and Decision Making in Healthcare

Abstract

EHR usability has been identified as a major barrier to care quality optimization. One major challenge of improving EHR usability is the lack of systematic training in usability or cognitive ergonomics for EHR designers/developers in the vendor community and EHR analysts making significant configurations in healthcare organizations. A practical solution is to provide usability inspection tools that can be easily operationalized by EHR analysts. This project is aimed at developing a set of usability tools with demonstrated validity and reliability. We present a preliminary study of a metric for cognitive transparency and an exploratory experiment testing its validity in predicting the effectiveness of action-effect mapping. Despite the pilot nature of both, we found high sensitivity and specificity of the metric and higher response accuracy within a shorter time for users to determine action-effect mappings in transparent user interface controls. We plan to expand the sample size in our empirical study.

Introduction

Electronic health record systems (EHRs) are the most critical form of IT penetration in the healthcare system. Being increasingly integrated across every aspect in healthcare delivery, EHRs are expected to be the powerful means to optimize quality of care. However, the National Research Council reported that current health IT applications provide little support for clinicians' cognitive tasks, increase the chance of error, and add to rather than reduce workload due to underutilization of human-computer interaction (HCI) principles [1]. Although usability, or human factors, has been widely recognized in the medical device industry since 1988 [2], it is currently a big challenge to improve the usability of complex EHRs.

Several differences between the medical device industry and the EHR industry shed light on practical solutions to the EHR usability challenge. First, unlike the former with human factors design process addressed in the American National Standard [3], AMIA Usability Task Force identified barriers in the EHR industry to adopt user-centered design process, such as different user types in one organization and significant process variations across different organizations [4].

Second, designing and evaluating EHRs become more complicated in organizations that are developing safer and more efficient processes. The uncertainty as to the fitness of EHRs for complex socio-technical systems coincides with the uncertainty of radical transformation in the organizations. A viewpoint from research scholars beyond the healthcare arena attempts to tackle such technology-and-institution co-evolving phenomena by engaging two fields --- IT research and organization studies --- in conversation [5].

Third, unlike ready-to-use medical devices delivered to the market, EHRs often require a significant amount of configuration made by local health IT personnel, especially for large healthcare systems. The functioning EHR that clinicians interact with is in fact the joint effort by the EHR vendor and the local IT department. Different from medical device manufacturers being held accountable for usability from early design phase [6] to post-market surveillance [7], it is yet not clear which entities should take the responsibility for improving EHR usability.

Last, unlike medical device manufacturers with established human factors programs and designated staff, healthcare organizations lack IT personnel with the expertise in usability engineering and usability evaluation. EHR analysts in healthcare organizations perform needs assessment, deliver ready-to-use EHRs, and directly support EHR use in

dynamic clinical and organizational contexts. However, only technical skills on EHR configuration and trouble-shooting in combination with a good understanding of clinical needs are currently required for this job role.

Due to the above usability challenges in the EHR world, practical tools that can be operationalized by IT personnel with minimum training in usability are in urgent demand. Healthcare organizations need to be the key players on assuring EHR quality use in dynamic contexts. Among numerous usability evaluation methods, analytical inspection is the efficient approach to screen for low-level predictable usability problems on end user interfaces (UIs). Since the scope of EHR configuration is often limited to textual labels of UI controls, cognitive ergonomics issues should be the first priority for healthcare organizations. Cognitive ergonomics focuses on the understanding of human cognitive abilities and limitations in the contexts of work in order to "improve cognitive work conditions and the overall performance of human-machine systems".[8] The goal of this research is to develop a set of reliable and valid inspection criteria for EHR analysts without systematic training in cognitive ergonomics.

Background

Usability is a multi-faceted concept in the quality model in ISO 25010 Software Quality Requirements and Evaluation (SQuaRE) and in the definitions published by research scholars [9, 10]. In spite of different opinions on usability components and measures, whether user interfaces (UIs) are easy for users to learn to use is one of the few in agreement. Clinicians highly desire EHRs that they can easily figure out how to use due to the time-pressured and interruptive work environment.

Analytical Inspection

Usability evaluation methods vary in the scope, theoretical basis, input, and output. Those that do not require user observation as the input during the evaluation are classified as analytical methods. Some are based on mathematical models such as GOMS model [11]; others are not and are generally referred to as *analytical inspection,* including heuristic evaluation [12], expert review, and cognitive walkthrough [13]. While saving a substantial amount of resources on collecting raw data from users and analyzing them, analytical inspection has limitations on the scope of usability issues being detected and evaluator effects [14].

In the effort of comparing different analytical inspection methods, a set of criteria based on the quantity of usability problems detected were proposed [15]. However, depending on the theoretical basis, these methods vary in the scope of targeted usability problems, and the concept of usability itself consists of multiple facets in very different nature. Simply using problem count as the indicator to compare inspection methods has been criticized [16].

A more helpful guidance for usability practitioners is to distinguish the scope of each method, that is, "what kinds of usability problems a method is and is not good for finding" [17]. Criteria based on the characteristics of the method, including reliability, validity, and downstream utility, are meaningful parameters for method evaluation. Reliability is the extent to which same input can yield same output regardless of the evaluator who applies the method, that is, there is very little evaluator effect [17]. Validity is "the extent to which the findings from analyses conform to those identified when the system is used in the 'real world'' [17]. Downstream utility describes the usefulness of the findings in informing redesign [17]. Downstream utility is especially important for analytical inspection methods because the ultimate goal of early-stage usability evaluation is design improvement.

Heuristic evaluation is a widely adopted inspection method. However, it relies on a limited set of loosely defined principles, which results in significant evaluator effect or low reliability across evaluators [14]. A method with low reliability means that the quality of its findings is highly inconsistent.

Cognitive walkthrough is an inspection method based on a cognitive model that guides evaluators through users' cognitive activities while performing tasks in "walk-up-and-use" applications (i.e. those that can be used with little training). In spite of the more structured evaluation procedure, usability problem detection is still subject to the evaluators' experience. Its focus on evaluating "the ease with which a user completes a task with minimal system knowledge" [18] makes it especially important to EHR usability.

The inspection criteria under development in this research are based on the same theoretical foundation as cognitive walkthrough and the goal is to improve the reliability and validity of usability inspection when the method of

cognitive walkthrough is employed. The inspection criteria are intended to provide guidance on identifying usability issues that make it difficult for users to figure out how to complete tasks, which will be further described in the following section.

Theoretical Foundation

To specify the different types of issues within this general scope and explain the theoretical basis of the proposed inspection tool, an overview of the resource model from the HCI field is provided here. The resource model explains how users figure out how to interact with UIs. The term *resource,* or task-critical information, is defined as the information that can be utilized by users to make interaction decisions [19]. The information on the UIs not only constrains what actions are possible but also helps users decide which action to take in order to accomplish their goals [20].

Users' interaction strategy varies depending on the available *resources* in the specific scenario [19]. *Plan following* strategy requires a pre-defined sequence of actions (i.e. *plan*), which is usually developed through systematic training. However, when there is adequate information on the UIs, users can figure out each action responsively without a pre-defined plan, which saves the mental cost of constructing and executing a plan. This less costly interaction strategy is named as *goal matching*. By matching the effect of a possible action with the current goal, users make interaction decisions without advance planning [19].

Table 1. Resource categories defined in the resource model [20]

Resource Category	Definition
Goal	A desired state of the world
Current state	A collection of attributes used to judge whether the state is closer to the goal
Plan	A pre-defined sequence of actions
Action possibility	Whether an action can be made or not
Action-effect mapping	Link between an action and its effect that will take place after the action is executed
Interaction history	A list of already-taken actions

The resource model reveals the association between *resource* category and interaction strategy. It provides a general guidance on how to determine what *resource* should be made readily available on the UIs in order to facilitate an interaction strategy. However, this theory fails to provide an operationalizable tool to determine whether or not task-critical information is effectively presented on user interfaces. If certain task-critical information is not effectively presented on the user interface, users have to retrieve that information from their memory, which incurs unnecessary cognitive effort.

As to EHRs, low cognitive effort of figuring out how to use is strongly desired because clinical documentation and business operations should not be an additional burden to clinicians' working memory which is already reaching its limit due to the high demand on productivity and safety in a work environment full of interruptions. According to the resource model, action-effect mapping is the key task-critical information that makes the less costly strategy *goal matching* feasible for users. Therefore, the analytical inspection criteria are developed particularly to provide guidance on how to determine whether or not action-effect mapping is effectively presented on user interfaces.

Methods

Preliminary Study

Initially, a metric for cognitive transparency was developed as a tool to guide the inspection of EHR user interfaces (UIs) for ease of use and ease of learning. As shown on Figure 1, the concept of cognitive transparency refers to clear representation of both operations and object-operation relations on UIs so that users can understand what will happen at the work domain level if they click on the UI controls. If the operation of the work that will be accomplished using the UI control is clear on the UI, the operation is considered as being externally represented according to the distributed cognition theory [21, 22]. If the operation is not clear on the UI and requires prior

knowledge in the user's mind, it is considered as being internally represented. Same applies to object-operation relation. When both of the operation and object-operation relation are externally represented on the UI, the UI control is considered as cognitively transparent. The concepts of operation, object, internal representation, and external representation will be illustrated using concrete examples given on Figure 2.

Metric for Cognitive Transparency		Representation of Object-Operation Relation	
		External (clear on the UI)	*Internal (memory required)*
Representation of Operation	*External*	**Transparent**	Not Transparent
	Internal	Not Transparent	Not Transparent

Figure 1. Initially proposed analytical approach: metric for cognitive transparency

On the top of Figure 2, the UI control with the textual label "Edit" is located next to a medication. Clicking "Edit" enables the user to edit that particular medication, where edit is the operation of the work that will be accomplished by clicking "Edit" and the particular medication is the object. In this case, both of the operation and object-operation relation are effectively presented on the UI, or externally represented. The UI control is cognitively transparent. On the bottom of Figure 2, the effect after clicking the UI control with the label of a medication name is not presented at all on the UI. It requires the user to have prior knowledge about the actual effect of this action. In this case, the object is the medication which seems clickable; however, the operation of the work that will be accomplished by clicking the medication name is internally represented. Thus this UI control is not cognitively transparent.

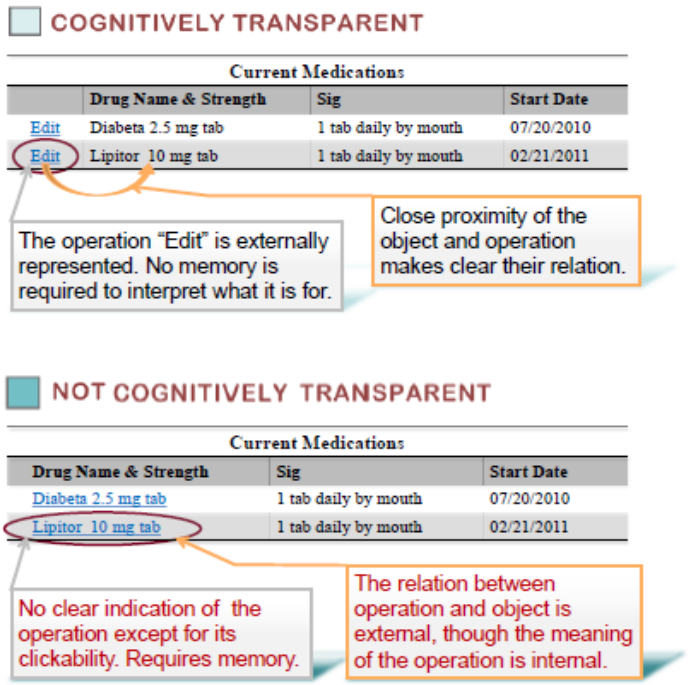


Figure 2. Examples of clickable user interface elements

The validity of this metric was tested in this preliminary study. One evaluator applied the metric to a set of clickable UI controls to predict if they are cognitively transparent or not. The UI controls were selected from E-Prescribing Use Case in three ambulatory EHRs. Three participants with general computer experience were recruited. With the original EHR screenshots presented, participants were asked to anticipate work domain effect of clicking the

selected UI controls, that is, "what do you think will happen in terms of the work that you are doing if you click this". Participants' anticipations were compared to the actual work domain effect of clicking those controls in the live EHRs. If the anticipations from all of the three participants match the actual effect, the UI control was empirically classified as cognitively transparent. If the anticipation from at least one participant does not match the actual effect, the UI control was empirically classified as not cognitively transparent. Subsequently, the predictions made using the metric were compared with the classification based on participants' anticipations. The specificity and sensitivity of the metric were analyzed.

Exploratory Experiment

Based on this initial metric, a set of inspection criteria were developed for health IT personnel with minimum training in cognitive ergonomics to identify issues of presenting action-effect mapping on UIs. According to the resources model, UIs where action-effect mapping is effectively presented are hypothesized to be easier for users to correctly match their goals with possible actions on the UIs.

The exploration for the validity of the inspection criteria on action-effect mapping began from UI controls that enable initiation of entry tasks. The proposed inspection criteria are explained in Table 2. Based on the proximity compatibility principle, the object can be directly presented on the label of the UI control or indirectly presented via immediate neighbor element(s). Published studies have demonstrated that unlabeled icons are significantly more difficult for novice users in all ages to interpret than textual labels [23]. Therefore, only texts or well-recognized symbols are considered as effective presentation.

An exploratory study was conducted to test whether user performance differs between EHR UI controls with effective presentation of action-effect mapping versus those without. Due to lack of published information on UI factors that affect human performance in relevant experiments, screenshots from several EHRs with multiple variations were used in this exploratory study.

Table 2. Proposed inspection criteria for EHR UI controls that initiate entry tasks

Effective Presentation*	**Location**	**Form**
Operation	On the same UI control	Text or symbol (e.g. "+" = add)
Object**	On the same UI control	Text
	Immediate neighbor	Text (usually as a header)
		View-only display of instances (e.g. a list of medications)

*A UI control is considered to be transparent only if both the operation and the object involved are effectively presented.
**The object is considered to be effectively presented as long as it falls in any of the three situations.

- *Stimuli*

A total of thirteen screenshots were extracted from four commercial EHRs. The work domain effects of selected UI controls were initiating prescription entry, initiating problem entry, initiating allergy entry, and initiating lab order entry. These screenshots were presented randomly to each participant via online first-click testing software Chalkmark (Optimal Product, New Zealand).

- *Participants and Procedures*

Four physicians with no prior exposure to these particular EHRs were recruited. During the experiment, they were asked to click on the presented screenshot to achieve a given goal. Prior to each screenshot being presented, one goal was indicated on the prompt, which corresponded to the work domain effect of the target UI control. An example prompt is "you would like to prescribe a medication (or order a laboratory test, etc.) to this patient. Where would you click to begin?". The location of their mouse click on each screenshot was collected together with the duration from the screenshot being presented to mouse click (i.e. response time).

- *Confounding Factors*

The variation in the quantity of textual words and pictorial icons on these screenshots was hypothesized to affect participant performance in addition to the independent variable in this exploratory study. The word count and icon count on each screenshot were analyzed with test patient data excluded.

Results

Preliminary Study

The percentages of UI controls with metric-derived prediction matching the empirical classification are presented on Figure 3. In EHR 1, among 35 target UI elements, 60% were predicted by the metric as transparent and had all of the three subjects' anticipations matching the work domain effect in the live EHR (i.e. empirically classified as transparent); 31% were predicted as not transparent and had at least one subject's anticipation not matching the work domain effect in the live EHR (i.e. empirically classified as not transparent); 6% were predicted as not transparent but empirically classified as transparent; 3% were predicted as transparent but empirically classified as not transparent. In EHR 2, among 17 target UI elements, 70% were predicted as transparent and empirically classified as transparent; 12% were predicted as not transparent and empirically classified as not transparent. In EHR 3, among 18 target UI elements, 50% were predicted as transparent and empirically classified as transparent; 39% were predicted as not transparent and empirically classified as not transparent.

		Observation from Experiment	
EHR 1 (total 35 UI elements)		Transparent (agreed by 3 subjects)	Not Transparent (at least 1 subject)
Prediction from Metric	Transparent	**60%**	3%
	Not Transparent	6%	**31%**
EHR 2 (total 17 UI elements)		**70%**	12%
		6%	**12%**
EHR 3 (total 18 UI elements)		**50%**	11%
		0%	**39%**

Figure 3. Proportions of UI controls with metric-derived prediction matching or not matching empirical classification

The specificity and sensitivity of the metric in each EHR are presented on Table 3. This preliminary study suggests that the cognitive transparency metric has high sensitivity and acceptable specificity.

Table 3. Specificity and sensitivity of metric for cognitive transparency

	Sensitivity	Specificity
EHR 1	91%	92%
EHR 2	92%	50%
EHR 3	100%	78%

Exploratory Experiment

Accuracy of response was determined by comparing the location of a participant's mouse click on the screenshot with the location of the target UI control. The small sample size (i.e. number of screenshots) in this exploratory

study makes it difficult to detect any statistical significance from the results. Basic data analysis was carried out to look for any potential difference in the two measures (i.e. accuracy of response and response time) between two groups of UI controls.

As shown in Table 4, the average accuracy of response for the group of UI controls with action-effect mapping effectively presented was 63%, whereas that for the group with action-effect mapping ineffectively presented was 33%. The average response time for the former group was 19 seconds, whereas that for the latter group was 28 seconds.

Table 4. Average accuracy of response and average response time for UI controls with effective or ineffective presentation of action-effect mapping

Presentation of Action-Effect Mapping	Number of Stimuli	Participant Performance		Average Word Count (Icon)
		Average Accuracy of Response	Average Response Time (seconds)	
Effective	10	63%	19	80 (30)
Ineffective	3	33%	28	75 (39)

UI controls with action-effect mapping effectively presented according to the proposed inspection criteria appeared to have a higher average accuracy of response with a shorter average response time. The average word count and average icon count were similar among the two groups of UI controls, suggesting that the quantity of words or icons may not be a contributing factor to the difference in participant performance.

Discussion

The EHR industry urgently needs HCI theories to be translated into inspection tools that can be operationalized by health IT personnel with minimum training in cognitive ergonomics, which empowers healthcare organizations to improve EHR usability and thereby assure EHR quality use in dynamic operational and clinical contexts.

Among different components of usability, clinicians strongly desire EHRs that are easy to figure out how to use without systematic training and memorization. Operating EHRs at the point of care should incur little unnecessary cognitive effort and clinicians' limited mental capacity should be focused on patient care. The theory-based criteria under development in this research are targeted at supporting analytical inspection of EHR UIs from this aspect.

Analytical inspection is the efficient approach to detecting low-level predictable usability problems because it does not require collection of raw data from user observation. The quality of findings relies on the theory basis of the inspection method or tool. Healthcare organizations can benefit the most from the HCI field by employing validated inspection tools in the EHR configuration process.

However, there is little agreement on how to empirically validate an inspection tool. The preliminary study and the exploratory experiment reported above are validation attempts using different research design. In the preliminary study, the target UI controls was pointed out on the screenshots presented to participants, and the raw data collected from participants were their anticipations of work domain effects after clicking those UI controls in live EHRs. In contrast, the exploratory experiment presented screenshots to participants without pointing out the target UI controls, but with specified goals that corresponded to the work domain effects of the target UI controls. The participants' choices of UI controls for the given goals were the collected empirical data. The research design in the exploratory experiment well simulated user interactions in the real world, compared to the preliminary study, thus will be adopted in future research on validating other proposed inspection tools.

Although objective performance measures, such as the location of clicking, are more robust indicators of the effectiveness of use than subjective measures, such as confidence rating, it is still important to include subjective measures in future research to understand participants' subjective perception with regard to the clarity of UIs. Other

UI factors that are different across EHRs and may affect human performance were not controlled in the exploratory experiment. A controlled experiment will be conducted in the future to establish the causal effect relationship between the effectiveness of action-effect mapping presented on UIs and participants' accuracy of response. Meanwhile, it is critical to expand the sample size so that statistical significance can be tested.

In addition to validity, reliability is another key criterion to evaluate inspection tools. Future research will test the reliability of the proposed inspection tools. Multiple evaluators with various levels of experience in usability will independently apply the inspection tools to detecting ineffective presentation of action-effect mapping on EHR UIs. Their detection results will be compared and analyzed statistically.

Acknowledgement
This project was supported by Grant No. 10510592 for Patient-Centered Cognitive Support under the Strategic Health IT Advanced Research Projects (SHARP) from the Office of the National Coordinator for Health Information Technology.

References

1. Stead WW, Lin HS, eds. Computational technology for effective health care: immediate steps and strategic directions. Washington, DC: National Academies Press. 2009.
2. Arnaut LY, Greenstein JS. Human factors considerations in the design and selection of computer input devices. In S. Sherr (Ed.), Input Devices 1988;71-121. San Diego, CA: Academic Press.
3. Human factors design process for medical devices. HE74:2001. ANSI/AAMI.
4. Middleton B, Bloomrosen M, et al. Enhancing patient safety and quality of care by improving the usability of electronic health record systems: recommendations from AMIA. J Am Med Inform Assoc 2013;20:e2-e8.
5. Orlikowski WJ, Barley SR. Technology and institutions: what can research on information technology and research on organizations learn from each other? MIS Quarterly. 2001;25:145-165.
6. U.S. Food and Drug Administration. Premarket Information - Device Design and Documentation Processes. http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/HumanFactors/ucm119190.htm.
7. U.S. Food and Drug Administration. Postmarket Information - Device Surveillance and Reporting Processes. http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/HumanFactors/ucm124851.htm.
8. Hoc JM. Towards ecological validity of research in cognitive ergonomics. Theoretical issues in ergonomics science 2001;2(3):278-288.
9. Nielsen J. Usability Engineering. AP Professional. 1993. New York.
10. Zhang JJ, Walji MF. TURF: Toward a unified framework of EHR usability. J Biomedical Informatics. 2011; 44(6):1056-1067.
11. Card SK, Moran TP, Newell A. Computer text-editing: an information-processing analysis of a routine cognitive skill. Cognitive Psychology. 1980;12 (1):32-74.
12. Nielsen J. Heuristic Evaluation. In J. Nielsen & R. Mack (Eds.), Usability Inspection Methods. New York: John Wiley. 1994:25-62.
13. Wharton C, Rieman J, Lewis C, Polson P. The cognitive walkthrough method: A practitioner's guide. In J. Nielsen & R. Mack (Eds.), Usability inspection methods. New York: John Wiley. 1994:105-140.
14. Hertzum M, Jacobsen NE. The evaluator effect: a chilling fact about usability evaluation methods. Int. J. Human-computer Interaction. 2001;13(4):421-43.
15. Hartson HR, Andre TS, Williges RC. Criteria for evaluating usability evaluation methods. Int. J. Human-computer Interaction. 2003;15(1):145-181.
16. Gray WD, Salzman MC. Damaged merchandise? A review of experiments that compare usability evaluation methods. Human Computer Interaction Journal. 1998; 203-261.
17. Blandford AE, Hyde JK, Connell I, Green TRG. Scoping analytical usability evaluation methods: a case study. Human Computer Interaction Journal. 2008;23(3):278-327.

18. Mahatody T, Sagar M, Kolski C. State of the art on the cognitive walkthrough method, its variants and evolutions. Intl. J. Human-Computer Interaction. 2010; 26(8): 741-785.
19. Wright P, Fields B, Harrison M. Analyzing human-computer interaction as distributed cognition: the resource model. Human-computer Interaction. 2000;(15):1-41.
20. Fields B, Wright P, Harrison M. Designing human-system interaction using the resource model. APCHI'96: First Asia Pacific Conference on Human Computer Interaction. 1996;181-191.
21. Zhang JJ, Norman DA. Representations in distributed cognitive tasks. Cognitive Science. 1994;18:87-122.
22. Zhang JJ, Patel VL. Distributed cognition, representation, and affordance. Pragmatics & Cognition. 2006;14(2): 333-341.
23. Wiedenbeck S. The use of icons and labels in an end user application program: an empirical study of learning and retention. Behavior and Information Technology. 1999;18(2): 68-82.

Automatically Detecting Acute Myocardial Infarction Events from EHR Text: A Preliminary Study

[*]Jiaping Zheng, MS[1], [*]Jorge Yarzebski, MD, MPH[2], Balaji Polepalli Ramesh, PhD[2], [**]Robert J. Goldberg, PhD[2], [**]Hong Yu PhD[1,2]

[1]University of Massachusetts, Amherst, MA

[2]University of Massachusetts Medical School, Worcester, MA

***author contributed equally; **corresponding author**

Abstract

The Worcester Heart Attack Study (WHAS) is a population-based surveillance project examining trends in the incidence, in-hospital, and long-term survival rates of acute myocardial infarction (AMI) among residents of central Massachusetts. It provides insights into various aspects of AMI. Much of the data has been assessed manually. We are developing supervised machine learning approaches to automate this process. Since the existing WHAS data cannot be used directly for an automated system, we first annotated the AMI information in electronic health records (EHR). With strict inter-annotator agreement over 0.74 and un-strict agreement over 0.9 of Cohen's κ, we annotated 105 EHR discharge summaries (135k tokens). Subsequently, we applied the state-of-the-art supervised machine-learning model, Conditional Random Fields (CRFs) for AMI detection. We explored different approaches to overcome the data sparseness challenge and our results showed that cluster-based word features achieved the highest performance.

Introduction

WHAS[1,2] is an ongoing population-based investigation examining changing trends in the incidence rates, hospital and post discharge death rates, occurrence of major clinical complications, and use of different management approaches in residents hospitalized with independently validated AMI at all metropolitan Worcester hospitals in Massachusetts. It has been used to study comparative change in attack and survival rates of AMI, the impact of age on the incidence and prognosis of initial AMI, etc, and has been contributing significantly to cardiovascular diseases population studies for decades. Various patient information has been collected, including demographics, medical history symptoms, laboratory and physiologic measures, medications, diagnostic procedures, coronary interventions, hospital length of stay, and hospital survival status.

Despite the large amount of data manually collected in the WHAS study, the data is not geared towards automatic information extraction from the clinical narratives. The assessment starts with the medical records of patients hospitalized with possible AMI in the dataset where clinicians manually review, validate, and extract information according to pre-defined diagnostic criteria. The corresponding text references in the medical records (or the annotation) were not recorded.

In this study, we report the preliminary development of supervised machine learning models for AMI information extraction from EHR DSs. We developed annotation guidelines and reported inter-annotator agreement. With an annotated EHR DSs data, we developed natural language processing (NLP) systems for automation of case validation of, and data abstraction from, EHR of patients hospitalized with AMI.

Our main goal is to develop and evaluate a machine-learning based NLP system to automatically extract AMI related information to facilitate AMI manual review. Our major contributions are the following:

a. We built an EHR-based guideline for annotating diagnostically relevant AMI variables including acute symptoms and electrocardiographic (ECG) and laboratory findings.

b. Although EHR annotation is not new, we are the first group to report the annotation and excellent inter-annotator agreement of AMI information using EHR DSs.

c. We are the first group to report automated AMI detection from EHR DSs. In our supervised machine learning approaches, we explored cluster-based learning features, which have not been widely used in clinical NLP tasks, and found that the features improved the performance.

Related Work

Concept identification has been a key NLP task since its inception. In the biomedical domain, many NLP systems have been developed to extract concepts based on the rich resources provided by the Unified Medical Language System (UMLS)[3] from unstructured clinical narratives. MetaMap performs lexical and syntactic analysis on input text and phrases are mapped to the UMLS concepts[4]. MedEx recognized medication information, such as drug name, dose, frequency, route, and duration[5]. cTAKES analyzes clinical free text using models specifically trained for the clinical domain and maps phrases to the UMLS concepts with additional attributes such as negation, uncertainty, and conditional, etc[6]. KnowledgeMap is a collection of NLP tools for clinical text[7], one of which is the Concept Index tool which identifies biomedical concepts and maps them to the UMLS. ARC aims at building a generic platform of information retrieval without custom code or rule development by the end user. It automates feature (including concepts extracted by cTAKES) and algorithm selection using existing medical NLP tools[8].

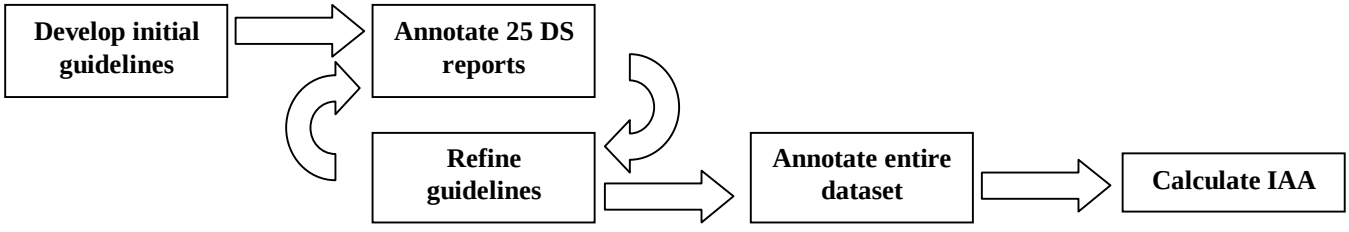
The Center of Informatics for Integrating Biology and the Bedside (i2b2) organized challenges to extract clinical entities, including medication[9] and problems, tests, and treatments[10] from EHR narratives. In the 2010 challenge, it has shown that supervised machine learning systems performed well for extracting medical concepts (problems, treatments, and tests) from EHR narratives, from which CRF models[11] stand out being the most successful models[12–14]. Features commonly used in these models include lexical and morphological features (word form, prefix/suffix), grammatical features (parts of speech), and semantic features (concepts extracted using rule-based approaches). Ensemble methods that combine multiple CRF-based and other rule-based models also showed success in this task[15].

Rule-based approaches are fast to deploy. However, for a focused domain such as AMI, a general-purpose tool would extract much more information than necessary, thus requiring careful filtering or building and tuning the rules. Therefore, we designed a supervised machine learning based system. Annotated data is essential for supervised machine learning. Although the biomedical annotation efforts have been deep and wide,[16] we are not aware of studies that annotate named entities for AMI. With the approval from the Institutional Review Board of the Medical School of the University of Massachusetts, we conducted annotation and NLP development for automatically recognizing AMI information.

Material and Methods

Pre-defined diagnostic criteria are key information in the WHAS database. These criteria include a clinical history of prolonged chest pain not relieved by rest or use of nitrates, serum levels of various biomarkers in excess of the upper limit of normal as specified by the laboratory at each greater-Worcester area hospital, and serial electrocardiographic tracings during hospitalization showing changes in the ST segment and Q waves typical of AMI. At least two of these three criteria need to be satisfied for study inclusion. Cases of perioperative-associated AMI are not included in the study sample. However, these criteria appear mainly in the narrative text of discharge summaries (DS), not in the structured data. Since there is no annotation at the word level in the discharge summaries, we first developed the annotation guidelines. The discharge summaries are then annotated by two annotators according to these guidelines. This process is illustrated in Figure 1. In the following we first describe the annotation guidelines and then report inter-annotation agreement, and our NLP approaches.

Figure 1. Illustration of the annotation process.



Data

We obtained a set of 105 discharge summary reports by filtering the reports that contain the ICD-9 diagnostic rubric 410 (AMI) as the primary or secondary discharge diagnosis. Five types of entities are annotated in these reports: symptoms, ECG findings, lab observations, ICD diagnosis, and Catheterization lab findings. These entity types are

validation criteria to rule in patients as AMI cases, which were developed by the World Health Organization[1], and more recently, by the Third Global MI Task Force[17].

Annotation Guidelines

The annotation guidelines were developed by a physician, a linguist, and three biomedical informaticians through an iterative process. Following the initial version of the annotation guidelines, two annotators first independently annotated 25 DS reports, and then resolved the disagreements. The guidelines were revised to reflect the discussions during the consensus session. Table 1 shows the annotation guidelines for the five entity types.

Annotation Process

An expert physician designated as *AnnPhy*, annotated 105 DS reports, including the 25 reports used to develop the guidelines. A linguist, designated as *AnnLing*, also annotated the same 25 reports. The inter annotator agreement was calculated using the 25 reported independently annotated. The annotated corpus was used to build machine-learning models to identify the AMI information. There were a total of around 135K word tokens in the corpus of 105 DS reports with an average of 1285.6 ± 440.5 word tokens per report. The subset of 25 reports consists of a total of 29.5K word tokens with an average of 1178.2 ± 377.0 word tokens per report. We report the Cohen's κ[18] to calculate the IAA between annotators.

Table 1. Definitions and examples of annotation categories.

Entity type	Definition	Select examples
Symptoms	Presence of symptoms of myocardial ischemia	Abdominal pain, chest pain, nausea, vomiting, shortness of breath
ECG Findings	Acute or evolving changes in the ST-T wave forms and Q waves, Bundle branch block	Q wave MI, T wave inversions, New LBBB
Lab Observations	Biomarker/cardiac enzyme detection of myocardial injury with necrosis	Cardiac enzymes, Elevated troponins
ICD diagnosis	ICD9 discharge diagnosis	410 Acute myocardial infarction, 410.00 episode of care unspecified
Catheterization Lab Finding	Description/Mention of AMI	Location of AMI (e.g., anterior, inferior, posterior, lateral)

Supervised Machine Learning Approaches

We developed a supervised CRF models to identify the clinical entities as shown in Table 1. CRFs are widely used in various NLP tasks, and have been shown to be among the best models for NER[10]. CRFs predict entity types from a sequence of input word tokens by optimizing for the conditional probability of the label given the observed data.

We used the ABNER[19] package as our CRF implementation. The features in this implementation include word token, word type, punctuation, capitalization, prefix, suffix, roman letters, and features from the previous and following words. We trained our systems to predict the symptoms, ECG findings, Lab observations, and ICD diagnosis entities. Catheterization lab finding is excluded because it is very rare in our dataset.

Our basic CRF implementation only utilizes lexical and morphological features. To explore the effects of various features on the data, we incorporated both syntactic (part of speech tags) and semantic (word representation) features into this basic CRF model as additional features. The parts of speech are obtained using the maximum entropy based classifier from OpenNLP with a model trained with both general English and clinical text[6] (including GENIA, Penn Treebank and anonymized medical records).

Each word in a corpus is conventionally represented as one dimension in the feature vector. Thus, the feature vector has the same length as the vocabulary size. Parameters for the rare words are poorly estimated from the data. It can not handle new words in the test data either. Word representation features can overcome this data sparseness problem. We explored two types of word representation features. The Brown cluster-based approach represents words in a hierarchical cluster, which maximizes the mutual information of bigrams. We learned Brown clusters from a collection of approximately 100,000 clinical notes (46 million tokens), including discharge summaries,

cardiology reports, emergency notes, history and physical exams, operative reports, progress notes, radiology notes, and surgical pathology notes. To compare with word representations induced from general English text, we also trained a CRF model with Brown clusters from the RCV1 news corpus[20,21] (approximately 810,000 news stories) provided by Turian et al[22].

The other approach to overcome data sparseness is to induce word embeddings. This approach represents a word as a dense low-dimensional vector. These dimensions capture latent features of the words. Word representation features have been shown to improve performance in NLP tasks including chunking and entity recognition. In our system, we incorporated a word embeddings representation included in SENNA[23] that was learned from English Wikipedia text.

Results

Corpus Characteristics and Annotator Agreement

Table 2 below shows the characteristics of the data: entity types, the number of entities annotated in each category and the inter annotator agreement in terms of Cohen's κ. We report both strict κ, where there is an exact match between the entities annotated and un-strict κ, where there is at least one word token overlap between the annotated entities. *ECGFindings* had the highest strict κ value of 0.86 followed by *LabObservation* with κ value of 0.84. Whereas, *LabObservation* achieved the highest un-strict κ value of 0.98 followed by *ECGFindings* with a κ value of 0.95 for un-strict criteria. Overall, all the entities except for *CatheterizationLabFinding* achieved a strict κ of over 0.74 and an un-strict κ of over 0.9. Since *AnnPhy* annotated only one entity in *CatheterizationLabFinding* category, the κ value was low for *CatheterizationLabFinding*. Both strict and un-strict F1 scores are reported as well.

Table 2. Annotated named entities, number of instances, and inter-annotator agreement.

Entities	AnnPhy		AnnLing	κ (strict)	κ (un-strict)	F1 (strict)	F1 (un-strict)
	Number of entities		Number of entities on 25 reports				
	25 reports	105 reports					
Symptom	153	651	244	0.74	0.93	0.36	0.71
ECG Findings	75	368	64	0.86	0.96	0.68	0.83
Lab Observations	105	370	66	0.84	0.98	0.20	0.61
ICD diagnosis	15	60	41	0.75	0.96	0.18	0.44
Catheterization Lab Finding	1	1	3	0.13	0.76	0.00	0.50

Experiment Results

We report recall, precision, and the F1 score using 10-fold cross-validation. Table 3, system setting b shows the performance of our basic CRF model using only the lexical and morphological features from ABNER. As a baseline, we implemented a rule-based method that matches terms defined in the annotation guidelines. Since only AMI related entities are considered, a general dictionary lookup system such as MetaMap will produce a large number of irrelevant entities---only 12% of the entities of type symptom, findings, lab results, and lab procedures from MetaMap appear in the annotation guidelines. The results of our baseline are shown in Table 3, system setting a. The results show that the rule-based dictionary-lookup has the worst performance in all five categories. The overall F1 score using the dictionary lookup was 29.74%, which is significantly lower than 62.90% F1 score of a CRF model trained on lexical and morphological features. The most noticeable difference is in the Lab observations category. There is an almost 60% absolute difference between the dictionary matching and the machine learning approaches.

As shown in Table 3, system setting c, adding syntactic features slightly improves the overall performance by 0.3%. There are minor decreases in two categories (symptom and ECG findings), however the improvements in the Lab observations category outweigh the decreases. The ICD diagnosis categories remain largely unchanged. These

results are consistent with the nature of our annotations. Lab observations often contain phrases such as "troponin bumped to 26 from 0.21". Part of speech features can capture the shallow syntactic structure of these phrases, thus enabling the CRF model to learn from them. On the other hand, ICD diagnosis is more homogeneous, and usually follows a predictable pattern, for example, a number followed by "acute myocardial infarction". Little variation in the syntactic structures does not leave much additional information for the POS tagger to provide.

Contributions from semantic features can be seen in Table 3, system settings d to f. We induced Brown clusters from the the clinical text corpus, using a minimum word frequency of one, and two difference sizes (100 and 1000). Prefixes of length 4, 6, 10 and 20 were incorporated into the basic CRF model. These word representations outperformed the basic system, and achieved similar or better results compared to the systems trained on the syntactic features. Both cluster sizes consistently outperformed our basic CRF model across all entity categories. Using a large number of clusters resulted in a 3.10 percentage gain in overall F1 measure. However, the trend was reversed when word embeddings representation was incorporated into the basic CRF model. This representation was learned from Wikipedia text. Our system performance suffered from a decrease in the symptoms category, which accounts for 45% of the total entities in our dataset. This resulted in a lower overall performance. One reason the word embeddings proved not beneficial is that only 38.7% of the entity tokens in our corpus appear in the word embeddings. Moreover, some medical abbreviations could have different meanings in general text. For example, "abd" is an abbreviation for abdomen or abdominal. However, in Wikipedia, the disambiguation page for this abbreviation links to over 10 pages, none of which is abdomen. Therefore, the low-dimensional vector learned for this word would not be able to capture its meaning in a medical context.

Table 3. System performance

System Setting		**Symptom**	**ECG Findings**	**Lab Observations**	**ICD Diagnosis**	**All**
a. Dictionary Lookup Baseline	Precision	22.71	46.13	19.38	100.00	28.65
	Recall	30.25	44.11	16.62	47.46	30.91
	F1	25.94	45.10	17.89	64.37	29.74
b. Basic CRF model	Precision	53.73	78.31	86.41	87.98	69.77
	Recall	42.80	72.10	73.17	81.67	58.30
	F1	47.00	74.20	77.70	83.90	62.90
c. Basic CRF + POS	Precision	53.48	77.75	87.54	87.98	70.11
	Recall	41.63	71.48	76.32	81.67	58.59
	F1	46.10	73.80	80.20	83.90	63.20
d. Basic CRF + 100 Brown clusters	Precision	55.12	77.65	85.55	89.31	69.81
	Recall	44.35	71.99	71.36	81.67	58.57
	F1	48.60	73.90	76.80	84.50	63.10
e. Basic CRF + 1000 Brown clusters	Precision	59.29	79.17	90.14	89.60	73.16
	Recall	49.22	73.57	72.19	84.92	61.58
	F1	53.00	75.44	79.00	86.44	66.00
f. Basic CRF + Wikipedia Word Embeddings	Precision	48.21	75.00	90.00	100.00	65.18
	Recall	40.91	70.59	72.00	100.00	56.59
	F1	44.00	72.00	80.00	100.00	60.00

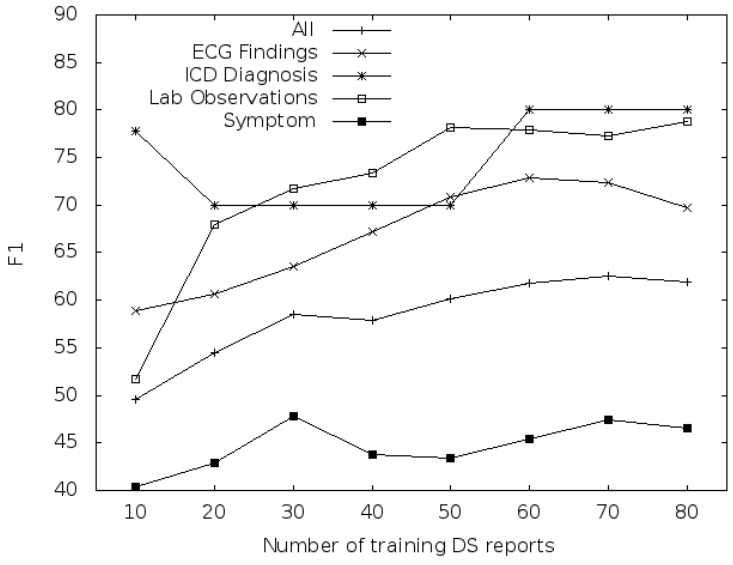
We have shown that semantic features obtained from general English corpora such as Wikipedia provides little benefit, if any, to our system. To further investigate the effects of using general domain word representation, we incorporated Brown clusters induced from newswire text. The performance from these clusters shows no improvement or a decrease over our basic CRF model. The results are listed in Table 4.

Table 4. System performance using Brown clusters induced from newswire text.

System Setting		**Symptom**	**ECG Findings**	**Lab Observations**	**ICD Diagnosis**	**All**
Basic CRF + 100 Brown clusters (newswire text)	Precision	52.47	79.02	85.18	89.31	69.14
	Recall	43.12	72.35	74.51	81.67	58.74
	F1	46.70	74.80	77.80	84.50	62.90
Basic CRF + 1000 Brown clusters (newswire text)	Precision	52.58	78.43	85.25	88.81	69.04
	Recall	41.97	71.23	72.77	80.00	57.66
	F1	46.10	73.80	77.00	83.30	62.10

Figure 2 shows the system F-1 scores using the best performing setting (basic CRF with 1000 Brown clusters induced from clinical text) with different sizes of training data. 20% of data is reserved for testing. Due to the small amount of ICD Diagnosis annotations, the performance varies more than the other categories.

Figure 2. F1 score of best performing setting with varying size of training data.



Discussions

Despite the lower inter annotator agreement (IAA) on the initial 25 DS reports, ICD diagnosis is consistently the highest performing category among all the entities we annotated. This could be explained by the fact that these entities exhibit a particular pattern, a number followed by one of a set of AMI related phrases, in capital letters. Lexical and morphological feature in our basic CRF model already captures such patterns. The low IAA is due to the linguist annotator annotating other occurrences of these AMI phrases as ICD diagnosis, or including more words.

The lab observations category sees the highest improvement with the addition of the part of speech features. Both precision and recall increased over the basic CRF model, suggesting the model benefited from the syntactic structures. Semantic features improved the precision score, at the same time decreased the recall score.

Errors of the other two types of entities (symptom and ECG findings) mainly stemmed from false negatives and false positives. There are in total 157 cases of false positives and 318 cases of false negatives. One reason for the false positives is that the model learned from the modifiers of the annotations but labeled the modifiers only, as in "not radiate to jaw or left arm". The most frequent false negatives errors are due to conjunctions or lists of symptoms, such as "chest pain, SOB, Diziness, Syncope, Palpitation or dizziness". Another reason is sentence-like entities. For example, "did not have a markedly elevated cardiac enzymes", and "trended his troponins and they peaked at 35". Miscategorization only accounts for a small percentage of the errors. In our basic CRF model, only 17 tokens, or 7 entities, were classified into a wrong category. One example is where the ECG findings entity "No ST elevations or depressions" was recognized as a symptom entity.

Another source of errors is in-exact boundary detections. There are 161 entities that match one boundary of the human annotations. Many of these boundary mismatches are due to the modifiers. For instance, our system recognized "acute onset of chest pressure" as a symptom entity whereas the gold standard only contains "chest pressure". This is consistent with the large discrepancy between the strict and non-strict inter-annotator agreements on the symptoms category.

Conclusion and Future Work

Worcester Heart Attack Study has been used to study comparative change in attack and survival rates of Acute Myocardial Infarction. However, the EHR of patients hospitalized with possible AMI in the WHAS dataset are manually reviewed and validated. This process is time-consuming and cumbersome. An automatic system that can extract key information from the EHR can speed up this manual process.

In this study, we evaluated CRF models on clinical entity recognition, specifically AMI related concepts that are used to validate patient's AMI status. We investigated the contributions of syntactic and semantic features. We demonstrated in our experiments that both types of features could improve the overall system performance. The F1 measure achieved a 3% to 3.10% increase when the features are trained from EHR, whereas features induced from general English text (newswire and Wikipedia) were not beneficial. The models are available upon contacting the authors due to IRB concerns.

There are many avenues to explore in future work. We will conduct a more thorough evaluation of annotation consistency. We will investigate contributions of word embeddings induced from clinical text, and compare with those induced from general English text. As discussed in the previous section, annotations with modifiers pose a challenge to learn an effective model. We will study how syntactic structures deeper than parts of speech can improve system performance. We will also explore the performance of these models in the context of the overall validation process.

References

1. Floyd, K. C. *et al.* A 30-year perspective (1975-2005) into the changing landscape of patients hospitalized with initial acute myocardial infarction: Worcester Heart Attack Study. *Circ. Cardiovasc. Qual. Outcomes* **2,** 88–95 (2009).
2. Goldberg RJ, Gore JM, Alpert JS & Dalen JE. Recent changes in attack and survival rates of acute myocardial infarction (1975 through 1981): The worcester heart attack study. *JAMA* **255,** 2774–2779 (1986).
3. Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* **32,** D267–270 (2004).
4. Aronson, A. R. & Lang, F.-M. An overview of MetaMap: historical perspective and recent advances. *J. Am. Med. Inform. Assoc. JAMIA* **17,** 229–236 (2010).
5. Xu, H. *et al.* MedEx: a medication information extraction system for clinical narratives. *J. Am. Med. Inform. Assoc. JAMIA* **17,** 19–24 (2010).
6. Savova, G. K. *et al.* Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J. Am. Med. Inform. Assoc. JAMIA* **17,** 507–513 (2010).

7. Denny, J. C., Irani, P. R., Wehbe, F. H., Smithers, J. D. & Spickard, A., 3rd. The KnowledgeMap project: development of a concept-based medical school curriculum database. *AMIA Annu. Symp. Proc. AMIA Symp. AMIA Symp.* 195–199 (2003).
8. D'Avolio, L. W., Nguyen, T. M., Goryachev, S. & Fiore, L. D. Automated concept-level information extraction to reduce the need for custom software and rules development. *J. Am. Med. Inform. Assoc. JAMIA* **18,** 607–613 (2011).
9. Uzuner, O., Solti, I. & Cadag, E. Extracting medication information from clinical text. *J. Am. Med. Inform. Assoc. JAMIA* **17,** 514–518 (2010).
10. Uzuner, Ö., South, B. R., Shen, S. & DuVall, S. L. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J. Am. Med. Inform. Assoc.* amiajnl–2011–000203 (2011). doi:10.1136/amiajnl-2011-000203
11. Lafferty, J. D., McCallum, A. & Pereira, F. C. N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. in *Proc. Eighteenth Int. Conf. Mach. Learn.* 282–289 (Morgan Kaufmann Publishers Inc., 2001). at <http://dl.acm.org/citation.cfm?id=645530.655813>
12. Jiang, M., Chen, Y. & Liu, M. Hybrid approaches to concept extraction and assertion classification - Vanderbilt's systems for 2010 i2b2 NLP Challenge. in *2010 I2b2VA Workshop Chall. Nat. Lang. Process. Clin. Data* (2010).
13. Gurulingappa, H., Hofmann-Apitius, M. & Fluck, J. Concept identification and assertion classification in patient health records. in *2010 I2b2VA Workshop Chall. Nat. Lang. Process. Clin. Data* (2010).
14. Kang, N., Barendse, R. & Afzal, Z. Erasmus MC approaches to the i2b2 Challenge. in *2010 I2b2VA Workshop Chall. Nat. Lang. Process. Clin. Data* (2010).
15. Kang, N., Afzal, Z., Singh, B., van Mulligen, E. M. & Kors, J. A. Using an ensemble system to improve concept extraction from clinical records. *J. Biomed. Inform.* **45,** 423–428 (2012).
16. Bada, M. *et al.* Concept annotation in the CRAFT corpus. *BMC Bioinformatics* **13,** 161 (2012).
17. Thygesen, K. *et al.* Third Universal Definition of Myocardial Infarction. *Circulation* **126,** 2020–2035 (2012).
18. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* **20,** 37–46 (1960).
19. Settles, B. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics* **21,** 3191–3192 (2005).
20. Rose, T., Stevenson, M. & Whitehead, M. The reuters corpus volume 1 - from yesterday's news to tomorrow's language resources. in *Proc. Third Int. Conf. Lang. Resour. Eval.* 29–31 (2002).
21. Lewis, D. D., Yang, Y., Rose, T. G. & Li, F. RCV1: A New Benchmark Collection for Text Categorization Research. *J Mach Learn Res* **5,** 361–397 (2004).
22. Turian, J. *et al.* Word representations: A simple and general method for semisupervised learning. in 384–394 (2010).
23. Collobert, R. *et al.* Natural Language Processing (Almost) from Scratch. *J Mach Learn Res* **12,** 2493–2537 (2011).

A Comparison of Data Driven-based Measures of Adherence to Oral Hypoglycemic Agents in Medicaid Patients

[1]Vivienne J. Zhu, MD, MS, [1]Wanzhu Tu, PhD,
[1]Marc B. Rosenman, MD, [2]J. Marc Overhage, MD, PhD,
[1]Regenstrief Institute, IN and [2]Siemens Healthcare, Malvern, PA

Abstract: *We evaluated and compared different methods for measuring adherence to Oral Antihyperglycemic Agents (OHA), based on the correlation between these measures and glycated hemoglobin A1C (HbA1c) levels in Medicaid patients with Type 2 diabetes. An observational sample of 831 Medicaid patients with Type 2 diabetes who had HbA1c test results recorded between January 1, 2001 and December 31, 2005 was identified in the Indiana Network of Patient Care (INPC). OHA adherence was measured by medication possession ratio (MPR), proportion of days covered (PDC), and the number of gaps (GAP) for 3, 6, and 12-month intervals prior to the HbA1c test date. All three OHA adherence measurements showed consistent and significant correlation with HbA1c level. The 6-month PDC showed the strongest association with HbA1c levels in both unadjusted (-1.07, P<0.0001) and adjusted (-1.12, P<0.0001) models.*

Background

Medication non-adherence is a major problem in health care, especially among patients with chronic conditions like diabetes, which has estimated non-adherence rates between 36% and 87%.[1] Non-adherence to prescribed oral antihyperglycemic agents (OHA) can cause serious consequences to diabetic patients, with higher rates of micro- and macro-vascular complications, increased emergency medical events and a higher mortality rate.[2] In addition, the costs of poor medication adherence for all conditions are estimated at hundreds of billions of US dollars per year.[3] Despite the known consequences, one study reported that the medication adherence rate for patients with diabetes has not improved for over 30 years.[4]

Although an effective adherence intervention may have a greater effect on population health than many other medical treatment improvements,[5] information about patients' adherence is usually not available to health care professionals. Patient self-reported medication adherence is sometimes used to estimate patient medication taking behavior, but they are subject to recall bias and do not correlate well with other methods of assessing adherence.[6-7] While successful interventions have combined convenient care, information, reminders, self-monitoring, and counseling,[8] studies also suggest that developing a data-driven approach to better measure adherence and initiating interventions in clinical practice can potentially improve both adherence and clinical outcomes.[9] An accurate assessment of OHA adherence and understanding its association with glycated hemoglobin A1C (HbA1c level), which is one of the objective measures of glycemic control for diabetic patients, is the first step towards improving OHA adherence.

The definition of adherence varies, and there is no consensus on the best method of measurement. The medication possession ratio (MPR) reflects the patient's overall accordance with the prescribed dosing regimen and disregards the timeliness of particular refills. On the other hand, both the proportion of days covered (PDC) and the gap (GAP) focus on duration or continuation of prescribed treatment, and they both take the timeliness of each refill into account.[10] In addition, different time frames are used to measure adherence; the most frequently used is 12-months.[11] However, some studies showed significant improvement in health outcomes if patients have good medication adherence in 3-month or 6-month intervals. In order to identify the most helpful feedback to physicians, the adherence measures that are best correlated with the HbA1c level, PDC, MPR, and GAP across different intervals should be analyzed and compared.

Previous studies have demonstrated the significant association between OHA adherence and HbA1c level in clinical trials or specific diabetes management programs.[12-14] However, the subjects of these studies were followed for short periods of time or they were informed that their medication use was being monitored. The extent to which these design features affect the validity of study findings remains unclear. As a result,

conclusions drawn from these studies provide somewhat limited insight into the long-term effectiveness of drugs in real-world populations and settings.

In order to identify the measure of patient adherence best suited for providing feedback to physicians, we undertook a study using data from a population-based health information exchange (HIE). We calculated OHA adherence using three different measures (PDC, MPR, and GAP) across three time intervals (3, 6, and 12 months) utilizing longitudinal HIE data. We also analyzed and compared the effects of these objective adherence measurements and patient factors on HbA1c level based on laboratory test results from our local, operational HIE.

Methods

Data sources and settings

We extracted patient information from Indiana Medicaid data which contained demographic (race, gender and age), diagnosis, and treatment information over time. The OHA prescription claims records include refill dates, days of supply, dose, and frequency. HbA1c test results were retrieved from the Indiana Network of Patient Care (INPC). The INPC is an operational regional clinical informatics network that has served Indianapolis for more than fifteen years (and now includes more than 90 Indiana hospitals and more than 22,000 physicians among its members). This system delivers medical record information from hospitals, laboratories, imaging centers, pharmacies, and physician offices, including registration records, laboratory tests, radiology reports, diagnosis and administrative data.[15] The claims-based medication dispensing data was linked to the INPC laboratory data by medical record number. The medical record number is assigned to patients once they visit any facility in the INPC institutions, such as hospitals, laboratories, and clinics. This study was approved by the Institutional Review Board of Indiana University and the INPC Management Committee.

Eligibility Criteria

The study sample was limited to patients with Medicaid coverage who were prescribed an OHA and who had HbA1c data in the INPC. Inclusion criteria were established as follows for the study period January 1, 2001 through December 31, 2005:

1: 18-64 years old in Indiana Medicaid data during the study period.
2: Have at least one *International Classification of Diseases, Ninth Revision, Clinical Modification* (ICD_9_CM) code for Type 2 diabetes (250.X0 or 250.X2) in Medicaid claims for either inpatient or outpatient encounters.
3: Have at least one First Databank *Standard Therapeutic Code* (STC: 71) for OHAs: Biguanides, Sulfonylurea (SU), Thiazolidinedione (TZD) and other OHAs (Meglitinides and α-glucosidase) in Medicaid medication claims.
4: Does not take fixed-combination OHA regimens.
5: Does not take insulin (STC: 0177).
6: Have a medical record number in the INPC from one of the major hospital systems in Central Indiana.
7: Have at least one HbA1c test result from the INPC during the study period.
8: Have at least one OHA prescription prior to HbA1c test date.

Measurements

The independent variable, medication adherence, was measured by calculating PDC, MPR, and GAP. PDC is defined as the total number of medication-covered days divided by the number of days in a certain time period. PDC can be calculated if a subject has even one fill and has been used increasingly to measure patient medication adherence for quality assurance.[16] MPR is commonly calculated as the total number of days supplied by all refills divided by the number of days between the first and last refill, and it usually requires at least two refills date to be caculated.[17] Both PDC and MPR range from 0 to 1. GAP assesses any lapse in medication therapy. GAP is measured in days with various lengths where 30 days is considered significant enough to cause suboptimal clinical outcomes. [16] Both MPR and GAP need at least two refill dates to be calculated.

The dependent variable was the patient HbA1c level based on the INPC laboratory test results. To dynamically and accurately reflect the effect of OHA adherence on HbA1c level, we defined the HbA1c

test date as the index date, and then traced back the patient medication adherence prior to this index date. For each patient, MPR, PDC, and GAP were calculated for 3-month, 6-month, and 12-month intervals prior to each HbA1c test date. For patients who were taking multiple OHAs, the average adherence was counted to reflect the overall medication taking behavior.

In order to control for possible confounders which may influence patient HbA1c levels, [18-19] we analyzed age, gender, race, duration of OHA treatment, and number of concurrent OHAs. The prescribed OHA drug classes included Biguanides, Sulfonylurea, Thiazolidinedione, other OHAs, and multiple classes.

Statistical Analysis

Levels of patients' medication adherence were assessed through three different metrics: PDC, MPR, and GAP, measured over 3, 6, and 12-month intervals. Average HbA1c values were calculated and reported. We examined the associations between HbA1c level and various adherence metrics using mixed effect generalized linear regression models. Random subject effects were used in these models to accommodate the potential association among observations contributed by the same study subjects. All analyses were implemented using SAS 9.1 (*SAS Institute, Cary, North Carolina*). *p*-values less than 0.05 were considered significant.

Table 1. Summary of selected characteristics of subjects and their hemoglobin A1c

	Number of Subjects (Percentage) (n=831)	HbA1c (%) Mean (95% CI)
Demographics		
Age (year) mean= 48		
18-30	71 (8.54%)	7.98 (7.64-8.31)
31-40	150 (18.05%)	7.87 (7.66-8.07)
41-50	295 (35.49%)	7.62 (7.52-7.73)
51-64	314 (37.78%)	7.53 (7.44-7.63)
Gender		
Female	570 (68.69%)	7.66 (7.57-7.73)
Male	261 (31.31%)	7.62 (7.51-7.73)
Race		
African-American	371 (44.64%)	7.88 (7.79-7.98)
Hispanic	7 (0.80%)	7.78 (6.90-8.63)
Asian	4 (0.48%)	6.55 (6.20-6.90)
Other	11 (1.30%)	6.98 (6.45-7.50)
White	438 (52.70%)	7.42 (7.33-7.51)
Diabetes Severity		
Duration of OHA Treatment (Year)		
0-3	514 (61.85%)	7.60 (7.51-7.69)
3-6	241 (29.00%)	7.61 (7.49-7.72)
6-9	76 (9.15%)	7.97 (7.77-8.17)
Number of Concurrent OHAs		
1	507 (61.01%)	7.29 (7.18-7.41)
2	253 (30.44%)	7.87 (7.75-7.99)
>=3	71 (8.55%)	7.98 (7.75-8.22)
OHA Classes		
Biguanides Only	224 (26.96%)	7.54 (7.42-7.67)
Sulfonylurea Only	214 (25.75%)	7.80 (7.69-7.97)
Thiazolidinedione Only	25 (3.00%)	7.66 (7.45-7.87)
Other	44 (5.29%)	6.81 (6.08-7.53)

Results

Demographic and Clinical Characteristics

A total of 831 subjects met all inclusion and exclusion criteria. The average entry age of study subjects was 48 years. Female subjects accounted for 68.7% of the sample. The average HbA1c level of the study

population was 7.60% (95% CI: 7.58%-7.71%). The average duration of OHA treatment was 2.09 years. The average number of HbA1c tests was 3.5 per patient. The majority of the study sample (61.0%) was taking one medication: 27.0% were prescribed Biguanides, 25.8 % SUs, 3.0% TZDs and 5.3% other drugs. More than one OHA was being taken by 39.0% of patients; such patients had a slightly lower HbA1c level than patients who were treated by SU only (Table 1).

OHA Adherence, Other Covariates, and Their Association with HbA1c Control
From January 1, 2001 to December 31, 2005, a total of 1,721 to 2,934 observations of OHA adherence and HbA1c results were formed for 831 subjects. Table 2 summarizes frequency and average value of adherence (PDC, MPR and GAP) at time intervals of 3, 6, and 12-months. The average adherence ranged from 39% to 85%. In unadjusted analyses, all three OHA adherence measurements for 6 or 12 months showed consistent and significant associations with HbA1c control. In adjusted analyses, PDC and MPR measured for 6 or 12 months were significantly correlated with HbA1c. The 6-month PDC showed the greatest association with HbA1c control in both unadjusted and adjusted analyses (Table 3).

Table 2. Patient adherence measures mean value and the 95% confidence interval across three time intervals

	3-month		6-month		12-month	
Adherence	**Mean (95 CI)**	**Freq**	**Mean (95 CI)**	**Freq**	**Mean (95 CI)**	**Freq**
PDC	0.60 (0.59-0.61)	2,795	0.51 (0.49-0.52)	2,838	0.39 (0.38-0.41)	2,934
MPR	0.85 (0.84-0.86)	1,721	0.82 (0.81-0.83)	2,117	0.79 (0.78-0.80)	2,336
GAP	0.13 (0.12-0.14)	1,721	0.18 (0.16-0.20)	2,117	0.25 (0.24-0.28)	2,336

Freq =Frequency
PDC and MPR ranged from 0 to 1, and GAP ranged from 0 to 5

Table 3. Unadjusted, adjusted coefficients and 95% confident intervals between OHA adherence and HbA1c control

Adherence	3-month	6-month	12-month
Unadjusted estimate			
PDC	-0.98 (-1.20, -0.76)	-1.07 (-1.28, -0.87)	-1.01 (-1.21, -0.81)
MPR	-0.51 (-0.94, 0.07) †	-0.92 (-1.29, -0.56)	-0.90 (-1.20, -0.59)
GAP	-- --	0.25 (0.12, 0.38)	0.19 (0.11, 0.29)
Adjusted estimate			
PDC	-0.89 (-1.12, -0.67)	-1.12 (-1.35, -0.91)	-1.20 (-1.42, -0.96)
MPR	-0.29 (-0.72, 0.14) †	-0.68 (-1.06, -0.32)	-0.87 (-1.19, -0.55)
GAP	0.19 (-0.002, 0.39) †	0.05 (-0.10, 0.20) †	0.05 (-0.06, 0.17) †

-- data did not converge
† *p*-value is greater than 0.05. *p*-value for any other unadjusted coefficients is smaller than 0.0001.

Adjusted with other factors, the coefficients between the 6-month adherence (PDC, MPR and GAP) and HbA1c level are shown in Table 4. The 6-month PDC showed greatest association (-1.12, P<0.0001) with HbA1c control. Among patient factors, increased age was correlated with better HbA1c control (P<0.0001). African-Americans had a higher average HbA1c level as compared with Whites (P<0.0001). The associations between number of medications and HbA1c level were about 0.41 (p<0.0001). Compared with patients treated with multiple OHA classes, patients treated with SU had slightly higher HbA1c levels (P<0.0001). Gender and duration of OHA treatment had no effect on HbA1c level.

Discussion
We have two main findings from this study. First, increased PDC and MPR are strongly correlated with lower HbA1c level while increased GAP relates to higher HbA1c level. Second, across different adherence measures and different time frames, 6-month PDC is more correlated with HbA1c level than other measures in both unadjusted and adjusted models.

The primary goal for this study is to analyze and compare associations between measurements of OHA adherence and HbA1c level among patients with Type 2 diabetes using real-world clinical data. In most cases (except in the 3-month models), PDC and MPR produced significant negative coefficients with HbA1c level, meaning that increased PDC or MPR is related to decreased HbA1c level. On the other hand, GAP produced positive coefficients, which means that increased GAP is associated with increased HbA1c level. The coefficients in the unadjusted model indicated that a 10% increase in PDC/MPR is related to a 0.09-0.10% reduction in HbA1c. In contrast, an increase in GAP of one is correlated with a 0.19-0.25% increase in HbA1c level. The adjusted model demonstrated similar results except that GAP is not significantly correlated with HbA1c These findings are consistent with results from previous studies that OHA adherence was independently associated with HbA1c control: HbA1c decreases 0.10% to 0.16% for each 10% increment in OHA adherence.[20]

Table 4. Adjusted coefficients between adherence, covariates and HbA1c level at 6-month interval

Predictors	PDC	MPR	GAP
Intercept	7.30 ± 0.44 ††	7.08 ± 0.58 ††	7.13 ± 0.21 ††
6-Month Adherence	-1.12 ± 0.10 ††	-0.68 ± 0.18 ††	0.26 ± 0.06 ††
Age *	-0.09 ± 0.03 ††	-0.11 ± 0.04 ††	-0.16 ± 0.04 ††
Gender			
Female	0.10 ± 0.6	0.16 ± 0.7	0.15 ± 0.7
Male	--	--	--
Race			
African-American	0.28 ± 0.06 ††	0.24 ± 0.07 ††	0.30 ± 0.06 ††
Hispanic	-0.10 ± 0.30	-0.34 ± 0.32	-0.26 ± 0.30
Asian	-0.45 ± 0.69	-0.11 ± 0.71	-0.36 ± 0.70
Other	-0.30 ± 0.35	-0.36 ± 0.36	-0.38 ± 0.36
White	--	--	--
Duration of OHA Treatment	0.20 ± 0.10 †	0.24 ± 0.11	-0.08 ± 0.13
Number of Medications	0.38 ± 0.04 ††	0.41 ± 0.05 ††	0.42 ± 0.05 ††
OHA Class			
Biguanides Only	1.04 ± 0.35	0.86 ± 0.53	0.77 ± 0.36
Sulfonylurea Only	1.46 ± 0.35 †	1.23 ± 0.53 †	1.18 ± 0.36 †
Thiazolidinedione Only	0.94 ± 0.36	0.67 ± 0.53	0.71 ± 0.37
Other	-0.03 ± 0.48	-0.37 ± 0.62	-0.34 ± 0.48
Multiple Classes	--	--	--

Age***** units in 10 years
† *p*-value is greater than 0.0001 and smaller than 0.05
†† *p*-value is smaller than 0.0001
-- = reference

All three measures of adherence, PDC, MPR, and GAP, were significantly correlated with HbA1c control but in different degrees. PDC had the biggest coefficient values in both adjusted and unadjusted models which indicated that PDC was most correlated with HbA1c level (Table 3, Table 4). These results are consistent with how PDC, MPR and GAP are calculated. MPR is calculated by adding the days' supply for all medications and then dividing over a certain period of time. [17] It assumes that all drugs eventually get used within the time period, which may overestimate the actual adherence if patients refill their medication before the last date of the preceding prescription. In contrast to MPR, PDC looks at each day to determine if the patient has one or more dispensed drugs and then determines the proportion of days that a patient has a drug available in a study interval. [16] Theoretically, PDC more accurately reflects patient adherence behavior, and it more effectively handles drug switching and prescription overlaps. GAP is simply measured by calculating the number of medication lapses greater than 30 days,[16] and it can be used as a reference to confirm the pattern of PDC or MPR (Figure 1).

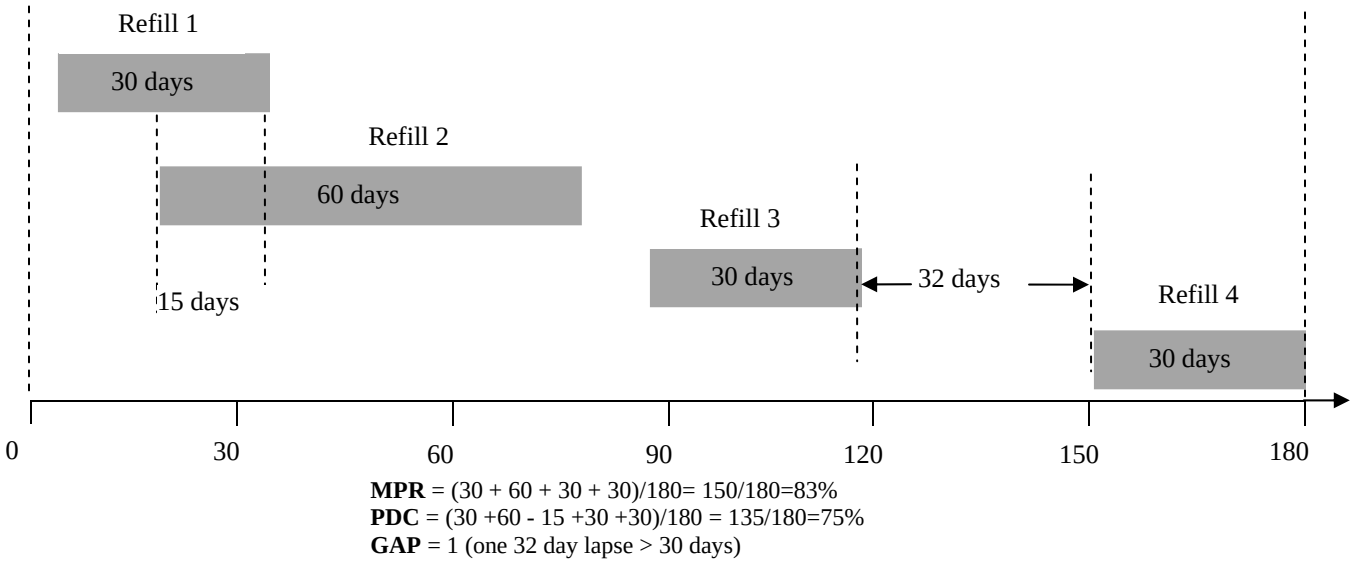


Figure 1. An example of calculating MPR, PDC and GAP.

This study also compared OHA adherence with HbA1c control across three time intervals. In the unadjusted model, significance disappeared for both MPR and GAP in the 3-month interval. Further investigation showed that 11% of refills were prescribed with a 90-day supply, resulting in insufficient information to calculate the MPR and GAP since they both would need at least two refill dates during the 3-month interval. The coefficients were close for PDC, MPR, and GAP across the 6-month and 12-month intervals. However, from a clinical perspective, the 6-month interval can provide patient adherence information in a more timely fashion. In summary, the 6-month PDC most accurately reflects OHA adherence and is the measure most closely associated with HbA1c level for patients with Type 2 diabetes in our study.

It is well-known that race is one predictor of suboptimal HbA1c control.[20] In our study, compared with Whites, African-Americans had a significantly higher HbA1c level (by 0.28-0.30%, $p<0.0001$). Sociodemographic, behavioral, genetic, or biological factors may independently or partially affect HbA1c level.[21-23] Our study also found that increased age is related to better HbA1c control, which is generally consistent with the notion that young patients are less likely to benefit from OHA therapy.[22] We additionally observed positive associations between number of concurrent OHAs and HbA1c levels. Patients were prescribed one additional OHA when their HbA1c level increased by 0.4%. A plausible explanation might be that patient may have been prescribed additional OHAs because they were not responding to one. [25]

The main findings from this study provided evidence and knowledge to establish intervention in medication adherence to improve health outcomes for patients with Type 2 diabetes. Challenges of medication adherence in diabetes are at patient, medication and provider levels, and a multi-dimensional approach is required to establish efficient interventions. Health information technology (HIT) and health information exchange (HIE) offer great potential to establish such a system. First, objective and data-driven approaches can be programmatically established through an HIE. These objective adherence measures enable accurate assessment of patient medication taking behaviors, which is the essential for physicians to estimate the effectiveness of treatment. This study provides specific information on which to base a choice for adherence measures in clinical practice. Second, a clinical decision support system may deliver patient adherence information and generate relevant recommendations in routine clinical practice. In addition, a well-established HIE/HIT supports patient-centric and team-based care that better engaged patients, providers and health care systems for improving medication adherence.

Limitations

Certain limitations should be recognized. First, the study population was Indiana Medicaid members

younger than 65, with relatively low socioeconomic status and with at least one HbA1c laboratory result in the INPC which covers the central Indiana region most closely. Therefore, the findings from this study may lack generalizability to all patients with Type 2 diabetes, including those who had no HbA1c results recorded in the INPC. Second, dispensing claim-based measures may not be equivalent to measures of the actual ingestion of medication. Nevertheless, filling a prescription is usually consistent with taking medication.[26] It also should be noted that it is possible we did not capture dispensing information about free samples or free medications offered by providers or pharmacies. Third, the average HbA1c levels of this study population were elevated (>7.0%) even though the average OHA treatment duration was 2.09 years. Provider inertia may be playing a role in inadequate glycemic control apart from patient adherence. However, we lack the necessary data to further analyze if providers have failed to adequately intensify OHAs, or to initiate insulin or HbA1c tests.[27] Similarly, the comparisons of different OHA adherence measures mainly rely on their effects on HbA1c levels, and we did not study pharmacodynamics/kinetics contributions, which may also affect patient HbA1c levels. Fourth, we excluded patients using insulin. However, insulin is commonly administered to patients with Type 2 diabetes whose HbA1c is poorly controlled. One study suggested that insulin use may affect adherence rates to oral medications.[28] Whether there is a significant association between insulin use and OHA adherence is an area for further study.

Conclusion

By evaluating real-world clinical data from the INPC and Medicaid claims data, this study confirmed the strong association between OHA adherence measured by PDC/MPR and HbA1c level among Medicaid patients with Type 2 diabetes. The 6-month PDC is the best measure of OHA adherence in this population. This study also suggested that linking HIE laboratory data with claims data is a helpful approach for comparing medication adherence and clinical phenomena.

Acknowledgements

We would like to thank Roberta Ambuehl (senior data analyst) at Regenstrief Institute, Inc.; the NLM Medical Informatics Fellowship Program (5T15LM007117), Regenstrief Institute, Inc. Vivienne Zhu researched data, initiated discussion, and wrote the manuscript. Wanzhu Tu researched data, contributed to discussion, and reviewed/edit the manuscript, Marc Rosenman contributed to discussion and reviewed/edit the manuscript, J. Marc Overhage led manuscript definition of intellectual content, researched data, contributed to discussion, and wrote/reviewed/edit the manuscript. J. Marc Overhage, Wanzhu Tu, and Vivienne Zhu are responsible for the contents of the article. No author has potential conflicts of interest to be disclosed.

References

1. Lee WC, Balu S, Cobden D, Joshi AV, Pashos CL. Prevalence and Economic Consequences of Medication Adherence in Diabetes: A Systematic Literature Review. Manag Care Interface. 2006 Jul;19(7):31-41.
2. Ho PM, Bryson CL, Rumsfeld JS. Medication adherence: its importance in cardiovascular outcomes. *Circulation*. 2009 Jun 16;119(23):3028-35.
3. Egede LE, Gebregziabher M, Dismuke CE, et al. Medication nonadherence in diabetes: longitudinal effects on costs and potential cost savings from improvement. *Diabetes Care*. 2012 Dec;35(12):2533-9.
4. Rubin RR. Adherence to pharmacologic therapy in patients with type 2 diabetes mellitus. *Am J Med*. 2005;118(suppl 5A):27S- 34S.
5. World Health Origanization. Adherence to Long-term therapies: Evidence for action. 2003.
6. Cook CL, Wade WE, Martin BC, Perri M 3rd. Concordance among three self-reported measures of medication adherence and pharmacy refill records. *J Am Pharm Assoc (2003).* 2005 Mar-Apr;45(2):151-9.
7. Hansen RA, Kim MM, Song L, Tu W, Wu J, Murray MD. Comparison of methods to assess medication adherence and classify nonadherence. *Ann Pharmacother*. 2009 Mar;43(3):413-22.
8. Lars Osterberg. Adherence to Medication. N Engl J Med. 2005:353:487-97.
9. Cohen HW, Shmukler C, Ullman R, Rivera CM, Walker EA. Measurements of medication adherence in diabetic patients with poorly controlled HbA(1c). *Diabet Med*. 2010 Feb;27(2):210-6.

10. N.M. Vink, O.H.Klungel, R.P. Stolk, P. Deng. Comparison of various measures for assessing medication refilladherence using prescription data. *Pharmacoepidemiol Drug Saf*. 2009 Feb;18(2):159-65.
11. Caetano PA, Lam JM, Morgan SG. Toward a standard definition and measurement of persistence with drug therapy: Examples from research on statin and antihypertensive utilization. *Clin Ther*. 2006 Sep;28(9):1411-24;
12. Rozenfeld Y, Hunt JS, Plauschinat C, Wong KS. Oral antidiabetic medication adherence and glycemic control in managed care. *Am J Manag Care*.
2008 Feb;14(2):71-5.
13. Kimberley Krapek, Kathleen King, Susan S Warren, et al. Medication adherence and associated hemoglobin HbA1c in Type 2 diabetes. *Ann Pharmacother*. 2004;38(9):1357-62.
14. Lawrence DB, Ragucci KR, Long LB, Parris BS, Helfer LA. Relationship of oral antihyperglycemic (sulfonylurea or metformin) medication adherence and hemoglobin A1c goal attainment for HMO patients enrolled in a diabetes disease management program. *J Manag Care Pharm*. 2006 Jul-Aug;12(6):466-71.
15. Clement J. McDonald, J. Marc Overhage, Michael Barnes, Gunther, Schadow LB, Paul R. Dexter, Burke Mamlin. The Indiana Network For Patient Care: A working local health information infrastructure. *Health Aff (Millwood)*. 2005 Sep-Oct;24(5):1214-20.
16. National Committee for Quality Assurance. Pharmacy Quality Alliance (PQA) Demonstration Project. 2008.
17. Warren J, Warren D, Yang HY, Mabotuwana T, Kennelly J, Kenealy T, Harrison. Prescribing history to identify candidates for chronic condition medication adherence promotion. *Stud Health Technol Inform.* 2011;169:634-8. J.
18. Yi Yang, Vennela Thumula, Patrick F. Pace, Benjamin F. Banahan III, Noel E. Wilkin, and William B. Lobb. Predictors of medication nonadherence among patients with diabetes in medicare part D programs: A retrospective cohort study. *Clin Ther*. 2009 Oct;31(10):2178-88.
19. Curkendall SM, Thomas N, Bell KF, Juneau PL, Weiss AJ. Predictors of medication adherence in patients with type 2 diabetes mellitus. *Curr Med Res Opin*. 2013 Oct;29(10):1275-86.
20. Joel M. Schectman, Mohan M. Nadkarni, John D. Voss. The association between diabetes metabolic control and drug adherence in an indigent population. *Diabetes Care*. 2002 Jun;25(6):1015-21.
21. Sequist TD, Fitzmaurice GM, Marshall R, et al. Physician performance and racial disparities in diabetes mellitus care. *Arch Intern Med*. 2008 Jun 9;168(11):1145-51.
22. Elizabeth Selvin MwS CMB, Hoogeveen, Josef Coresh, Frederick L. Brancati. Racial differences in glycemic markers: A Cross-sectional analysis of community-based data. *Ann Intern Med.* 2011;154(5):303-9.
23. Chandalia M GS A-HB, Abate N. Ethnic differences in the frequency of ENPP1/PC1 121Q genetic variant in the Dallas Heart Study cohort. *J Diabetes Complications*. 2007;21(3):143-8.
24. Odegard PS, Capoccia K. Medication taking and diabetes: A systematic review of the literature. *Diabetes Educ*. 2007 Nov-Dec;33(6):1014-29.
25. Yurgin N, Secnik K, Lage MJ. Antidiabetic prescriptions and glycemic control in German patients with type 2 diabetes mellitus: A retrospective database study. *Clin Ther*. 2007 Feb;29(2):316-25.
26. Steiner JF, Prochazka AV. The assessment of refill compliance using pharmacy records: Methods, validity,and application. *J Clin Epidemiol*. 1997 Jan;50(1):105-16.
27. Vinik A. Advanced therapy in type 2 diabetes mellitus with early, comprehensive progression from oral agents to insulin therapy. *Clin Ther*. 2007 Jun;29:1236-53.
28. Gérard Reach, Véronique Le Pautremat, Shaloo Gupta. Determinants and consequences of insulin initiation for type 2 diabetes in France: analysis of the National Health and Wellness Survey. *Patient Prefer Adherence*. 2013; 7: 1007–1023.

Building a Treebank of hospital discharge summaries

Min Jiang, M.S.[1],
Yang Huang, Ph.D.[2], Jung-Wei Fan, Ph.D.[2], Elly W Yang M.A.[2], Hua Xu Ph.D.[1]
[1]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA
[2]Kaiser Permanente, San Diego, CA, USA

Introduction In order to build high-performance syntactic parsers in the medical domain, Treebanks that contain manually annotated parse trees of clinical sentences are needed. Current efforts on clinical Treebank include the MiPACQ corpus that consists of various types of clinical notes about colon cancer and a small collection of 25 progress notes that we built previously. In this abstract, we describe our current effort on building a Treebank of hospital discharge summaries, an important type of clinical documents that cover a broad type of clinical concepts.

Method In our previous study, we developed a parse tree annotation guideline[1] for clinical sentences that are often ill formed. Here we applied this guideline to annotating discharge summaries contained in the 2010 i2b2 clinical NLP challenge, a total of 237 documents. Every sentence in these discharge summaries was preprocessed by the Stanford parser and a researcher in clinical NLP manually reviewed each parse tree generated by the Stanford parser and corrected them based on the annotation guideline. We used WordFreak, an annotation tool, to perform our manual annotation. When there is any question about a parse tree, a linguist and a physician were consulted for correct annotation. To ensure the quality of annotation, a small subset of sentences in discharge a second NLP researcher also annotated summaries.

Results Thus far, we have annotated 60 discharge summaries, accounting for about 6000 sentences. Compared with progress notes, discharge summaries show some different characteristics. In average, the sentence in the progress notes contains 8.1 tokens, which is far less than the one in the discharge summaries (15.3 tokens). Furthermore, Table 1 shows a preliminary comparison of the distribution of main syntactic constructs (constituents) in two Treebanks. Compared to the progress notes Treebank, Noun Phrase has the biggest increase of the portion (increase from 40.0% to 46.8) in all the constituents of discharge summaries Treebank and Verb Phrase drops the most on the percentage of all the constituents in discharge summaries Treebank.

Discussion Due to the inherent differences in the use of progress notes and discharge summaries, there is a difference in the sentence structure found in these two types of notes. Progress notes are usually written in an informal, brief manner, which explains the difference in the sentence length from the discharge summaries. On the other hand, discharge summaries, as a summary of all the clinical reports, contain more comprehensive clinical concepts and probably more syntactic patterns. Therefore, a Treebank of the discharge summaries would be a valuable data set for building more generalizable parsers for clinical notes. We plan to finish annotating all discharge summaries in the i2b2 data set and make the Treebank freely available to the research community.

Table 1 – Statistics of distribution of main syntactic constituents in two Treebanks

Constituent label	**S**	**Frag**	**NP**	**VP**	**PP**	**ADJP**	**ADVP**	**SBAR**
Constituent description	Sentence	Fragment	Noun phrase	Verb Phrase	Prepositional Phrase	Adjective Phrase	Adverb Phrase	Clause
Progress notes	12.2	7.7	40.0	16.8	8.7	4.5	2.8	1.4
Discharge summaries	9.4	7.9	46.8	13.9	10.5	2.8	3.6	1.3

1. Fan, J.W., et al., Syntactic parsing of clinical text: guideline and corpus development with handling ill-formed sentences. J Am Med Inform Assoc, 2013. 20(6): p. 1168-77. 2013

The Pareto Principle in the ICU: A Model for Knowledge Resource Health Information Technologies

Christopher A. Aakre, MD, Marc A. Ellsworth, MD, Brian Pickering, MD, Vitaly Herasevich, MD, PhD
Mayo Clinic, Rochester, MN

Abstract: *The frequency and distribution of medication orders and lab results in intensive care units follow the Pareto Principle, colloquially known as the "80/20 Rule". Utilizing this knowledge is helpful in the design of health information technologies, most specifically point-of-care knowledge resource tools, which limit information overload and provide clinical users with the most pertinent and needed clinical information.*

Introduction: Clinical providers in the intensive care unit are confronted daily with large amounts of data. This data must be distilled into clinically useful information in real-time for use at the bedside. When designing a health information technology solution to reduce information overload and user-interface clutter, it is important to understand the frequency and distribution of the clinical data available to providers at the point-of-care. We hypothesized that the frequency of medication orders and lab results in the neonatal intensive care (NICU), pediatric intensive care (PICU), and medical intensive care (MICU) units follow the Pareto Principle.[1] This concept has previously been validated in the outpatient setting,[2] but it has not yet been validated in a real-time data-intensive inpatient setting such as the intensive care unit.

Methods: The Multidisciplinary Epidemiology and Translational Research in Intensive Care (METRIC) DataMart, combines real-time ICU clinical information from many EMR systems at the Mayo Clinic into one searchable integrative database.[3] This DataMart was used to retrieve all medication orders and lab results that occurred in our institution's NICU, PICU, and MICU from January 1, 2012 – December 31, 2012. Our simple Pareto Plot was created by comparing the frequency of each specific medication and lab test against the cumulative number of medication orders and laboratory test results.

Results: During the study period the total number of medication orders and unique medications were as follows: NICU – 9,065/179; PICU – 25,054/478; MICU – 83,997/700. The total number of lab results and unique lab tests were: NICU – 65,044/459; PICU – 89,202/900; MICU - 367,905/1,033. Within their respective units, the top 20.1% (NICU), 15.2% (PICU) and 9.9% (MICU) of unique medications accounted for 80% of all medication orders. Similarly, 4.6% (NICU), 3.2% (PICU) and 3% (MICU) of all unique laboratory tests were responsible for 80% of all laboratory results.

Discussion: In contrast with the outpatient setting previously studied by Wright et al, the intensive care unit is an environment with a high volume of time sensitive data. The large volumes of complex data in this setting make point-of-care knowledge resource tools attractive to busy clinicians. Our data demonstrates that the frequency of medication orders and lab results in the intensive care unit follow the law of the vital few, otherwise known as the Pareto Principle. When designing point-of-care knowledge resource tools for the intensive care setting, the application of the Pareto Principle can help identify the most common clinical data for inclusion - thereby reducing interface clutter from less frequently used items. In most common data scenarios confronting the intensive care clinician, the Pareto Principle suggests that the most common 20% of clinical data will generate a vast majority of the alerts. A careful design of the point-of-care knowledge tool's interface may help minimize alert fatigue attributed to the most common data obtained in the intensive care unit. Further study is needed to characterize clinician attitudes on frequency of alerts pertaining to less familiar data.

References

1. Dugmore CR. The 80-20 phenomenon (80:20 distribution of caries)--myth or fact. Br Dent J. 2006;201(4):197-8.
2. Wright A, Bates DW. Distribution of Problems, Medications and Lab Results in Electronic Health Records: The Pareto Principle at Work. Appl Clin Inform. 2010;1(1):32-37.
3. Herasevich V, Pickering BW, Dong Y, Peters SG, Gajic O. Informatics infrastructure for syndrome surveillance, decision support, reporting, and modeling of critical illness. Mayo Clin Proc. 2010;85(3):247–54.

Electronic Dental Record Research: descriptive review of current status

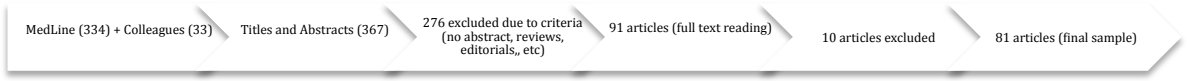
Renata Abramovicz-Finkelsztain MDS, PhD candidate [1],
Claudia G N Barsottini MSc, PhD, Associate Professor[1],
Heimar Fatima Marin MSc, PhD, Full Professor[1]
Department of Health Informatics, Universidade Federal de São Paulo, Brazil

Abstract

As the adoption of Electronic Dental Record (EDR) increased during the last decade, so did the research in the field. In order to describe the current status of the research related to EDR we developed a descriptive bibliometric review followed by a description of the methodology and object of the found studies.

Introduction and Methods

During the last decade surveys indicated a worldwide trend in the use of Electronic Dental Record (EDR) for the management of patient clinical data.(1-4); and, as a consequencet so does the research related to this field. In order to describe the current status of the electronic dental records research we did a bibliometric research(5). In addition, we also described the articles according to their methodology and object of the studies. A research strategy was developed and the data source used was Medline and known authors from the field contribution.



Results

The 81 articles included in the review were published among 41 different journals. The Journal of the American Dental Association, The Journal of Dental Education and the Journal of Public Health Dentistry held 13, 10 and 4 articles each respectively. From the 41 journals, 30 were indexed both in Journal Citation Reports (JCR) and Scientific Journal Ranking (SJR), with an index factor ranging from 0,771 to 7,735; 9 were indexed only in SJR , with index factor ranging from 0,119 to 3,758 and the remaining 2 were not indexed in neither of the databases. The articles selected for the review were analyzed based on the country of origin (Figure 1), the methodology used and the object of the study (Figure 2). A research trend during the last decade is also shown. (Figure 3).

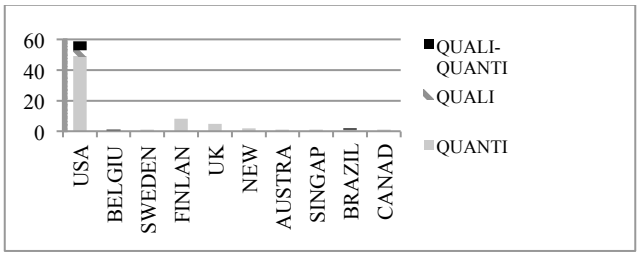


Figure 1- Country X Methodology

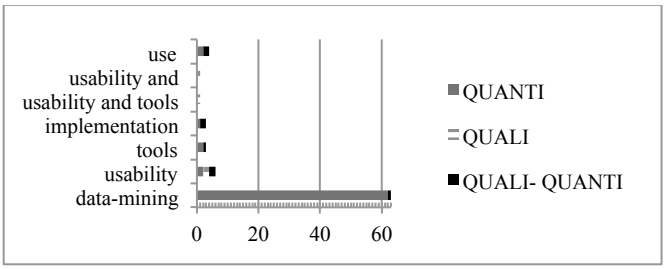


Figure 2- Methodology X Object of Study

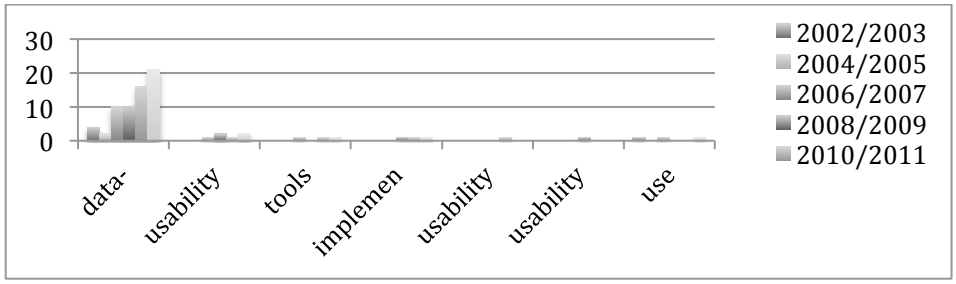


Figure 3- Research Trends

Discussion

While there has been a large number of published studies on EDR, the majority (78%) are editorials, reviews or description of use of the systems and only 22% represent research involving EDRs. From theses studies, only 20% (14 articles) studied EDRs themselves, i.e., evaluated use, usability, implementation and/ or tools. For this reason, it is our view that there is a lack of research on EDRs use and we need the urge the scientific community to develop more studies in order to increase and improve EDR's adoption.

References

1. John JH, Thomas D, Richards D. Questionnaire survey on the use of computerisation in dental practices across the Thames Valley Region. British dental journal. 2003;195(10):585-90; discussion 79.
2. Flores-Mir C, Palmer NG, Northcott HC, Huston C, Major PW. Computer and Internet usage by Canadian dentists. Journal (Canadian Dental Association). 2006;72(2):145.
3. Schleyer TK, Thyvalikakath TP, Spallek H, Torres-Urquidy MH, Hernandez P, Yuhaniak J. Clinical computing in general dentistry. Journal of the American Medical Informatics Association : JAMIA. 2006;13(3):344-52.
4. Center ADAS. 2006 Technology Survey. Chicago: American Dental Association, 2007.
5. Ugolini D, Neri M, Casilli C, Ceppi M, Canessa PA, Ivaldi GP, et al. A bibliometric analysis of scientific production in mesothelioma research. Lung cancer (Amsterdam, Netherlands). 2010;70(2):129-35.

Model selection for EHR laboratory variables: how physiology and the health care process can influence EHR laboratory data and their model representations

David Albers[1], PhD, Rimma Pivovarov[1], MA, Noémie Elhadad[1], PhD, George Hripcsak[1], MD MS
[1]Department of Biomedical Informatics, Columbia University, New York, NY

Abstract

Patient data are present in an electronic health record (EHR) for many reasons; the data in an EHR are a function of a complex set of processes ranging from physiology to the health care process (HCP). Understanding the impact of processes governing the data accumulated in the EHR on large-scale measurement analysis is an open research problem. Here we focus on modeling laboratory measurements independent of time. Phenotypes are collections of summary variables: a glucose measurement summarizes or maps the complex endocrine system to one number; a set of glucose values can then be condensed by parameterization from the set to a few parameters (e.g., a mean). Because EHR data can be influenced by physiology, the HCP, and context, selecting the most useful parameterization can be non-trivial. Moreover, the choice of parameterized model can have a significant impact on what we can observe and how well we can stratify a population into different phenotypes. Here we find that glucose in a broad EHR is best represented, summarized, and modeled by an generalized extreme value (GEV) distribution rather than by a mean and variance.

Introduction

The EHR can provide a platform for large-scale phenotyping. We believe that because EHR data are biased in the way they are recorded, data modeling must account for these biases. Our focus here concerns laboratory test data across a large patient population and developing a method for deconvolving biases and generating phenotypes. Here we show that the HCP intervenes in such a way that outside of an ICU, only extreme values of glucose are captured. As a consequence, large scale phenotype studies should use location and scale instead of mean and variance to model, summarize, and stratify populations.

Methods

Our analysis was conducted using data from the clinical data warehouse at our institution. We observed distributions of glucose in two contexts: an ICU population that is more likely to represent pure physiology, and a broad EHR population that is more likely to be influenced by the HCP. In addition, we modeled the individuals and the populations using normal and GEV distributions. These distributions are generated by distinctly different processes; importantly, GEV distributions are generated by measuring relative maxima of any distribution and are the only distributions that can model extrema of distributions. Moreover, glucose physiologic driven distributions show that mean- and variance-like quantities increase simultaneously with acuity [1]. We will use this to help understand how different statistical models could influence phenotyping though different model parameterizations.

Results

The ICU distribution of glucose is unimodal and symmetric, matching the distribution of glucose captured in a controlled setting where only physiology affects the measured values [1]. The broad EHR glucose distributions do not resemble the ICU glucose distributions but are well fit by a GEV distribution. The linear correlation (LC) between the mean and variance of glucose for the EHR population is 0.63 (95% CI 0.62-0.65). The LC between the location and scale, the mean and variance analogs for the GEV, are 0.82 (95% CI 0.82-0.83). The larger LC and smaller CIs imply the GEV is the natural parameterized family to model glucose in an EHR

Conclusion

The glucose distribution in the ICU represents a physiologic process. The glucose in the broad EHR is a function of both physiology and the HCP; the HCP intervenes such that only relative maxima of individual's glucose values are captured. The GEV is the only distribution that can represent extrema of distributions, implying that the EHR glucose data are collections of relative extreme values of glucose, or glucose captured when patients are ill. Because of this, we hypothesize that HCP biases the glucose in the EHR by only capturing individuals when they are acutely ill. This hypothesis is supported by the fact that location and scale have a strong positive LC that implies and signals increased acuity. With a weaker LC and wider CI, mean and variance do not distinguish healthy and sick individuals as well as location and scale. This means that for phenotyping studies, relative to glucose, it is better to use location and scale of glucose as summary variables for stratifying a population than mean and variance. It is likely that the most representative model will be lab and context dependent.

Acknowledgments

We acknowledge NLM grant R01 LM06910, NLM R01 LM010027, and NSF IGERT #1144854 for support.

References

1. DJ Albers and G Hripcsak, "Population physiology: Leveraging electronic health record data to understand human endocrine dynamics," PLoS One, 7 (12), e48058, 2012.

Health Information Technology Adoption in Home Health ... Research in Progress

Dari AlHuwail[1], Gunes Koru, PhD[1], Ahmad Alaiad[1], Anthony Norcio, PhD[1], Maxim Topaz[2]
1 Department of Information Systems - University of Maryland - Baltimore County, Baltimore, MD, ; 2 School of Nursing - University of Pennsylvania, Philadelphia, PA

Motivation

U.S. health care systems are constantly looking to provide higher quality of care, improve the health outcomes of patients, and reduce overall costs. Home Health Agencies (HHA) are key players in this ecosystem and are an important component of the patient-centered continuum of care. In 2012, Home Health expenditures were $77.8 billion and CMS spending alone accounted for approximately 81% of total home health care spending. [1] Health Information Technology (HIT) has "the potential for achieving numerous health care goals—including efficiency, cost-effectiveness, and patient-centered care". [2]

Research Aims

Today HHAs are ineligible to receive incentives under the Health Information Technology for Economic and Clinical Health (HITECH) act; many HHAs operate with very limited resources and expected to improve quality of care while lowering their costs. HIT adoption can potentially help HHAs to reap the benefits and become more effective. In this evidence-based investigation, we aim to uncover the challenges and opportunities for HHAs to e effectively adopt Health IT solutions. The research proposal was funded through the Agency for Healthcare Research and Quality (AHRQ).

Methods

We used qualitative methods in order to obtain rich and contextual information and inform the limited research in this domain. Purposeful sampling based on agency size, business model, organization structure, and geographical areas was implemented for agencies in the State of Maryland. Participants were recruited through the Maryland National Home Care Association and Maryland Health Care Commission. The unit of analysis was a single HHA. Semi-structured interviews were conducted with CEOs, CIOS, and clinical providers (including nurses and physical therapists). The interviews were recorded and transcribed. Member-checks were also conducted with participants to ensure findings' validity and reliability. The Framework method is used for analysis of the data. Such method is widely used by qualitative researchers allowing to organize raw data to uncover themes and concepts without loosing sight of original data through the different levels of abstraction.[2]

Research Status and Preliminary Findings

All interviews and member-checks were completed and transcribed. Certain themes are starting to emerge.

- For HHAs, compliance with regulatory requirements is complex, time consuming and costly process.
- Patients with varying demographics (including gender, age, and education,) introduce complex needs.\
- Coordination of care across the continuum of healthcare is critical for positive patient outcomes
- Physicians are resistant to use HHAs's HIT
- High turnover of HHA staffing and big age gap amongst clinical providers

Conclusion

HHAs struggle to keep up with complex regulations and requirements while operating with limited resources. Many operational and clinical challenges are complex to solve; effective and efficient HIT adoption is amongst such challenges. Evidence-based findings from this study will uncover issues of HIT adoption in HHAs to inform policy makers and HHA administrators.

References

1. Centers for Medicare & Medicaid Services (CMS), "National Health Expenditures 2011 Highlights," 2011.
2. N. Ruggiano, E. L. Brown, V. Hristidis, and T. F. Page, "Adding Home Health Care to the Discussion on Health Information Technology Policy," Home Health Care Serv. Q., vol. 32, no. 3, pp. 149–162, 2013.
3. J. Ritchie and J. Lewis, Qualitative Research Practice: A Guide for Social Science Students and Researchers, 1st ed. Sage Publications Ltd, 2003.

Dental Information Needs – a Survey of the Medical Health Providers

Márcia C. Almeida, DMD, UCP[1], Amit Acharya, BDS, MS, PhD[2] , Altamiro C. Pereira, MD, PhD[3], André R. Correia, DMD, PhD, UCP[1]

[1]Portuguese Catholic University, Viseu, Portugal; [2]Marshfield Clinic Research Foundation, Marshfield, Wisconsin, USA; [3]CIDES, Porto, Portugal

Abstract

The aim of this study was to understand and analyze dental information needs of medical providers at the Portuguese National Health System and register their opinion of an integrated medical-dental electronic health record. The majority of the providers considered the implementation of such integrated system useful to support bi-directional flow of patients' health information between medical and dental providers.

Introduction

The feasibility of integrating medical and dental patient data should be considered in a holistic patient care approach due to the oral-systemic relations of several pathologies [1, 2]. The aim of this study was to analyze medical providers' core dental information needs and to get their opinion of an integrated medical-dental electronic health record in their workflow.

Methods

A twelve-question survey based on previously reported open/closed ended interviews [2] was distributed (face-to-face) to a sample of 346 physicians in Primary Health Care Centers and in St. Teotónio Hospital, in Viseu, Portugal. The survey focused on medical providers' dental information needs, referral to dentists and their opinion and considerations about the use of an integrated medical-dental electronic health record in their workflow.

Results

There was a 51% overall response rate. Eighty-nine percent of the respondents that included Otolaryngologists, general physicians and pediatricians considered bi-directional flow of patient data between dental and medical care providers essential to provide effective medical care. There was a significant difference in the attitudes of the physicians towards incorporating the different components of the dental data. About 49% of the respondents requested a dental appointment monthly or less, and 14% reported weekly. It was seen that 92% recognized the significance of implementing an integrated medical-dental electronic health record to facilitate access to dental data, specially the "oral health status" (Figure 1). About 58% of respondents requested dental clinical history and dental diagnosis to be displayed in the integrated dental and medical record. Furthermore, about 59% of respondent considered problem list to be significant. Pulmonologists, Internists and Physicians with fewer years of medical practice tended to devalue the sharing of this kind of information.

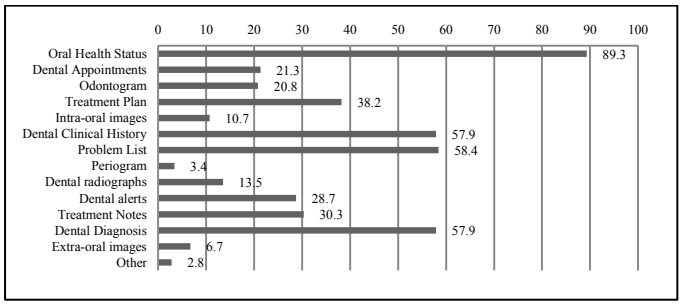


Figure 1. Medical provider's dental information needs

Conclusion

The results of this study indicate that implementation of a secure and confidential integrated medical-dental electronic health record may facilitate bi-directional access to relevant patients data, contributing to a more effective and efficient medical treatment. Capturing structured dental diagnosis in the dental practice is warranted since this was one of the top information that the medical providers find it important. A quantitative analysis of the advantages vs. disadvantages of an integrated medical-dental electronic health record environment can further be conducted to explore the feasibility of such an environment.

References

1. Powell V, Din F, Acharya A, Torres-Urquidy M. Integration of Medical and Dental Care and Patient Data. Hannah K, Ball M, editors. London: Springer; 2012.

2. Acharya A, Mahnke A, Chyou PH, Rottscheit C, Starren JB. Medical providers' dental information needs: a baseline survey. Stud Health Technol Inform. 2011;169:387-91.

A Qualitative Study Exploring The Vulnerabilities Of Computerized Physician Order Entry Systems in the U.S. and Canada

Mary G. Amato PharmD MPH[1,2], Sarah P. Slight MPharm, PhD, PGDip[1,3], Tewodros Eguale MD, PhD [1,4], Andrew C. Seger PharmD [2], Diana L. Whitney BS[5], David W. Bates MD, MSc[1,6], Gordon D. Schiff MD[1,6].

[1] The Center for Patient Safety Research and Practice, Division of General Internal Medicine, Brigham and Women's Hospital, Boston, MA, USA; [2] MCPHS University, Boston, MA, USA; [3] School of Medicine, Pharmacy and Health, The University of Durham, UK; [4] McGill University, Montreal, Quebec, Canada; [5] Baylor College of Medicine, Houston, TX, USA; [6] Harvard Medical School, 250 Longwood Ave, Boston, MA, USA.

Abstract: *Computerized physician order entry (CPOE) systems can prevent medication errors but they can also introduce new types of errors. We aimed to test the vulnerabilities of a wide range of CPOE systems to medication errors, and to develop a more comprehensive understanding of how their design could be improved. We found a high degree of variability in alerting between different systems, which represents a safety concern. Safeguards need to be put in place to ensure safer prescribing.*

Introduction: CPOE systems with embedded clinical decision support (CDS) can play a major role improving patient safety. However, these systems are not failsafe and can also introduce new types of errors. The Institute of Medicine report *Health IT and Patient Safety: Building Safer Systems for Better Care* recommended that specific examples of potentially unsafe processes and risk-enhancing interfaces be identified and shared amongst the health IT community. This study aimed to test the vulnerabilities of a wide range of CPOE systems to medication errors, and to develop a more comprehensive understanding of how their design could be improved to advance patient safety.

Methods: As part of a National Patient Safety Foundation-funded project, we examined 13 unique leading vendor and homegrown CPOE systems in diverse organizations in United States and Canada. Typical users at each of 16 sites were asked to enter 13 different erroneous orders on test patients in the usual and customary way, and where necessary perform workarounds. A research pharmacist and research assistant independently observed test users enter each order, and rated the ease or difficulty of these entries using standardized operational definitions. An Excel file was created and detailed descriptions of users' observations and verbalizations were recorded. Overarching themes relevant to interface design and usability / workflow issues were identified, reviewed as a group, and any discrepancies resolved by discussion.

Results: We found these systems often failed to detect these errors. Firstly, the generation of electronic alert warnings varied widely between systems, and appeared to depend on how the order information was entered into the system (i.e., in a structured or unstructured way); whether a specific alert functionality (e.g., duplicate-drug checking) was operational in the system; and which drugs or drug combinations were included in the CDS algorithms. Secondly, the wording of alert warnings was found to be confusing, with unrelated warnings appearing on the same screen as those more relevant to the current erroneous entry. Thirdly, the timing of alert warnings differed across CPOE systems, with many dangerous drug-drug interaction warnings displayed only *after* the order was placed (e.g. both Imdur® (isosorbide mononitrate) and Revatio® (sildenafil) had been entered and the order signed off in two systems). Fourthly, alert warnings varied in their level of severity in different systems and even within the same institution (outpatient vs. inpatient system). Finally, users developed workarounds to avoid getting duplicate drug alert warnings, such as entering the brand name of the drug for the morning dose and the generic name for the evening dose.

Conclusion: We found a high degree of variability in ordering and alerting between different CPOE systems, which creates major vulnerability. Detailed qualitative analysis of both the observed CPOE functionality and users think-aloud comments provided rich insights into the ways both systems and users were susceptible to ordering errors. System developers and users need to assess these findings and build in safeguards to ensure safer prescribing for patients.

Public Perceptions of Privacy and Healthcare Quality Effects of Electronic Health Records

Jessica S Ancker, MPH, PhD,[1] Samantha Brenner, MD, MS,[2]
Michael Silver, MS,[1] Joshua Richardson, MLIS, PhD[1]

1. Weill Cornell Medical College, Department of Healthcare Policy and Research, New York, NY
2. Stanford Hospitals, Department of Internal Medicine, Stanford, CA

Abstract: *Our objective was to track consumer perceptions of quality effects and privacy risks associated with health information technology during the initial years of the federal electronic health record (EHR) incentive ("meaningful use") program. National random-digit-dial telephone surveys were conducted annually for three consecutive years, 2011 through 2013. During these years, consumers became more likely to report having a physician who used an EHR. Over time, fewer people expressed concern about privacy risks of electronic information, but simultaneously, slightly fewer endorsed the statement that EHRs would improve care. In all years, consumers whose doctors used EHRs were more confident that EHRs would benefit them. Public opinion about EHRs may be evolving over the three-year period coinciding with the beginning of "meaningful use."*

Introduction: Surveys have suggested that much of the American public agrees with federal policymakers that electronic health records (EHRs) will improve the quality of healthcare. Nevertheless, public concerns about risks to the privacy and security of electronic medical information remain common. Our objective was to track consumer perceptions of quality effects and privacy risks associated with health information technology during the initial years of the federal EHR incentive ("meaningful use") program.

Methods: National random-digit-dial telephone surveys were conducted annually for three consecutive years, 2011 through 2013. Both landlines and cell phones were sampled, and the sample size of 1000 provided a 3.1% margin of error each year.

Results: Response rates for each year of the survey ranged from 63% to 64%. In 2011 and 2012, 64% of respondents reported seeing at least one doctor with an EHR, and this percentage rose to 70% in 2013 ($p = .003$). In the first two years of the survey, 66% of respondents endorsed the belief that EHRs would improve the quality of healthcare they received, but this percentage dropped to 61% in 2013 ($p = .02$). Also in the first two years of the survey, the percentage who believed that EHRs would worsen the privacy and security of medical information remained approximately stable (47% and 50%), but dropped to 41% in 2013 ($p < .001$). In each of the three years, consumers who had experience with doctors using EHRs were more likely to believe EHRs would improve healthcare quality than those whose doctors did not use EHRs (e.g., 64% vs. 53% in 2013; odds ratio = 1.5; $p = .04$). In 2013, consumers whose doctors used EHRs were less likely to believe that EHRs threatened privacy and security (38% vs. 55% in 2013; odds ratio = 0.5; $p = .001$). This association did not appear in previous years.

Discussion: During the early years of the meaningful use program, consumers became more likely to report having a physician who used an EHR. Consumers with this exposure to EHRs were more confident that EHRs would benefit them. Both confidence in the quality effects of EHRs and concerns about privacy risks became somewhat moderated over time. Although causal relationships cannot be determined definitively from repeated cross-sectional surveys, it is possible that increased familiarity with EHRs over time is associated with evolution in public opinion about health information technology.

Emergency Department Information System Selection: A Structured Approach

Christine K Anderson, RN, BSN; Graduate Student, Clinical Informatics and Patient-Centered Technology;
University of Washington, Seattle, Washington

Project Aims: 1. Identify key elements in an effective emergency department information system (EDIS). 2. Use a participatory approach, including emergency department (ED) physicians and staff, to rank and assign weights to the key elements of an EDIS. 3. Use the information to develop a vendor assessment tool for the evaluation and selection of an EDIS for a critical access hospital in rural Alaska.

Background: Federal healthcare legislation requires meaningful use of a certified electronic health record (EHR) for a healthcare organization to receive maximum reimbursement for their services.[1] Urban and rural hospitals alike are expected to meet these standards by 2015 or face payment penalties. Multiple professional organizations recognize the EHR as a support to safer and more efficient patient care. Lack of efficient and usable EHRs is identified as a primary reason for delayed adoption.[2] Software designers and developers must consider inclusion of fundamental principles of design identified by professional health information organizations in their EHR product development. Including staff in the EDIS selection process helps build support of the product and its effective use.[3] Use of these fundamental design principles along with the inclusion of clinical staff in system selection of an EDIS will support successful selection and implementation.

Methods: A survey-based field study was conducted to develop a rubric for EDIS selection. The survey was developed following review of the literature identifying essential features of an effective EDIS and finalized with input obtained from experienced EDIS users via semi-structured interviews. The survey was available via *Catalyst*, an electronic survey application, and a paper format, to increase the survey response rate.

Results: With a survey response rate of 51% (20 of 39 eligible respondents) the essential features of an effective EDIS were identified and evaluated by clinical staff scoring each item as most important to least desired. Top 10 feature ratings came out of staff survey responses, (a) being most important and carrying a weighted score value of 10, and so on.

Criteria	Criteria Score	Weighted Score	TWCS*
a) Rapid access to past medical history using an interface between EHR systems, if needed		10	
b) Multi-user platform allowing more than one simultaneous user per chart		9	
c) Supports bedside data entry & use of mobile devices for EHR access		8	
d) Intuitive system design requiring minimal initial user training		8	
e) Electronic WhiteBoard for tracking patient status and diagnostic results		6	
f) Patient education materials linkage for diagnoses, medications and discharge planning		5	
g) Flexible EDIS platform for facility specific set up and use		4	
h) Embedded CDS processes to support evidence-based and patient safety initiatives		3	
i) CPOE functionality specific to ED processes		2	
j) Coding & billing linkage to documentation supporting efficient & optimal reimbursement		1	

* Total Weighted Criteria Score (TWCS = Criteria Score X Weighted Score)

Vendor	a	b	c	d	e	f	g	h	i	j	Vendor Total Score
1											
2											
3											

Conclusions: Implementation of an EDIS is in the future for all facilities that have yet to move into the digital age. Selecting an EDIS is a serious process for careful consideration. Patient safety and staff efficiencies are called into play when making this significant process change. Allowing staff to be part of the product selection helps engage them for successful implementation. A needs assessment and vendor assessment tool supports the objective and thoughtful process of vendor selection. The vendor assessment tool developed from the results of the staff survey can be utilized to for this purpose.

References

1. HealthIT.gov. 2013. Available at: http://www.healthit.gov/providers-professionals/meaningful-use-critical-access-hospitals-and-other-small-hospitals. Accessed August 21, 2013.
2. HiMSS Usability Task Force. *Defining and Testing EMR Usability: Principles and Proposed Methods of EHR Usability Evaluation and Rating*. HiMSS. 2009.
3. McDowell, S.W., Wahl, R., & Michelson, J. Herding Cats: The Challenges of EMR Vendor Selection. *Journal of Healthcare Information Management*. 2003; 17:3; 63-71.

Pediatric Dose Range Checking with Hierarchical Rules to Provide Value Added Alerts

Charles H Andrus, MHA [1], Kevin O'Bryan MD [1], Patrick Feldman RPh [1], S. Paul Hmiel MD, PhD [2], Phillip Asaro MD [2], Feliciano Yu, MD, MSPH [2,1]

[1] St. Louis Children's Hospital, St. Louis MO; [2] Washington University School of Medicine, St. Louis MO

Abstract

Medication dosing rules for infants and children depend on many parameters, including patient weight, age (chronologic, and for infants, gestational), and indication, with per dose and daily limits. We describe a robust dose range checking system that allows for complex, hierarchical rules, while allowing near real time monitoring of alerts. Implementation was associated with decreased alert override rates, an increased acceptance of dosing recommendations, and an additional prescribing check on high risk medications.

Introduction

Clinical decision support (CDS) engines provide a mechanism to utilize the data elements of the EMR to provide relevant guidance to clinicians providing complex care. These complexities are exemplified by medication dosing rules for infants and children[1], as the clinician needs to consider age (chronologic age and for infants, gestational age), weight, indication, and per-dose and daily maximal dose limits, for many high risk drugs. We describe a robust dose range checking system that allows for complex rules, [1]systems (EMRs), while simultaneously limiting nuisance alerts.

Methods

All drug orders were tagged with a unique RxNorm code within the EMR. Dosing rules were developed from existing pediatric clinical pharmacology resources, by clinical pharmacists in collaboration with physician experts. Analysis of these rules revealed that any dose check could be represented as a combination of two types of rules: criterion rules and dose range rules. The rules were translated into criterion rules contained a criterion type such as flat dose, total daily dose, weight -based dose, or modifications based on gestational age or renal function. Criterion rules were developed with parent and child rules allowing complex dose-rule hierarchies including multiple input modalities. When a dose range check rule was applied to an order, it could result in any of 3 alert types: warning (user can proceed without a comment), soft stop (user must comment to proceed), hard stop (user cannot proceed).

Clinical Decision Support (CDS) logic for the dose range checking system is represented in Arden Syntax, and is separated from clinical content, which is held in tables within the EMR database. Thus alert messages and rules can be built and customized without MLM code changes by using a custom configuration tool. A new criterion rule type would require only one new Arden Syntax Medical Logic Module (MLM) to provide the logic of that specific criterion type. This approach provides extensibility and ease of maintenance. The .NET configuration tool further allows a user to aggregate historical data and model future dose rules to assess predicted fire rates. A real-time dashboard assesses alert effectiveness and user response. The initial trial consisted of twenty-three commonly ordered drugs.

Results

After 15 months, only 24% of dose range alerts were overridden. For acetaminophen, the sensitivity of alerts increased from 5.2% to 100%, and the percentage of alerts changing provider practice rose from 43.9% to 97.8%. Overall the number of alerts per day for acetaminophen decreased from one per day to one every three days. The next most frequently alerted medication was morphine, accounting for nearly 17% of the total; of these, 7.8% were for orders exceeding recommended maximal doses.

Conclusion

Creating robust dose range checking requires careful planning to develop effective dose rules. Utilizing a real-time dashboard to monitor alerts helps assess and reduce alert fatigue. Quick changes made by content experts (pharmacists and physicians) with the custom configuration tool were vital to provider acceptance. Further development is necessary to handle more complex if/else hierarchical dose rules.

References

1. *Scharnweber C. Lau BD. Mollenkopf N. Thiemann DR. Veltri MA. Lehmann CU Evaluation of medication dose alerts in pediatric inpatients. International Journal of Medical Informatics. 82(8):676-83, 2013*

Implementing a wireless-distributed EMR for a traveling student-run global health clinic

Emeka C. Anyanwu, MD[1]; Cheryl Thompson, PhD[1]; John T. Gale, PhD[2]
[1]Case Western School of Medicine, Cleveland, OH; [2]Cleveland Clinic, Cleveland, OH

Introduction

Since 2009, the Peru Health Outreach Project has offered students and health professionals in Cleveland, OH the opportunity to provide care in the Sacred Valley Region of Peru. Formerly the project made use of paper charting, which was later transcribed in to an electronic spreadsheet. In 2011, the group decided that it could benefit from an electronic record workflow and began to pursue options.

Early Considerations & Constraints

In considering electronic medical record solutions the group took in to consideration the following requirements:

- Independence from stationary power and network infrastructure was essential
- As a volunteer/student-run initiative the solution needed to be extremely cost efficient
- The solution needed to provide data security for our patients and their protected health information (PHI)
- Easy integration into the clinic's dynamic workflow with minimal to little training
- Platform and device agnostic, as the devices used would be volunteer owned and operated

OpenMRS, an open-source medical record system led by the Regenstrief Institute and Partners In Health fit these criteria and allowed us to prototype and deploy our system rapidly

Taking OpenMRS Mobile

To achieve necessary mobility the OpenMRS stack was deployed on a laptop equipped with an additional external battery. Consumer-grade wireless routers were modified to be powered via USB. An alternative firmware installed on the routers allowed the use of wireless distribution system (WDS) protocols to distribute the network over a greater range by linking the routers together creating a mesh network. Data security was achieved by disk and network level encryption.

Our Experience

Over the course of a month more than 1500 unique patient encounters were logged at over a dozen clinical sites. Only a single day of downtime was experienced. EMR kiosks were setup at patient intake, physician encounter, and pharmacy encounter stations. Immediately, the group recognized the gains in efficiency and data quality, including:

- Weekly report generation allowed the tracking and anticipation of medication supplies
- Data workflow adjustments could be made to encounter forms each day as the group saw fit
- Patients requiring further follow-up were more readily documented and tracked
- Data entry for IRB approved research was facilitated by a computer-based workflow

Difficulties & Lessons Learned

Considering the group's success, several critical lessons were learned during this experience:

- There are issues with portability between server computers and software versions of OpenMRS
- Incremental backups of the OpenMRS database are simple to create and restore to the same machine with basic SQL knowledge.
- Wireless network penetration can be greatly improved using mesh-linked consumer routers
- Similar traveling implementations could benefit from a mobile device ready interface as these devices offer greater battery life and portability.
- Installation and maintenance of OpenMRS requires intermediate-advanced technical knowledge, making implementation difficult for even a small traveling clinic
- The OpenMRS project would benefit from continuing its efforts to simplify the implementation process for newcomers through improved documentation

Protecting Patient Data and Maintaining Site Autonomy: Managing Project Access in a Multi-Site i2b2 Database

Nate C. Apathy[1], Abu Saleh Mohammad Mosa[2,3], Kelly J. Ko, PhD[1],
[1]Cerner Corporation, Kansas City, MO; [2]Institute for Clinical and Translational Science, School of Medicine, University of Missouri, Columbia, MO; [3]Informatics Institute, University of Missouri

Abstract

Organizations often establish an enterprise data warehouse then develop smaller databases for specific projects. However, doing so increases the need to manage multiple databases, including access. We developed a secure web application leveraging organizational parameters within the i2b2 Project Management tool to divide users and projects by organizations or sites. Through the secure web application users can now manage access at the project level instead of relying on the hosting institution to manage user accounts.

Introduction

Informatics for Integrating Biology and the Bedside (i2b2) is an open source tool developed by Partners Healthcare, Inc. and the NIH, to help manage clinical data for research purposes. Commonly, organizations will establish an enterprise data warehouse containing all relevant clinical information and then develop smaller databases for individual projects. However, as contributing institutions and projects increase, so does the need for data protection and decentralized management of data access. In order to address this issue, we developed an approach to manage access to a large multitenant database with various disparate projects and unique user groups.

Architecture

Through a secure web application (Cerner i2b2 User Management Tool), institutions can assign user permissions, provide access relevant to the users' role, and monitor user activity. This can be managed across individual users as well as multiple projects at the same site. By leveraging organizational parameters already established in the opensource i2b2 Project Management tool, site administrators can manage i2b2 access for only their institution or projects, while protecting other projects and data associated with other sites, investigators, or data domains (Figure 1).

By democratizing user management, we remove the onus of user maintenance and provisioning from a single, centralized i2b2 resource and allow project teams to maintain project-level access. Project-specific resources (i.e., principal investigator) can now control user access at the project level, rather than having those responsibilities rest in the hands of a centralized resource who may be managing multiple databases, to provision and manage user accounts on their behalf. Doing so helps to expand knowledge of i2b2 at the site level and also increases the efficiency of account maintenance.

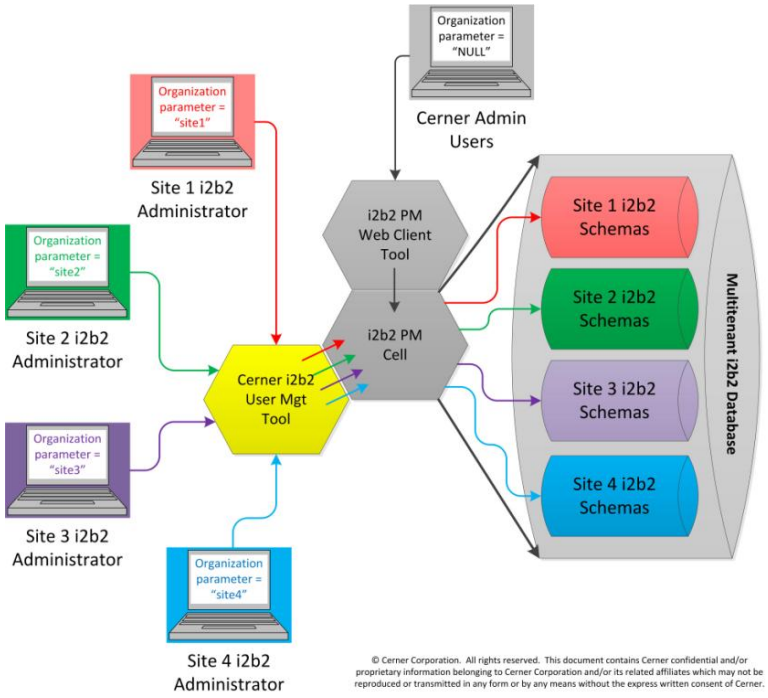


Figure 1. Schematic illustration of Cerner i2b2 User Management Tool architecture.

Conclusion

Although there are many different approaches to protect patient data, managing roles and subsequent access is particularly important when data reside in a single multitenant database. However, managing access across multiple projects through a centralized resource may prove difficult, especially as the number of projects increases. Through developing a secure web application, we are able to manage access based on organizational affiliation as well as role at the project level.

References

1. Murphy SN, Mendis M, Hackett K, Kuttan R, Pan W, Phillips LC, Gainer V, Berkowicz D, Glaser JP, Kohane I, Chueh HC. Architecture of the open-source clinical research chart from Informatics for Integrating Biology and the Bedside. AMIA Annu. Symp. Proc. 2007;11:548–52.

Experimental Protocol to Assess Comprehension and Perceived Ease of Comprehension of Tailored Health Infographics Compared to Text Alone

Adriana Arcia, PhD, RN[1] & Suzanne Bakken, PhD, RN[1,2],
[1]School of Nursing and [2]Department of Biomedical Informatics, Columbia University, New York, NY

Introduction

As part of the Washington Heights/Inwood Informatics Infrastructure for Comparative Effectiveness Research (WICER) project (R01HS019853), we created a series of infographics to return self-reported health data back to the >5,800 survey participants in an easily-comprehensible and actionable format. The purpose of this abstract is to describe an experimental protocol for the assessment of comprehension and perceived ease of comprehension of tailored infographics as compared to text alone.

Experimental Protocol

Principal aims are: 1) to evaluate the efficacy of infographics (experimental condition) compared to text alone (control condition) at aiding participants' comprehension of personal health status in the context of community/aggregate data and/or national standards; 2) to determine if participants' ratings of ease of comprehension are associated with the mode of information presentation (infographics vs. text alone); and 3) to determine what effect(s), if any, age and health literacy level have on comprehension.

Table 1. Stratification by age group and literacy level

Age group	Inadequate health literacy	Marginal + adequate health literacy
18-60	*n* = 36	*n* = 36
>60	*n* = 36	*n* = 36

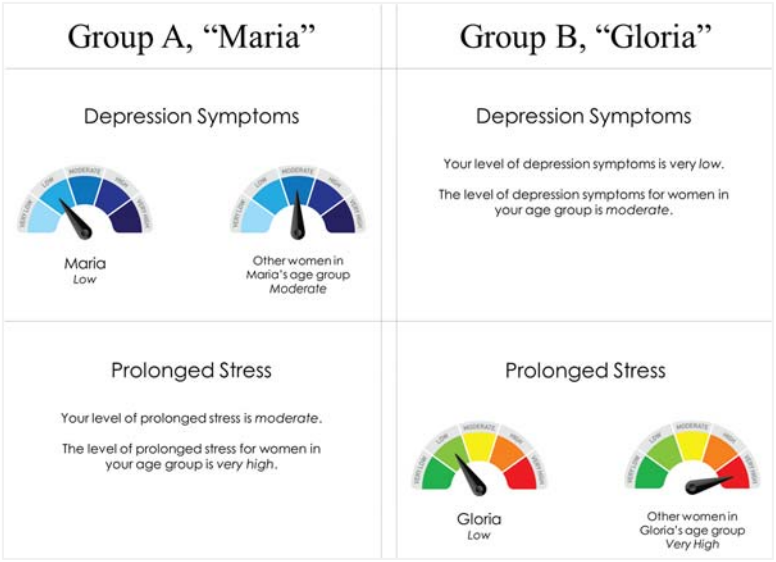


Figure 1. Visually and conceptually similar paired items

We will conduct a controlled trial with stratified randomization. Participants (*n* = 144) will be stratified by age group (18-60 vs. >60) and by health literacy level (inadequate vs. marginal + adequate) (see Table 1) prior to randomization to four groups: A1, A2, B1, and B2. Group A will serve as group B's control, and vice versa; group 1 will see text first, group 2 will see infographics first. For example, group A will be presented with an infographic on depression symptoms and text-only on prolonged stress while group B will be presented with text-only on depression symptoms and an infographic on prolonged stress (see example in Figure 1). Decks of infographic and text-only slides will be tailored for participants with their own survey data. For each slide, participants will complete one or more comprehension questions (e.g., "Is your BMI category underweight, normal, overweight, or obese?") and a rating of the ease of comprehension of the item ("Very difficult" to "Very easy" to understand). Summated comprehension and ease of comprehension scores will be analyzed with *t*-tests; a balanced two-way analysis of variance design will be used to test for main and/or interaction effects of age and health literacy on comprehension scores.

Discussion

The study design features, especially the use of visually and/or conceptually similar matched pairs of infographics, will be employed to maximize scientific rigor and minimize systematic bias.

Acknowledgments

Study supported by Washington Heights/Inwood Informatics Infrastructure for Community-Centered Comparative Effectiveness Research (R01HS019853) and WICER 4 U (R01HS022961). Dr. Arcia is supported by T32 NR007969.

Finding Different Types of Medical Conditions: From Data Generation to Automatic Classification

Nathan Artz[1], Jinho D. Choi, PhD[2], Stephen Doogan[1]
[1]Real Life Sciences, New York, NY; [2]Emory University, Atlanta, GA

Abstract

One of the challenges analyzing medical conditions in unstructured data is in determining whether they are reported as indications or side effects. Classifying different types of medical conditions enables a deeper understanding of disease manifestation and treatment risk, helping further research in this field. In our study, we collect medical forum posts, annotate different types of medical conditions, and classify medical conditions into indications and side effects using natural language processing and machine learning techniques.

Introduction

When analyzing texts from patient reporting systems (e.g., FAERS, social media), researchers often utilize medical databases such as MedDRA to produce ranked frequency counts of medical conditions[1]; however, this approach fails to differentiate treatment indications and side effect conditions, which is a critical part of assessing treatment outcomes. In this study, we aggregated and annotated social media posts from several online medical forums. We have also developed a novel system to classify mentions of conditions into indication and side effect groups.

Data Creation

We aggregated online medical forum posts from 22 different sources, and normalized them using ETL extraction infrastructure. For this study, we focused on a class of antibiotic drugs, identifying 19,313 posts with mentions of the target treatment entities. We then used a custom medical dictionary (MedDRA in addition to verbatim terms that were manually selected) to identify 5,179 unique conditions across this dataset, and manually annotated a dataset of 760 sentences that contained at least one condition and treatment. Each sentence was then given a binary label indicating whether the condition played the role of an "indication" or "side effect" of the given treatment. Finally, specific treatment and condition mentions in text were replaced by generic text labels (i.e. _TREATMENT, _CONDITION) to prevent overfitting to the antibiotic drug class.

Automatic Classification Approach

Various combinations of syntactic and semantic features are extracted from dependency graphs generated by a NLP toolkit, ClearNLP[2], and used for generating a statistical model trained by support vector machines (SVM). This statistical model is used for classifying medical conditions in raw texts. Figure 1 shows the overview of our approach.

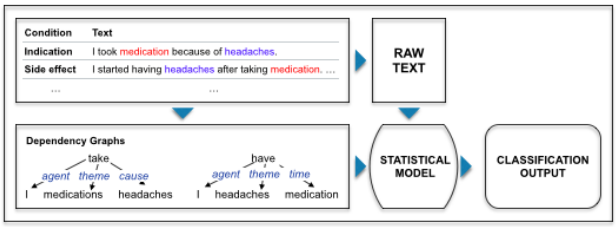


Figure 1. The overview of our classification approach.

Results

We used an SVM classifier with 5-fold cross validation and averaged the outcomes of the folds to determine the F1 scores (see Figure 2).

We manually reviewed some of the most impactful features of the SVM to see which are most important when differentiating the two classes of conditions. For conditions playing the role of indication, we see interesting features such as a Bigram(*treatment*, for) and syntactic_parent(*condition,* for) as in "on steroids for the inflammation". For side effects, we observe features such as SyntacticParent(*condition*, make) or SyntacticParent(*condition*, give) as in "treatment made me sick" or "the drug gave me an allergic reaction" which heavily influence the decision boundary of the SVM.

Experiment	Features	F1 Score
1	Baseline: Bag of Lemmas	78.23
2	Bag of Lemmas + Syntactic Features	84.26
3	Bag of Lemmas + Syntactic Features + Syntactic Roles	83.3

Figure 2. F1 scores for three experiments using different feature sets

Conclusion

We believe this is a feasible approach to differentiating the roles of medical conditions. This approach is suggested for use in spontaneous reports where many of the conditions can be assumed to play side effect or indication roles, versus other texts where the relations between treatment and conditions could be highly variable. Our study can enable researchers and pharmaceutical companies to better monitor health risks for patients by providing, for instance, a means to filter out instances of indications and focus on side effects related information in spontaneous reports. In future studies, we will use use our model to extract indications and side effects across various drug classes and compare our results to existing drug labels.

References

1. Gurulingappa H, Toldo L, Rajput AM, Kors JA, Taweel A, Tayrouz Y. Automatic detection of adverse events to predict drug label changes using text and data mining techniques. Phar. Drug Saf. 2013;22(11):1189-94.
2. Choi JD, McCallum A. Transition-based Dependency Parsing with Selectional Branching. ACL'13. 2013; 1052-1062.

Linking Provider Documentation to Handoff – the Status View

Phillip Asaro MD [2], Charles H Andrus, MHA [1], Kevin O'Bryan MD [2], Feliciano Yu, MD, MSPH [2]
[1] St. Louis Children's Hospital, St. Louis MO
[2] Washington University School of Medicine, St. Louis MO

INTRODUCTION: The road to effective implementation of electronic medical records remains challenging in most settings. The billing document use-case often trumps clinically oriented use-cases and leads to charts bloated with text that is perceived as necessary for reasons other than clinical care. Effective handoff of patient care from provider to provider is critical to quality patient care and various tools have been developed to assist in the transfer of useful information during handoff. These tools are typically not integrated with provider documentation and require duplicate documentation, further frustrating clinicians.

GOAL: Our overarching goal is the development of a functional framework for provider documentation that supports efficient workflow and optimal information flow, utilizing appropriately visible and enduring information components that integrate patient-care planning, documentation tasks and handoff.

METHODS: We are designing a documentation framework for a complex academic environment in which many of the patients are cared for by multiple clinical services and within each clinical service there may be multiple provider types working in a cooperative hierarchical relationship.

We began by eliciting information flow requirements from a variety of clinical services. We compiled a list of the standard components of clinical notes and supplemented this with additional components required to support the elicited information flows. Documentation tasks considered included History and Physicals, Consult Notes, Progress Notes, Discharge Notes, Off-Service Notes, Rounding Worksheets, and Handoff Sheets. We characterized each information component according the its scope, described in terms of which document types the component will appear in, which provider types would create, edit, and/or consume the information in that component, and durability of information in that component. Then through brainstorming between a software architect and clinicians (the authors), we developed a vision of an intuitive interface for maintaining the information in these components.

THE VISION: The functional framework that emerged is a single-screen integrated interface with context-oriented views. The interface consists of movable widgets that can be selected and arranged as needed according to the current view. Views and widget functionality are further customized according to the needs of each clinical service. Each of the widgets is designed independently with functionality to allow intuitive entry of the data needed for that specific component while meeting billing and regulatory requirements. Differing needs of various services are handled primarily at the widget and view levels. Most of the information components contain information entered specifically by each clinical service, but a few components, related to discharge planning for example, may contain shared information across services.

Typical Use-Cases: When in History & Physical or Consult Note views, information components are selected and arranged approximating the layout of the relevant document. The provider enters the relevant information in each component. When the information is complete, a "Document Save" button is activated producing the History & Physical or Consult document. The following day, before and during patient-care rounds, the Status View is used. This view displays all information components arranged according to user preference. Durable information entered the day before remains in the widgets. Interim history is entered in the Interim History widget. New results are reviewed using the Result Viewer which also allows the user to select results for inclusion in the progress note for the day. Problem Assessments and Plans are updated as needed to reflect the current state of the patient and current decision-making. When ready to create a progress note for the day, the provider selects the Progress Note view, further tweaks any information as needed and activates the Document Save button to create the note. The progress note is thus a snapshot of the current state of the patient's condition and current decision-making related to that patient. Later that day during handoff, a handoff sheet is created utilizing the information from the current state at that time. We believe that this vision will lead to an intuitive user interface in an information-flow-oriented framework that will support management of clinical content via easily configurable and modifiable components.

Reducing Complexity of Breast Cancer Treatment Regimen Representation in Tumor Registries

Ravi V. Atreya[1], Mia A. Levy, MD, PhD[1]
[1]Vanderbilt University School of Medicine, Nashville, TN

Abstract

To rapidly improve clinical decision-making and implement new initiatives in healthcare, we must be able to constantly learn from data generated through the course of patient care. Cancer treatment regimens are complex and must be effectively identified in data found in the electronic medical record (EMR). Tumor registries are curated patient care records and can be used to generate gold standards to better analyze EMR data. In this study, we seek to assess the ability of various methods to reduce the complexity of the treatment regimens stored in the tumor registry. Regimens were stored as strings of successive treatment events and were simplified by removing repetitive treatments and by modeling the order of treatment events. This work demonstrates the ability of basic methods to simplify tumor registry treatment regimens and shows that as regimens grow more complicated, it is important to develop methods to represent them effectively.

Introduction & Background

The rapid learning cancer system concept provides an opportunity to rapidly improve clinical decision making and implement new initiatives to meet financial and quality challenges facing healthcare organizations. In order to accomplish this, it is vital to discover treatment regimens from often complex data in the electronic medical record (EMR). Tumor registries are patient records that curated by nurses and contain data on patients' treatment history. While these registries are difficult to maintain, they contain a wealth of organized data on patients' cancer history. Tumor registries, while complex, can serve as a tool to filter and understand even more multifaceted EMR data. In this study we assess methods that utilize the importance of treatment event sequence to identify and simplify the treatment regimens for patients with breast cancer.

Methods

4,160 records of patients with breast cancer undergoing 16,054 treatment events between 1999 and 2012 were extracted from the Vanderbilt University Medical Center tumor registry. A basic simplification of the data was conducted to merge chemotherapy and immunotherapy events due to shifting registry definitions. Character strings were used to represent treatment regimens where the treatment events (S - surgery, R - radiation, H - hormone, C - chemotherapy) were placed in order of occurrence. Complexity was assessed for two additional treatment regimen simplification methods. First, repeating treatments (i.e., SSSC) were reduced to a single treatment event (i.e., SC). Another method utilized the ordering of the first occurrence of the four therapies (i.e., for SSSC, S occurred before C) to organize the treatment sequences. These methods were assessed by how the new representations were able to simplify the treatment regimens by assessing the count as well as distribution across patients.

Results and Discussion

This work demonstrates how treatment regimens extracted from the tumor registry can be simplified by focusing on event ordering to serve as an effective gold standard for complex EMR data. Using only the order of the first occurrence of treatment events demonstrated the greatest reduction in complexity (figure 1). From 1999 to 2012, the number of treatment regimen representations has doubled, along with the number of new breast cancer patients seen per year. In order to keep pace with the growing complexity of cancer care, it is important to develop gold standards of various levels of complexity to aid in the analysis of EMR data. The methods described here are limited by not incorporating the time between events. Additional work will utilize clustering algorithms to demonstrate the complexity introduced by adding time between treatments.

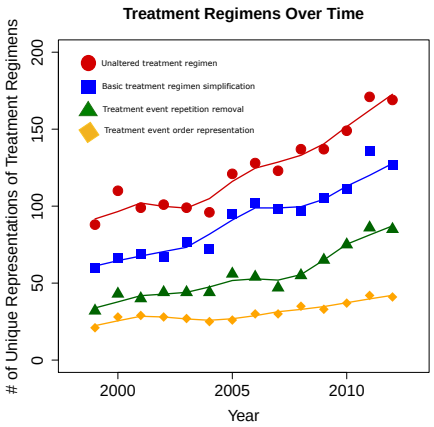


Figure 1: Number of unique representations of treatment regimens over time has increased. Number of patients seen has also increased from 209 in 1999 to 427 in 2012.

Picking a proxy on the web: Interactive Patient Interview Module for Health Care Proxy Documentation

Adarsha S Bajracharya, M.D., [1] Bradley H Crotty, M.D., M.P.H., [1] Hollis B Kowaloff, B.A., [1] Warner V Slack, M.D., [1] Charles Safran, M.D., M.S. [1]
1. Beth Israel Deaconess Medical Center, Harvard Medical School, Boston MA

Problem: When ill or incapacitated, patients may be unable to make their own medical decisions. A health care proxy (HCP) form is a legal document formalizing the appointment of a surrogate decision-maker. Prior studies show that patients who had given surrogate decision-making information to their providers lacked documentation of this in their medical records.(1) (2)Physician reminders alone are not associated with greater HCP appointment, but reminders plus mailings to patients of advance care planning forms and related materials have improved HCP appointment rates.(3)

Solution: We have developed an interactive interview to be available in a patient portal, through which patients can provide health care proxy information. Information provided by the patient will be stored in the patient's electronic medical record for clinician review. This module also has an integrated educational section, that answers common questions, provides background information about the importance of health care proxies, and advises patients how best to discuss their care preferences with loved ones.

Impact: Informed by prior experiences, we will measure the impact of our approach on rates of HCP appointments and the provision of completed of health care proxy forms. Through better information and documentation, we hope to improve care of all patients in line with their wishes.

Screen shot 2: Demographics

Screen shot 1: Patient options

REFERENCES

1. Yung VY, Walling AM, Min L, Wenger NS, Ganz DA. Documentation of advance care planning for community-dwelling elders. J Palliat Med. 2010 Jul;13(7):861–7.

2. Halpern NA, Pastores SM, Chou JF, Chawla S, Thaler HT. Advance directives in an oncologic intensive care unit: a contemporary analysis of their frequency, type, and impact. J Palliat Med. 2011 Apr;14(4):483–9.

3. Heiman H, Bates DW, Fairchild D, Shaykevich S, Lehmann LS. Improving completion of advance directives in the primary care setting: a randomized controlled trial. Am J Med. 2004 Sep 1;117(5):318–24.

Value Transparency in Health Information Exchange as a Mechanism to Enrich Patient Participation

Gina B. Baker, MSN[1], Shan He, PhD [1], Darren K. Mann [1] Pallavi Ranade-Kharkar, MS[1, 2], Jason Gagner, MBA[1, 2], and Sidney N. Thornton, PhD[1, 2],

[1]Intermountain Healthcare, Salt Lake City, UT; [2]University of Utah, Salt Lake City, UT

Abstract

Informing patients of the contributed value for each external datum obtained through Health Information Exchange (HIE) can increase their understanding of data exchange value, encourage their participation, and optimize data validity for clinical use. Values of external data, shown in dollars and patient time, can be inferred from the specific datum's relevance within a modeled care process and the associated treatment activity cost databases. The electronic Patient Visit Summary document can be extended to include itemization of the value contribution for each externally acquired datum pertinent to the encounter.

Introduction

Health Information Exchange (HIE) provides value to the healthcare system via shared data resulting in money and time saved. HIE, however, may also propagate data integrity issues as data are exchanged and integrated. To address these challenges, Intermountain Healthcare, in conjunction with the Care Connectivity Consortium[1], has developed novel HIE services to filter inbound data documents for new or amended data and to logically assess conflict and compatibility for the specific patient encounter[2]. The relevancy of the external data received and resulting value of the specific datum vary depending on the reason(s) and context for the current patient encounter. Patient engagement in HIE, specifically data review and stewardship, provides a way to validate information used in the electronic summary views[3]. To encourage patient engagement and promote data validation an additional section is included in the electronic Patient Visit Summary document. This section provides the value contribution of each data element integrated from an external source and looks much like a shopping receipt showing transactional savings. Recognition of this value presented in familiar units of dollars and time increases patient interest and results in their engagement to validate data thereby increasing its value for future encounters.

Methods

Connections were established between local HIE data processing services, a framework for understanding current patient encounter context, and a repository of treatment-based costs. From these connections logic was developed to compute the value of each externally acquired datum relevant to the current encounter. Then a novel section was appended to the Patient Visit Summary, listing the value contribution of each externally acquired datum in a way that patients and providers can discuss.

Results and Discussion

Prescreening of HIE data for de-duplication and relevancy to the current encounter is helpful to provide a clean record for the clinician and increases value for the use of HIE by patient and provider. While the value of the data changes depending on the clinical context of the current encounter, the section listing value of the data received in relation to the current encounter provides the opportunity for recognition of cost savings associated with HIE participation. This recognition of the role and value of HIE contributes to increased patient participation by verification or correction of data.

Conclusions

The value contribution of each datum originating from external sources can be quantified using time and money measurements and summarized for the current patient encounter. The itemized summary of integrated external data from HIE provides the clinician and the patient insight into the efficiency of the visit.

References

1. Care Connectivity Consortium: http://www.careconnectivity.org/. Accessed on March 12, 2014.
2. Ranade-Kharkar P, Mann D, Thornton S. Data adjudication architecture for health information exchange (hie): a case of adjudicating and storing hemoglobin A1c values. AMIA Symposium 2013.
3. Tripathi M, Delano D, Lund B, Rudolph L. Engaging patients for health information exchange. Health Affairs 2009;28(3):435-443.

Implementing SNOMED CT in Laboratory Specimen and Source Tables – If the Shoe Fits…

Pamela Banning, BS (Biology) MT (ASCP)^{CM}, PMP, Elva Knight, BS (Biology) MT (ASCP)
3M Health Information Systems, Inc., Salt Lake City, UT

Issue: Meaningful Use core measures and secondary use in public health reporting require the addition of the SNOMED CT terminology standard to legacy system embedded source and specimen lists. 'Sources' pertain to the anatomic site of concern (e.g. arm, wrist), while 'specimen' pertains to the form delivered to the lab (e.g. swab, aspirate, syringe). Facilities decide mapping scope based on content origination, use case, and system capability.
Method: Four sites submitted separate specimen and source lists in Excel format for terminology mapping. Resources included the SNOMED CT terminology content via the 3M Healthcare Data Dictionary, CliniClue freeware browser, and UMLS. SNOMED CT domain selection guidance was obtained from HL7 and Lab Communities of Practice[1] documents. Specimens were mapped to the SNOMED CT specimen domain. Sources were mapped to body anatomy, substance, or physical device/object domains. Laterality source modifiers were added as needed. Sites clarified ambiguous or nonspecific terms by a question and answer process.
Results: Specimen lists ranged in size from 32 to 111 terms, while source lists ranged from 243 to 995 terms. Both list types had data that crossed domains (e.g. wrist on specimen list; fluid on source list). Successful mapping occurred within all files, but with varying levels of completeness. Completion ranged from 75.7% - 96.7% in each file. Sites may build unconstrained content that is understood and used in their labs but lacks granularity or specificity for standardization (e.g. blue port, bagged, red lumen).

Table 1 - Mapping Examples across SNOMED CT domains

Site Term / SNOMED CT domain	**Body Structure/Substance Physical Device**	**Body Structure Option 2**	**Qualifier**	**Specimen**
Bronchial Washing, Bilateral	955009 Bronchial structure (body structure)	5926001 Structure of bronchial lumen (body structure)	257916006 Bilateral sampling (qualifier value)	122609004 Specimen from lung obtained by bronchial washing procedure (specimen)
Urine, Clean Catch	78014005 Urine (Substance)	--	--	122880004 Urine specimen obtained by clean catch procedure (specimen)

Additional complexity comes from multiple options. A source term of abdominal fluid could map to either:

- 113345001 – Abdominal structure (*body structure*)
- 83670000 – Peritoneal cavity structure (*body structure*)
- 32457005 – Body fluid (*substance*) + 113345001 – Abdominal structure (*body structure*)

By strict definition, a specimen term of abdominal fluid maps to 168139001 – Peritoneal fluid sample (*specimen*)
Conclusions: Full source description is frequently only achieved by combining SNOMED CT concepts (Table 1). Some terms required up to four concepts to provide clarification or best match. Many of the source terms include specimen types, morphologic abnormalities or collection procedures as identified by SNOMED CT hierarchy (abscess, cyst, washing, and lavage). Some mappings fall outside expected domains. The granularity of the terminology standard versus the site data are at different levels. Rebuilding site data may be required to meet standards requirements in the future.

References

1. Merrick, R. Lab Communities of Practice/Public Health Information Network. [Internet]. Sacramento, CA. 2010 Jul 10 [updated 2014 Feb 27; cited 2014 Mar 12] available at http://www.phconnect.org/group/laboratorymessagingcommunityofpractice

Analysis and Evaluation of methods of similarity in Nutritional Recommendations

Haroldo G. Barroso, Msc[1], Márcia Ito, M.D, PhD[2]
[1]Federal University of Maranhao, Sao Luis, MA; [2]IBM Research Brazil, Sao Paulo, SP

Introduction

A Nutritional Recommendation is understood as a source of information from the Nutrition expert, providing a therapeutic and educational treatment to the patient. Similar cases among patients require an understanding of a whole application domain (Nutrition and Related Pathologies) and their variables (clinical and personal data). For each case we can use some information that was retrieved from similar cases in order to make the nutritional recommendation. [1] This is done through methods of Similarity [2] based on an approach called Case-Based Reasoning. The divergence between the methods however, shows up in the axiom, which directly reflects the behavior of event and performance in the recovery of the same process, leading some methods to achieve further successes in relation to others.

Methods

There are a range of methods of Similarity, in this work, we use: Nearest Neighbor, Contrast model, Simple Matching Coefficient and Probabilistic Model. Although each method has different structures, all of them has the same goal, recover cases, which in turn will receive a nutritional recommendation. All methods are derived from a standard axiom defined by: $S(E, B) * W$

Where we have: S (Similarity Function, final result related to the performance of the method or amount of recovered cases), B (if previously indexed or old), E (Case of entry, to be compared) and W variable (weight, balance in a situation, it is strongly used in all tests). For Analysis and Evaluation of the performance of such methods, by default they use a technique called, Recall x Precision in a situation using 100 cases and assigning values factorial (n!) the weights, in order to validate the tests. This technique considers the points of greatest numbers x minor cases recovered in a situation where it is tested in each case by case method of similarity. [3] In our testing, we found considerable amounts of recovered cases (Σ), where: $\sum^{\text{nearest neighbor}} = \{78,1; 45; 43,1; 39,1; 37,4\}$; $\sum^{\text{contrast model}} = \{45; 43,1; 43,2; 43; 40\}$; $\sum^{\text{simple matching coefficient}} = \{54; 43,2; 43,1; 39,2; 37,6\}$; $\sum^{\text{probabilistic model}} = \{78,1; 58,9; 45; 43,2; 43,1\}$

Conclusions

In preliminary tests, it is concluded that the probabilistic model (Figure 1) reached a more cohesive result, recovering a greater number of cases and show that the methods of similarity have regular and stable behavior, leading to reflect on metrics recovery based not only on satisfactory results, as well as the regular results, avoiding a disparity in performance.

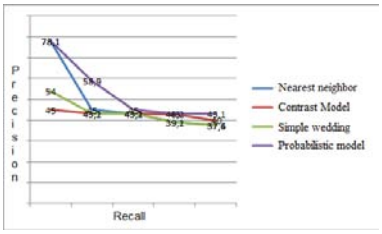


Figure 1. Performance graph of the methods of similarity

References

[1] HEIMBURGER, D. C.; ARD, J. Handbook of Clinical Nutrition. 4th edition. Elsevier. 2006.

[2]KOLODNER, J. **Case-based reasoning**. San Mateo: Morgan Kaufmann, 1993.

[3]LEE, L.G.L. Application of case-based reasoning to customer service. In**: Proceedings of the 3rd world congress on Expert systems.** pp.1143-1149, 1996.

A Proposed Protocol for Meaningful Use Audit Documentation

Robert C. Bell, MS[1], Fengwei Zhong, MD, MPH[1],
Debora J. Simmons, PhD, RN, CCNS[1], Adol Esquivel, MD, PhD[1]
[1]St. Luke's Health System, Department of Clinical Effectiveness and Performance Measurement, Houston, Texas

Abstract
In anticipation of Meaningful Use audits a system standardized, EHR agnostic documentation protocol has been developed to demonstrate compliance with the regulations.

Introduction
The Health Information Technology for Economic and Clinical Health Act (HITECH) put forth incentives for healthcare providers and hospitals to adopt electronic health record (EHR) systems. Along with adoption, "meaningful use" of these systems was specified as a set of goals in order to achieve improvements in care[1]. Meaningful Use started in 2011 and, as currently planned, is implemented in three stages over the course of many years, with each stage increasing in complexity. As a method of regulation of the distribution of incentives, CMS has partnered with Figliozzi and Co. to perform audits of providers and hospitals. There have already been cases where Meaningful Use incentives have had to be completely refunded; Health Management Associates of Naples, Florida reported returning $31M. Many participants are concerned with the lack of knowledge of what exactly triggers an audit[2] and how to demonstrate that the EHR system meets Meaningful Use requirements. To mitigate this risk, it is important to anticipate anything that could be asked for by auditors. To this end a comprehensive audit protocol has been developed that can be used to answer auditors questions, and provide any documentation to prove compliance with Meaningful Use.

Meaningful Use Audit Documentation Protocol
The Meaningful Use audit documentation protocol includes:

- Performance and pre-performance period reports and scorecards[3],
- Extensive supporting documentation for how the system meets the Meaningful Use measures
- The required security assessment and the final report and recommendations,
- Documentation of all Meaningful Use functionality as implemented in the providers' and hospitals' environment,
- Testing of the EHR to ensure that it meets the requirements of Meaningful Use
- NIST (National Institute of Standards and Technology) use case testing[4] performed in the hospitals' EHR system.
- EHR vendor supplied test scripts performed in the providers' and hospitals' EHR system, which verifies that the workflows indicated capture the Meaningful Use required data elements and report them properly.
- The sign off page by executive leadership in the organization.
- The documentation of attestations for Meaningful Use.

References
1. Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. *New England Journal of Medicine*. 2010;363(6):501–504.
2. Barr P. Auditable use. Meaningful use audits trip up some hospitals. *Hospitals & health networks / AHA*. October 2013:20.
3. Raiford R. Death, Taxes, and Meaningful Use Audits. 2012:1–9.
4. Lowry SZ, Quinn MT, Ramaiah M, et al. Technical Evaluation, Testing, and Validation of the Usability of Electronic Health Records. 2012.

A New Approach to Evaluating Mobile Applications in the Emergency Department: Extending Usability Testing and Clinical Simulation

Elizabeth M. Borycki, RN, PhD[1], Andre W. Kushniruk, MSc, PhD[2]
Judith W. Dexheimer, PhD[2]
[1]University of Victoria, Victoria, British Columbia, Canada; [2]Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

Abstract

There is a need to use new methodological approaches when evaluating mobile applications for the emergency department (ED). In this work we present one method that can be used to evaluate mobile device use in the ED.

Introduction

Evaluating mobile software applications in terms of their usability and effects on physician and nurse workflow can be difficult in most health care settings, but for evaluators it can be an especially complex undertaking when conducted in settings that are known for their dynamic nature such as the emergency department (ED). In this poster presentation we outline a new "in situ" clinical simulation approach that can be used to create simulated patient cases and to evaluate mobile phone and tablet software specific to an organization's ED visits.

Learning Objective

After reviewing this poster the learner will be better able to evaluate the usability of mobile devices for the ED.

Methodological Approach

In our work we focused on developing a new methodology that can be used to collect usability and workflow data to evaluate the effectiveness of mobile phone and tablet software in ED contexts. We extended Kushniruk and colleagues "in situ" methodology [1] used in physician office and hospital ward settings to collect usability and workflow data and have applied a clinical simulation component to the methodology so that participants in the studies fully experience the dynamic and time limited nature of a typical ED patient visit. In the first stage of the work both typical and atypical cases that are seen in the ED are identified by reviewing visit data and recording the range of patient visits arriving as non-urgent to urgent. Following this we identify typical ED rooms used to treat these types of cases. With the hospital ED organizational context in mind (i.e. the physical structure of the room), we then conduct a thematic analysis of the patient charts that represent each of the above mentioned patient visits to the ED. The thematic analysis is then used as the basis for the development of simulated patients that are similar to real-world ED patients. Therefore, organizational and patient contexts are similar to what a nurse or physician would encounter in an ED. As an added layer, to ensure the ecological validity of the simulated patient cases, we ask ED physicians to review the cases for their representativeness. Therefore, ED visits are used as inputs to the creation of simulated patient cases. Testing is then conducted in an ED room using the simulated patient cases.

Experiences to Date and Discussion

Participants will be trained on the use of mobile phone and tablet software using a face-to-face approach in a classroom setting. After the training is completed physicians and nurses will be invited to participate in the usability testing in a pediatric ED context. Once training is completed a video camera will be placed in the room to record user interactions and communication between the simulated patient (based on our work in reviewing ED visit records) and the mobile device. Screen recording software will be deployed to record participant interactions with mobile devices. The approach is expected to get at the nuances of pediatric care managed in conjunction with software support as compared to other forms of usability and workflow testing of software.

References

Kushniruk, Andre W., Elizabeth M. Borycki, Shigeki Kuwata, and Joseph Kannry. "Emerging approaches to usability evaluation of health information systems: towards in-situ analysis of complex healthcare systems and environments." *Studies in health technology and informatics* 169 (2010): 915-919.

Integrating Diverse HIV-associated Datasets via Semantic Harmonization

William Brown III, DrPH, MA[1,2], Chunhua Weng, PhD[1], David Vawdrey, PhD[1], Alex Carballo-Diéguez, PhD[2], Suzanne Bakken, PhD[1,3]
[1]Department of Biomedical Informatics, Columbia University, New York, NY; [2]HIV Center for Clinical and Behavioral Studies, NY State Psychiatric Institute & Columbia University, New York, NY; [3]School of Nursing, Columbia University, New York, NY

Abstract

The objective of this research is to integrate diverse HIV-associated research datasets. Thus, we have systematically identified and collected HIV-associated datasets based on NIH inclusion criteria and are semantically harmonizing variables across datasets by leveraging controlled vocabularies to formally represent synonymous variables.

Introduction

Integration of diverse HIV-associated datasets has the potential to increase knowledge by increasing the breadth of variables and statistical power critical for analysis, particularly for sub-groups. Thus, integrated analysis can guide the development of novel, high quality, effective interventions for disease prevention and treatment. This topic has been identified as an NIH research priority in the area of HIV-associated data science (RFA-MH-14-200), but few efforts have been made to combine data across HIV studies. To address this gap we have identified HIV-associated datasets based on the inclusion criteria outlined by NIH, and are employing methods of semantic harmonization.

Methods

Datasets were systematically identified using the advanced search form in Clinicaltrials.gov. Four hundred and fifty-three datasets were identified. Eight of 16 investigators who were contacted contributed a total of 17 datasets that met inclusion criteria, representing >4,300 subjects. The UMLS was used to identify controlled vocabularies for formal representation of HIV-associated neurocognitive and behavioral measures, diseases, laboratory results, and medications. Controlled vocabularies (CV) selected include: LOINC [>71,000 measures], SNOMED [>311,000 diseases/measures/medications], RxNorm [medications].

We started the variable formalization process by first standardizing the column, row, and heading format of the data dictionaries of all datasets. We then aggregated all data dictionaries into a variable harmonization table. To find variable related formalisms across different datasets, we used a semantic knowledge resource created by Columbia University, the Medical Entities Dictionary (MED), which is a large repository of medical concepts that are drawn from a variety of semantic knowledge resources including the New York Presbyterian Hospital and the UMLS (i.e. LOINC, SNOMED). The MED browser allowed us to simultaneously identify multiple semantically represented clinical formalisms. We also traversed the semantic structure of the MED to identify relational classifications to other potentially HIV-associated concepts (i.e. Parent nodes, Child node).

Results

The variable harmonization table includes the study ID, variable name, variable identifier or concept data entity (CDE), UMLS-CUI, MED code, LOINC ID, SNOMED ID, and RxNorm. These are further classified by NIH's four target variable classifications: Neuro-psychological, Psychosocial, Behavioral, and Biological/Biomarkers.

Conclusion

Next steps include continued progress towards formalizing >4,300 participants' data and completing the variable formalization table. We will also explore current ontology-based data integration methods.

Acknowledgements: Dr. William Brown III is supported by NLM research training fellowship T15 LM007079 and NIMH center grant P30 MH43520.

Development of an Antimicrobial Data Mart from the Electronic Medication Administration Record

Vladimir Bubelev PhD, Brian Myers BS, Jessica Johnston MS, Kurt Stevenson MD, MPH
The Ohio State University Wexner Medical Center, Columbus, Ohio, USA

Abstract: *Monitoring antimicrobial use is critical to the success of Antimicrobial Stewardship Programs (ASPs). We have created a data mart to automatically calculate antimicrobial use metrics based on the medication administration record (MAR) that is readily available to an ASP.*

Introduction: ASP and Information Warehouse (IW) at The Ohio State University Wexner Medical Center (OSUWMC) have partnered to provide access to antimicrobial use data in order to understand antimicrobial use patterns. Antimicrobial use significantly contributes to increasing rates of resistant pathogens and should be monitored at the local level to understand the relationship between antimicrobial use and emerging resistance.

Methods: An Antimicrobial Datamart (AD) was developed based on electronic medication administration records (MAR) and contains two main measures: days of therapy (DOT) and patient days on any antimicrobial (PD). In order to stratify reports by the fields of interest to ASP, the data was grouped by the following: ordering provider, MAR location, clinical service, route and site of administration, MAR date, and medication name. MAR and Admission/Discharge/Transfer (ADT) data were joined in order to report by location. To determine the rate of DOT, a simple proportion was constructed with the numerator as amount utilized and the denominator as number of patients eligible for antimicrobials. DOT is the number of days that a patient is receiving an antibiotic, regardless of dose, frequency, route or number of antimicrobials received on a given day. PD accounts for the receipt of multiple antimicrobials on the same day. When patients receive more than one antimicrobial on the same day, more than one DOT may be counted. Since antimicrobial consumption is directly proportional to the number of patients admitted to a hospital system, and as this number fluctuates, these measures are normalized by dividing by the number of patient-days multiplied by 1,000. Patient-days is the sum of all inpatients at midnight each day. For combining MAR and ADT only the date component of MAR taken date/time stamp for a given MAR was matched with the ADT event window record that spans midnight of the MAR taken date. This method could cause an under-count of DOTs as patients who received an antimicrobial during a given day and were then discharged before midnight of the same day would not contribute to the DOT count since they do not have a corresponding ADT record. Dynamic normalization was used.

Results: The data mart contains 1,477,483 MAR records from Nov2011 to Dec2013. Figure 1 shows an example of select antifungal DOT data.

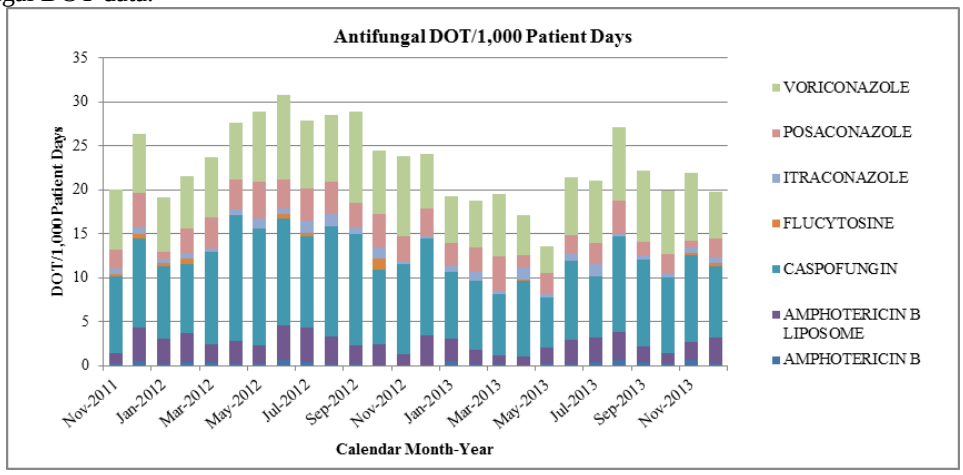


Figure 1. Example of select antifungal days of therapy (DOT)

Discussion: This is an important tool for quality improvement in healthcare as it provides a method for measuring antimicrobial use quickly and reliably allowing for quantification of the direct impact of ASP interventions.

Patient Screening Application to Identify Suitable Clinical Trials

Anca Bucur, PhD[1], Jasper van Leeuwen[1], Njin-Zu Chen[1], Brecht Claerhout[2], Kristof de Schepper[2], David Perez-Rey, PhD[3], Raul Alonso-Calvo, PhD[3], Kamal Saini[4]
[1]Philips Research Europe, The Netherlands; [2]Custodix NV, Belgium; [3]Universidad Politécnica de Madrid, Spain; [4]The Breast International Group, Belgium

Abstract

The population suited to be enrolled in a clinical trial is described by a set of free-text eligibility criteria. The assessment of the eligibility of a patient for a trial requires evaluating whether each criterion is satisfied by the patient information, and is often a time consuming and manual task. We propose a pragmatic and efficient solution to matching the two types of information to evaluate whether a patient satisfies the criteria of relevant trials.

Introduction

Our goal is to automate certain tasks of the patient screening procedure to enhance modern clinical trial recruitment. The application was developed to suit the needs of a large clinical research network in breast cancer. The solution covers the formalization of criteria and of other trial metadata and the efficient management of these representations. We rely on widely-adopted standards to represent the trial and patient information. Our semantic interoperability approach[1] provides shared semantics between formalized trial information and clinical data. This is leveraged by the application to implement automatic linkage and matching of trial criteria to patient data.

The Patient Screening Application

Figure 1 depicts the main components of the application. The front-end component enables the users to visualize available patients and their data, to evaluate for each patient whether they are eligible for one of the trials, to inspect the automatic evaluation results of each criterion and the evidence in the patient file, to update the patient information and to validate or override the recommendations of the system. The underlying Patient Screening service retrieves relevant trial information from the Trial Metadata Repository (TMR), such as the trial criteria and their associated formalism and execution logic. In the TMR criteria are mapped to templates and several formalisms and representations can be associated to a template. The Patient Screening service uses the Criteria Matcher to evaluate the criteria on the patient data retrieved from the Common Information Model (CIM). The CIM expands the set of initial concepts in the trial criteria based on the available domain knowledge representation provided by ontologies (e.g. SNOMED-CT) with relevant concepts whose instances are searched in the available patient data.

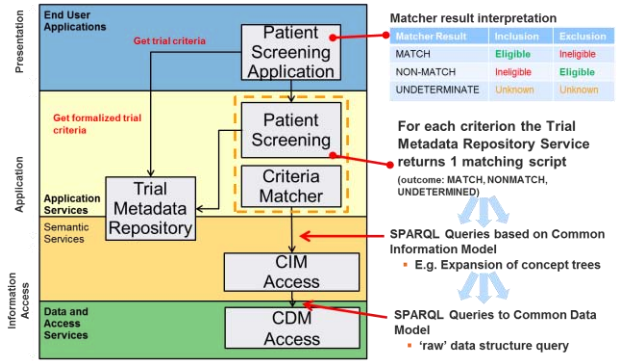


Figure 1. Architecture overview of the Patient Screening application

The application automatically identifies the trials for which a patient is eligible. We combine formal representations of criteria with a pragmatic and efficient solution in which templates are linked to scripts and extensively reused.

References

1. 1. S. Paraiso-Medina et al., "Semantic interoperability solution for multicentric breast cancer trials at the INTEGRATE EU project". In Proceedings of the HEALTHINF 2013, 2013.

Mapping Document Types using LOINC ® Document Ontology: a Case Study

Lindy Buhl RN, BSN[1]; Rachael Seeley RN, BSN[1]; Pat Wilson, RT(R), CPC, PMP[1]

[1]3M Health Information Systems, Inc. Murray, UT

Abstract

A case study was completed on the results of mapping 150 documents for a hospital. The findings were also compared with four previous studies. The case study results indicate the need for further expansion within the type of service and kind of document domains. With the feedback and requests from the industry there has been a constant progression and growth with the Document Ontology axis values in the past decade.

Introduction

The need to send, retrieve, and store health information can be greatly assisted with the coding of document types. Utilizing a universal standard for naming document types allows for information exchange and interoperability of disparate electronic systems[1, 2]. Logical Observations Identifiers Names and Codes® (LOINC®) provides a universal coding system that assists in the identification of related information within electronic messages[2].

Methods

A total of 150 document types from a large-scale hospital were evaluated and mapped using LOINC's Document Ontology. Each document item mapped was ranked by the level of specificity the LOINC code provided. The rankings included; specific, not specific, and not matched. A 'specific' ranking was for a mapped document item which met the criteria in each of the five LOINC document ontology attributes (kind of document, type of service, setting, role and subject matter domain) [1, 2]. With a 'Specific' ranking, the LOINC codes could include null values.

Results

The results of the authors' map based on the institution's use case, were as follows: Of the 150 document types, 98% (147) matched LOINC codes, leaving only 2% (3) that did 'not match'. Of the 147 mapped documents; 19% (28) were ranked 'specific', and 81% (119) were ranked 'not specific'. Of all 147 clinical documents; 17% (25) were mapped to "Note", 11% (16) to "Consultation note", and 7% (11) to "Study report". Also, of the total mapped; 6% had an unspecified provider, 35% had an unspecified setting and 30% had both unspecified provider and unspecified setting. There were instances of multiple site documents mapped to the same LOINC code. Less granular mappings of an institution's document to a LOINC code where less than three LOINC parts matched occurred most often from insufficient content of the Type of Service and Subject Matter Domains.

Conclusion

Industry-wide participation is of the utmost value with actual use cases detailing application of the ontology model, the map findings, and suggestions for improvement. In addition, evaluation of "administrative note" and potentially creating note types such as financial note, financial responsibility note, and various consents is suggested

References

[1] Vreeman DJ, LOINC tutorial: documents. 2013. LOINC website http://loinc.org/slideshows. Accessed Mar 8, 2013.

[2] Logical Observation Identifiers, Names, and Codes (LOINC) User's Guide. Regenstrief Institute. http://loinc.org/downloads/files/LOINCManual.pdf. Accessed Mar 1, 2013.Vreeman DJ, LOINC tutorial: documents. 2013. LOINC website http://loinc.org/slideshows. Accessed Mar 8, 2013.

Leveraging the Electronic Health Record to Identify Prescribing Errors through Rapidly Discontinued Medication Orders

Jonathan D. Burlison, PhD[1]; Lisbeth Bowlin, PharmD[1]; Donald K. Baker, PharmD, MBA[2]; Murad Hasan, MBBS, MS, DrPH[1]; R. Ben McDaniel, PharmD[1]; Jennifer J. Robertson, PharmD[1]; Scott C. Howard, MD[3]; James M. Hoffman, PharmD, MS[1]

Departments of [1]Pharmaceutical Sciences, [2]Information Sciences, and [3]Oncology, St. Jude Children's Research Hospital, Memphis, TN 38105

Background and Objective: Patient safety events, such as prescribing errors, provide learning opportunities to improve patient care systems. Traditional manual methods to identify these events produce incomplete results and are very time consuming, but electronic health records (EHRs) present opportunities to devise new, automated methods that can efficiently detect patient safety events. The "trigger tool" approach (i.e., selectively querying data elements that represent the possibility of an adverse event) has been automated in EHRs, and effectively captures events that are missed using other techniques.(1) The "trigger" *abrupt medication stop* was revised for use in EHRs by Koppel et al., where it was defined as "orders stopped within 120 minutes of being submitted."(2) Through a prospective evaluation, rapidly discontinued orders had a high positive predictive value (PPV) for detecting prescribing errors. We sought to validate this method in a retrospective manner that would allow for efficient ongoing data collection. The purpose of this study was to determine the feasibility and PPV of retrospectively identifying prescribing errors by retrieving and evaluating rapidly discontinued medication orders from an EHR.

Methods: Since 2010, all inpatient and outpatient prescriptions at our tertiary care pediatric hospital have used computerized prescriber order entry (CPOE). A query was developed to extract medication orders from the EHR that were canceled or discontinued within 120 minutes of being originally entered. Using a random number generator, orders from 28 randomly selected days were queried and evaluated as possible prescribing errors. Two pharmacists (LB & BM) reviewed the patient's health record and classified the abrupt medication stop as: 1) *most likely a prescribing error*, 2) *most likely not a prescribing error*, and 3) *not enough information to determine if an error occurred*. Discovered errors were assessed for potential clinical significance and cross-referenced with the hospital's voluntary electronic event reporting system to assess convergence of the two detection methods.

Results: 319 medication orders that were rapidly discontinued were retrieved and reviewed. Interrater agreement was 65% ($\kappa = 0.46$), which corresponds to a moderate level of agreement, and full agreement was reached by consensus in all cases after the reviewers consulted with each other. Roughly half ($n = 154$, PPV = 0.48) of the orders were determined to have been canceled or discontinued due to a prescribing error, 54 (17%) were non-errors and 111 (34%) were indeterminate. The data were divided by time from order submission to discontinuation into 15 minute intervals to examine trends over time. The PPV over time indicated that the trigger was most predictive for orders stopped within the first 90 minutes (PPV = 0.54 vs. 0.21 for 91-120 minutes). None of the detected errors were reported to the hospital's voluntary electronic event reporting system. The 154 errors included duplicate orders ($n = 46$, 30.1%), wrong drug ($n = 20$, 13.1%), wrong route ($n = 16$, 10.5%), and incorrectly ordered ($n = 14$, 9.2%, e.g., forgetting clarification instructions). Some of the errors ($n = 24$) might have been clinically significant had they reached the patient, including significant opioid analgesic dosing errors. 56% of the potentially significant order discontinuations occurred within 15 minutes of original submission.

Discussion: Similar to Koppel et al., our retrospective evaluation of rapidly discontinued medication orders identified many prescribing errors that were not detected by our voluntary error reporting system, and with a high PPV.(2) Along with the ability to detect potentially clinically significant errors, this method can be used to track error characteristics over time, such as error type, prescriber staff position, and drug class. While Koppel et al. demonstrated orders discontinued with 45 minutes of entry had the best PPV, our data indicate prescribing errors can be identified over 90 minutes with an acceptable PPV. Future work will refine this error detection method and seek ways to automate the error detection process even more, without sacrificing predictability.

Conclusion: Rapidly discontinued medication orders can be used to identify prescribing errors that are not detected through other methods. Applying this method retrospectively yielded a similar PPV to previous prospective evaluation, and a retrospective approach allows this error detection method to be used on an ongoing basis to inform systems and process improvement in the prescribing process.

1 Classen DC, Resar R, Griffin F, et al. 'Global trigger tool'shows that adverse events in hospitals may be ten times greater than previously measured. Health Aff. 2011;**30**(4):581-9.
2 Koppel R, Leonard CE, Localio AR, Cohen A, Auten R, Strom BL. Identifying and quantifying medication errors: evaluation of rapidly discontinued medication orders submitted to a computerized physician order entry system. J. Am. Med. Inform. Assoc.. 2008;**15**(4):461-5.

Meaningful Use Stage 1 Clinical Quality Measures: Comparison of Abstracted and Electronic Results

Zahid Butt MD[1], Sara K. Galantowicz, MPH[2], Sam Ogunbo, PhD[1], Ken McCormick[1], Sheryaar Butt[1]
[1]Medisolv, Columbia, MD; [2]Abt Associates, Cambridge, MA

Abstract

There is limited information on how individual measures abstracted from medical records compare, in a "live" setting, to the results generated through fully "re-tooled" electronic specifications. We compared performance of a case matched cohort of 15 abstracted "Core measures" to electronically-specified versions from Meaningful Use Stage 1 for 9 Eligible and Critical Access Hospitals; for ratio measures performance rates at the aggregate level yielded simple correlations ranging from .06 to .94. The continuous measures showed similar variability in correlation.

Methodology:
Nine hospitals were selected from Medisolv's client base that were live the longest with Medisolv end user applications designed to meet clinical quality measures reporting requirements for both CMS (Meaningful Use Stage 1 eCQMs for Eligible Hospitals) and The Joint Commission (Core Measures for accreditation and the CMS Inpatient Quality Reporting Program). Measure results for the abstracted measures were compared to eCQMs in a cohort of case matched patient populations over a six-month period. Correlations of results between the two methods were analyzed first at the measure level and included an aggregate of all facilities. This analysis was supplemented by a facility-level analysis. Performance rates, numerator, denominator and exclusion counts were compared for ratio measures (Stroke and VTE) and median times were compared for the continuous (Emergency Department) measures. The study included all patients 18 years or older at the beginning of the reporting period, regardless of payer and financial class, and discharged between January 1 and June 30, 2013. Because some hospitals use sampling for their abstracted measures, a matched case cohort was constructed comprising patients included in both measure calculations. The entire dataset for the project included 4,657 unique discharges.

Principal Findings:
Direct comparison of performance rates at the aggregate level yielded simple correlations ranging from .06 to .94 for the ratio measures. The continuous measures showed similar variability in correlation. Plotting the aggregate differences for the six months showed the differences to be statistically significant for many, but not all, measures. Analysis at the individual hospital level, however, showed several instances of perfect correlation for some measures, along with considerable variation in correlations across hospitals. Some measures had higher performance correlations across multiple hospitals. A more focused look at the results for one participating facility showed multiple factors may drive observed differences in results including hospital-specific factors, such as abstraction errors, documentation or workflow practices, and measure-specific factors, such as differences (sometimes quite subtle) in measure logic.

Table 1: Reasons for Variation in Hospital I (n=18 cases)

Reason	Frequency
Specification difference	12
Abstractor error	3
Documentation issue	2
Workflow issue	1

Conclusions:
These results suggest that measure "re-tooling" can be successful in a live setting and can replicate results of at least some medical record abstracted measures, provided the specifications match precisely and are implemented properly by the EH/CAHs. Study limitations included the small sample size and the six-month study period. The matched cohort design could also potentially introduce bias in favor of the chart-abstracted measures; this was not tested.

Prior Probability Assessment Wizard

Amos Cahan, MD,[1] James J Cimino, MD[1 2]
[1] Lister Hill National Center for Biomedical Communications and Clinical Center, National Library of Medicine, Bethesda, MD; [2]National Institutes of Health Clinical Center, Bethesda, MD

Bayesian probabilistic diagnosis relies on the clinician's estimated prior or pre-test probability for a disease. However, both overestimation of objective probabilities and poor between-physician agreement raises question about the validity of physicians' confidence judgments as a proxy for the "pre-test probabilities" in the implementation of the threshold approach[1,2].

As real-time, tailored and objective probability estimates are not currently available to clinicians at the point of care, we envisioned a clinical decision support-tool to correct or calibrate physicians' initial estimates. The tool we have developed, the Probability Wizard (PW), serves as a "compiler" that translates a physician-generated differential diagnosis (DD) list into probability estimates through an interactive process. The PW is designed to correct for probability over-estimation by eliminating subadditivity and is also likely to improve between-physician agreement. Starting from a physician-generated DD, the PW leads the physician through a graphic user interface-aided step-wise dialog, resulting in a DD list with numeric estimated probabilities assigned by the PW to each of the clinical entities in the list. Unlike other diagnosis support systems, the possible diagnoses are provided solely by the physician. Clinical validation is needed.

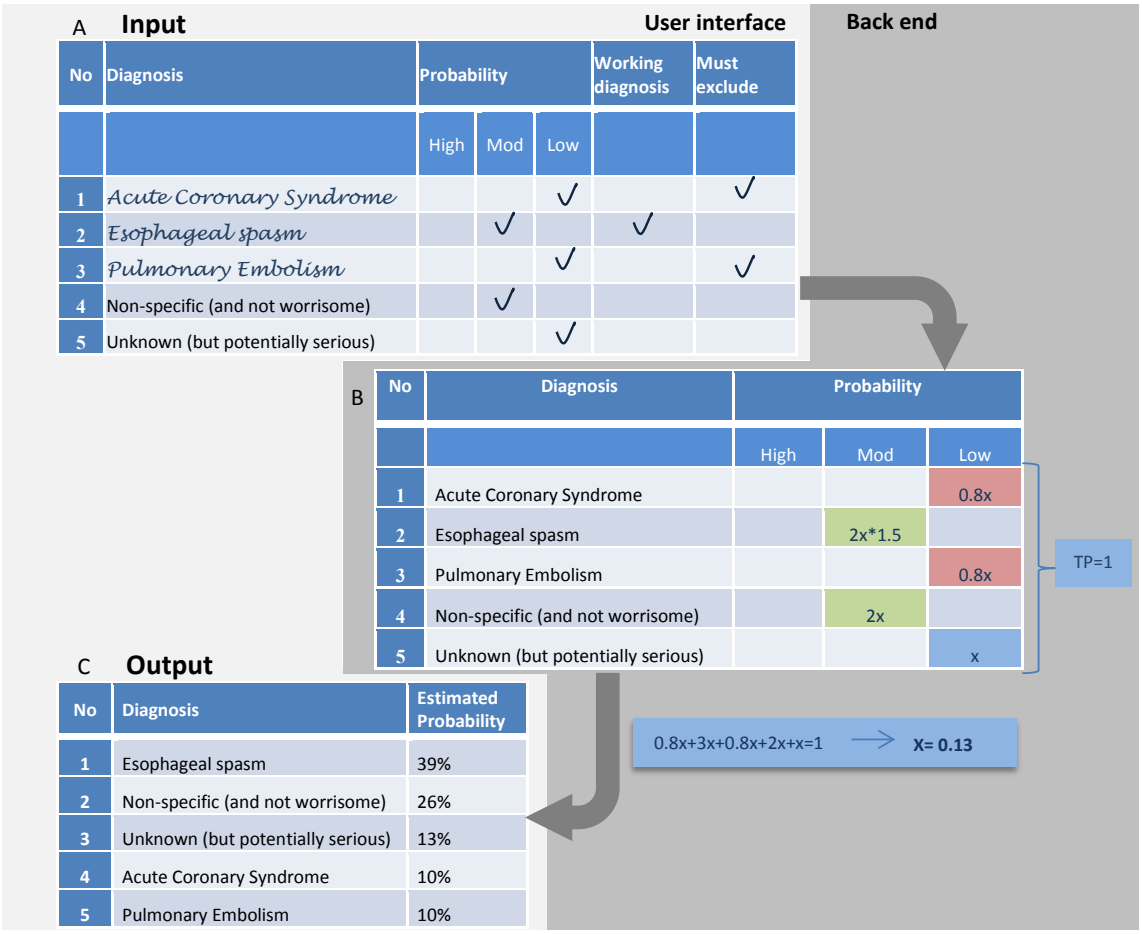


Figure: **Construction of numeric subjective pre-test probability estimates using the Probability Wizard (PW) interface**. In this example, a patient with chest pain is assessed. (A) A preliminary differential diagnosis list is formed by the physician and likelihood categories are checked. Two diagnostic categories are introduced by the PW to account for common clinical scenarios in which there is no specific diagnosis evident, namely "Non-specific" and "Unknown". A working diagnosis and diagnoses which must be excluded are checked if applicable. (B) The sum of all possible diagnoses (TP) is considered to be 1. Correction for moderate confidence (*2), working diagnosis (*1.5) and possible overestimation of the probability of hazardous conditions (*0.8) are applied and absolute numeric probability estimates are computed. (C) An adjusted differential diagnosis list with numeric pre-test probability estimates (in %) is presented to the physician.

1. Acad Emerg Med. 2004;11(6):692–4; 2. Med Decis Making. 1986;6(4):216–23.

Addressing some statistical challenges of using EHR data for clinical research

Fiona M. Callaghan, PhD[1], Dina Demner-Fushman, MD, PhD[1], Swapna Abhyankar, MD[1], Matthew T. Jackson, PhD[2], Mallika Mundkur, MD[1], Clement J. McDonald, MD[1], [1]National Library of Medicine, National Institutes of Health, Bethesda, MD; [2]Food and Drug Administration, Center for Drug Evaluation and Research, White Oak, MD

Abstract

Much has been said about the "big data" potential of large medical databases for clinical discovery, but even after the data are available there are still challenges for statistical analysis. We present two of the challenges here: combining unstructured and manually abstracted data, and imputing or modeling missing or unavailable data.

Introduction

We have conducted clinical projects based on MIMIC-II, a large, comprehensive intensive care unit (ICU) database that includes clinical notes, diagnosis codes, demographics and mortality information[1]. However, challenges remain.

Combining unstructured and manually abstracted data: metformin use and mortality risk in the ICU

We extracted information from clinical notes using natural language processing (NLP) or information retrieval (IR) techniques, and manually abstracted the "ground truth" for a subset of patients. The challenge was whether to use only the NLP/IR information, which is known for everyone but with less than 100% accuracy, or only manually abstracted data which is reliable but represents only a subset of the data, or somehow use a combination of both. We used a statistical misclassification method to combine abstracted and NLP/IR information to predict risk[2]. The estimates of risk, odds ratios (ORs) and confidence intervals (CIs), are more powerful and accurate using misclassification methods than those produced by using manual or NLP/IR data alone[2]. For example, we hypothesized that metformin before ICU admission would lower mortality risk due to its anti-inflammatory properties[3]. Metformin was extracted from admission notes using IR and manual abstraction. The metformin group saw a 70% reduction in risk with the manual data (OR=0.3, 95% CI 0.1-0.6, p=0.003), but a 50% reduction (and narrower CI) when IR and misclassification methods were added (OR=0.5, 95% CI 0.4-0.7, p<0.001).

Missing and/or uncertain data: imputation and modeling

Information may be entered as "unknown" or not entered at all (i.e., "truly missing"). A patient whose information is unavailable on *any* one predictor has all their information removed completely from a multivariable regression analysis, reducing the sample size and often introducing bias. In our obesity study, height (necessary for calculating body mass index) was missing for 25% of patients[4]. We solved the problem by imputing height based on age and gender and checking for bias. However, for another study on race and mortality, race was unavailable for 12% of patients and imputation was not possible. Responses included "declined to answer", "unknown", and "unable to obtain", as well as truly missing. Excluding the unknown race groups was not acceptable as they represented a significant proportion of the data and had a higher mortality rate compared to overall (28% versus 18%). We chose to address this issue by modeling the unknown race values as separate categories, but the work is ongoing.

Conclusion

Large, "real world", medical databases hold great promise for clinical research, but they must be analyzed with care. Often techniques from statistics and informatics have to be brought to bear in order to produce clinical results.

References

1. Saeed M, Villarroel M, Reisner AT, et al. Multiparameter intelligent monitoring in intensive care II: a public-access intensive care unit database. Crit Care 2011; 39: 952–960
2. Lederer W, Kuchenhoff H. simex: SIMEX- and MCSIMEX-algorithm for measurement error models. R package version 1.5. 2013. Available from: http://CRAN.R-project.org/package=simex.
3. Christiansen CF, Johansen MB, Christensen S, et al. Preadmission metformin use and mortality among intensive care patients with diabetes: a cohort study. Crit Care 2013; 17: R192
4. Abhyankar S, Leishear K, Callaghan FM, Demner-Fushman D, McDonald CJ. Lower short- and long-term mortality associated with overweight and obesity in a large cohort study of adult intensive care unit patients. Crit Care 2012; 16: R235

A Data Repository System for Translational Research

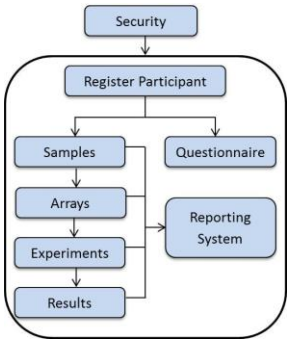
Frank J. Cammarata[1], Leonid Kvecher[1], Hallgeir Rui[2], Albert J. Kovatich[3], Leigh F. Campbell[4], Jeffrey A. Hooke[4], Norman P. B. Joseph[1], Craig D. Shriver[4], Richard J. Mural[1], Hai Hu[1]

[1]Windber Research Institute, Windber, PA; [2]Thomas Jefferson University, Philadelphia, PA; [3]MDR Global Systems, Windber, PA; [4]Walter Reed National Military Medical Center, Bethesda, MD

Abstract: *We developed a Data Repository System to manage, track, store, and process clinicopathologic, biospecimen, and experimental results information for integrative translational research. The system has 7 modules, including Patient, Biospecimen, Tissue Microarray, Experimental Results, and Reports. It takes input from multiple sources and formats and stores the information in a centralized database. The Data Repository System was designed to be secure, flexible, and extendable as the research expands.*

Introduction: A consortium headed by the Thomas Jefferson University, including clinicians and researchers from the Walter Reed National Military Medical Center, MDR Global Systems, and the Windber Research Institute (WRI) was awarded a Komen Promise Grant entitled "Therapy-relevant stratification of breast cancer patients: integrating pathology and biomarker analyses" to study the protein expression of 250 drug targets across 5000 invasive human breast cancer tissues using tissue microarray technology. The clinicopathologic data and sample information were collected at different clinical sites in different formats. The WRI team was tasked to develop an infrastructure system to support the study. We therefore have developed a system to manage, track, store, and process all of this information, including experimental results, to a centralized Data Repository, with multiple reporting capabilities.

Methods and Results: We initially designed an infrastructure composed of a laboratory information management system and a data warehouse system. As the project progressed, the team realized that the ideal design was too cumbersome to support a research project with limited funding. We finally decided to take the approach of developing a data repository incorporating the needed features of the two otherwise distinct systems, including tracking, storage, and reporting. The system was designed to be web-accessible. Java was used to build the graphical user interface and Oracle RDBMS was chosen as the backend database. Figure 1 shows the architecture of the system (A) and a screenshot of a user interface (B). The clinicopathological and sample information, data on tissue microarrays, antibodies, and experiments including high-resolution images and marker expression information as well as subcellular compartment information is collected at the clinical sites and loaded into the system. The accessibility to the data is controlled based on a user's defined credentials for the application. Three levels of reporting capabilities were developed, including data dump, ad-hoc query, and canned reports.



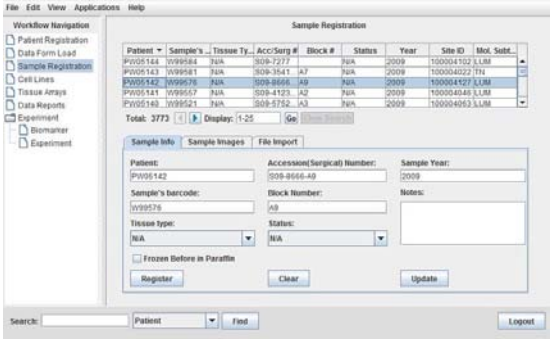


Figure 1(A) Workflow of the Data Repository System.
Figure 1(B) User interface of the Data Repository System.

Conclusion: We have designed and developed a Data Repository System that possesses the important functions of a data tracking system and a data warehouse. It's an innovative light-weight system capable of tracking, storing, and reporting information needed to support a translational research program, that manages information from multiple platforms, multiple sites, and multiple data formats.

Disclaimer: The views expressed in this abstract are those of the authors and do not reflect the official policy of the Department of Defense, or U.S. Government.

The H.O.P.E. Project: Trust and engagement in an online social networking intervention focused on enhancing HIV/STD resilience among African-American youth

Terrance R. Campbell, MA Ed, MSISM[1], Tiffany C. Veinot, PhD[2,3], Alison Grodzinski, MLIS[3]

[1]YOUR Center, Flint, MI, [2]School of Information, University of Michigan, Ann Arbor, MI; [3]School of Public Health, University of Michigan, Ann Arbor, MI

Problem Addressed:
Rates of sexually transmitted diseases among African American youth (aged 18-24) in Flint, MI are among the highest in the US. Our preliminary research regarding the technology-oriented practices of this population identified challenges and opportunities to their engagement with social network-based, digital sexual health promotion tools. We have argued for the value of a trust-centered. Such a framework should address young people's concerns for online safety along with trusted offline networks linkages. Additionally, we note a need to the online experiences to be as compelling as the online "drama" that both engages and repels them.

Specific purpose of the project:
The H.O.P.E. (HIV/STD Outreach, Prevention & Education) project investigates the effects of integrating social media components into an existing evidence-based HIV prevention program for African American youth. The H.O.P.E. project has two components: 1) face-to-face H.O.P.E. parties and 2) H.O.P.E. Online: an interactive website and social media pages. Created by YOUR Center in 1999, H.O.P.E. parties use naturally occurring friendship networks (FSN) to create opportunities for participants to assess their HIV risk, develop a risk reduction plan, and practice communication and decision making skills. Developed by a team of youth advisors, community organizations, and university-based researchers, H.O.P.E. Online includes peer blogs, expert advice, polls, local resource lists, a Facebook page and Twitter feed. This poster presents implementation results from one of our trust-centered design strategies: the use of social media contests to engage youth in disseminating HIV prevention messages and online referrals through trusted social networks.

Methods:
To engage youth in H.O.P.E. online, we conducted three social media contests that operationalized elements of our trust centered design framework, including strategy (participation and network embeddedness, and positive social influence) and functionality (offline networks and collective action). The first contest rewarded youth for posting content on the H.O.P.E. website, and for disseminating messages through Facebook and Twitter. The second contest focused on leveraging HOPE online participants' social networks to increase traffic to the website, Facebook and Twitter pages using text messaging, email and Facebook and Twitter inboxes to refer friends. The third contest rewarded both online activities and referrals. Contests were promoted at H.O.P.E. parties, via postcards to former party participants, and through H.O.P.E. online. We tracked participants' activities and awarded points per activity. For the first contest, all participants received a $10 gift card when they earned 50 points. The two contestants with the most points received $50 gift cards. For follow up contests participants were rewarded points based on referrals and activities. Participants with the most points were awarded $30 gift cards.

Results:
At a 3 month post party follow up survey, approximately 15% of H.O.P.E. party participants reported using the H.O.P.E. website or participating in other online activities. However, the overall contest period (following the 3 month follow up survey) was accompanied by a 103% increase in website visits and a 35% increase in the number of online participants; to include Facebook and Twitter (Shares, comments and retweets). Roughly, 54% of contest participants joined the Facebook page, 30.5% followed H.O.P.E. on Twitter and 27% used text messaging to send messages to H.O.P.E. staff and to their friends. Many also indicated this was the preferred this method of contact for follow ups. Yet, the number of new participants and online activities fell with successive contests and after the completion of all contests.

Conclusions:
Our findings suggest that social media contests can be an effective, trust-centered approach for increasing African American youth engagement in online HIV/STD prevention. These findings add implementation experience to previous theoretical literature regarding the potential role of trust enhancement in technology adoption and usage. Notably, however, the lack of sustained increase in usage over time suggests a potential ceiling effect, diminished interest over time, or the ongoing need for financial incentives. However, strategies for sustained engagement over time are needed; perhaps alignment with participants' daily routines or that of one of their FSN could assist. Additionally, we require insight into participants' subjective experiences of trust and safety while using the intervention. We will investigate these issues in our ongoing research.

Use of Enterprise Clinical Decision Support Infrastructure to Implement Pharmacogenomics at the Point of Care

Pedro Caraballo, MD, David Blair, MD, Michelle Elliott, MD, Robert Bleimeyer, John Crooks, Donald Gabrielson, PMP, Gaurav Jain, Wayne Nicholson, MD, PharmD, Charles Pugh, Padma Rao, Cloann Schultz, PMP, Lynn Summerlin, Joseph Sutton, Carolyn Rohrer-Vitek, Kelly Wix, PharmD, RPh, John Black, MD, Mark Parkulo, MD
Mayo Clinic

Abstract
Human cognition alone is insufficient to implement pharmacogenomics at the point of care. We report our progress accelerating translation of pharmacogenomics guidelines into clinical practice. We are using an enterprise clinical decision support infrastructure and adapting current functionality of commercially available electronic medical records to develop and implement several drug-gene interaction rules. We also combine the CDS alerts with additional educational efforts by using direct links to web resources.

Introduction
Pharmacogenomics (PGx) has the potential to positively impact the prescribing behavior of clinicians. Computer-based clinical decision support (CDS) integrated in the electronic medical record (EMR) could be used to support the implementation of the complex and increasing PGx knowledge. However, current commercially available EMRs have not been designed to implement PGx. We present our efforts to customize our EMRs and enterprise CDS infrastructure to implement PGx across our diverse clinical practice.

Enterprise Clinical Decision Support Infrastructure
Mayo Clinic is committed to promote and coordinate common development, implementation, evaluation, and maintenance of computer-based clinical decision support across the continuum of care. We utilize the operational infrastructure supporting real-time CDS integrated in commercially available EMRs to implement PGx alerts at the point of care. We have two main EMRs, Centricity Enterprise and Cerner Millennium with their respective integrated rule expert systems, Blaze Advisor and Discern Expert. These systems allow Mayo experts to represent knowledge, and develop, deploy and maintain clinical rules to improve quality of care and efficiency. These rules are developed by a multidisciplinary group including clinical experts, pharmacy informatics, nursing informatics, system engineers, software developers, and others, to define the functional and technical specifications of the clinical rules. Extensive inclusion and exclusion criteria are utilized to minimize incorrect or excessive alerts and to avoid alert fatigue. Transactional data associated to the alerts are collected to assess performance and clinical impact of the rules. This program has been very successful implementing multiple CDS rules across Mayo enterprise.

Pharmacogenomics at the Point of Care
Our institution has established the Center for Individualized Medicine and the Pharmacogenomics Task Force to accelerate the implementation of genomic science into routine clinical practice. Working together with the CDS Program, we have several drug-gene interaction rules at various levels of development and implementation across the enterprise (*HLA-B*57:01*-abacavir, *HLA-B*15:02*-carbamazepine, *TPMT*-thiopurines, *CYP2D6*-codeine/ tramadol/tamoxifen, *CYP2C19*-clopidogrel, *SLOC1B1*-simvastatin and *CYP2C9/VKORC1*-warfarin). We are adapting current commercial EMR functionality to accommodate the following specifications: 1) Alert provider if PGx testing is required before prescribing specific medication based on current clinical guidelines. 2) Transfer structure PGx test results to the EMR or capture outside results at the point of care. 3) Document in the problem list or allergy module the relevant genotype/phenotype. 4) Notify ordering provider of the PGx test result by special inbox which contains the genotype/phenotype and specific instructions related to drug-gene interaction. 5) Alert prescribers of a significant drug-gene interaction when ordering specific drug. 5) Provide a web-link in the inbox and alert to additional educational resources in an easy to use Q&A format (AskMayoExpert).

Conclusion
The complexity of PGx requires computer-based aids to become actionable in routine clinical practice. Commercially available EMR systems can be adapted to implement basic gene-drug interactions but more complex genomic data and multi-gene/multi-drug guidelines may require additional development of EMR functionality.

Adopting a Collaborative Program Evaluation Model to Aid Administration and Evaluation of a Large-Scale Public Health IT Grant

Jonathan H. Cardwell, MS[1], Kristina Doing-Harris, PhD[1], Megha Kalsy, MS[1], Wu Xu, PhD[1, 2, 4], Jennifer H. Garvin, PhD, MBA, RHIA[1, 2, 3]

[1]Department of Biomedical Informatics, [2]Division of Epidemiology, University of Utah, Salt Lake City, UT; [3]VA Health Care System, Salt Lake City, UT; [4]Utah Department of Health

Abstract

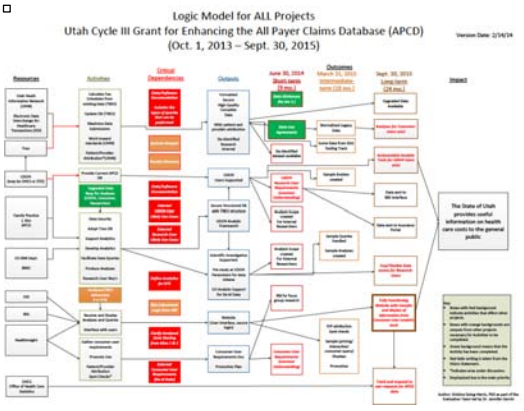
We adopted the Model for Collaborative Evaluations (MCE) developed by Rodriguez-Campos[3], as part of program evaluation for a large-scale, federally funded information system improvement project with the Utah Department of Health. In this poster we describe the use of logic models to facilitate application of the collaborative evaluation model. Reception has been mixed with the use of logic models particularly well received. This feedback indicates that ongoing communication about evaluation goals is necessary.

Introduction

Many projects in IT and Health IT are often not successful in development stages[1,2]. In order to have the best chance of success, we chose a collaborative evaluation model that involves a substantial degree of interaction between evaluators and stakeholders[3]. This approach is unique among evaluation models in that collaborative members (CMs) are stakeholders that work jointly with evaluators to help achieve the overall project and evaluation vision[3]. These CMs are members of relevant groups that provide immediate contribution and feedback throughout the duration of the evaluation. Through the use of collaboratively developed logic models, CMs are actively engaged in understanding the components of the complicated projects through a logic model, which include inputs, critical activities, outputs, and outcomes, both from their own groups' perspectives and with respect to the "big picture" or overall project perspective. We present our current progress in the use of the MCE in a large-scale public health IT project. We work with the Utah Department of Health on the "Enhancing Utah APCD for Healthcare Cost Transparency Cycle III" grant. We iteratively produced logic models for each Aim's work group, along with an overarching logic model for the entire Cycle III grant. The overall goal for the Cycle III grant is to improve the Utah All Payers Database[4] by producing online pricing/cost transparency reports for consumers, employers, researchers, and the general public in Utah. The grant consists of three Aims, each with their own scope and deliverables and also inter-depended with each other. The three aims relate to data quality and security, IT infrastructure, and dissemination.

Methods

Using the MCE requires a cyclical interaction with CMs providing feedback throughout the process. MCE goals include "identifying the situation," "clarifying expectations," "establishing a shared commitment," "ensuring open communication," "encouraging best practices," and "following specific guidelines." To facilitate these goals, logic models are collaboratively developed. Evaluation team members use grant documents as reference points, observe and participate in grant meetings, and review vision statements and work plans to create initial Logic Models. These models depict inputs, critical activities, outputs, and outcomes (see an example in figure 1). Logic models are shared through individual and group meetings with CMs and iteratively refined. Each model is refined until no additional feedback is provided. Success is measured by tracking our evaluation questions throughout the program, such as monitoring the use of best practices in large-scale Health IT implementations and data quality, tracking security and privacy, and facilitating best practices in dissemination – all of which is part of our evaluation plan.



Results

The current overall logic model for all Aims in Cycle III is shown in figure 1. Three other models have been developed for the Aims themselves. We solicited viewpoints from the principle investigator and the leads for each aim regarding the evaluation model as part of our formative evaluation process. Satisfaction with the model and understanding of the use of it are mixed, half of the respondents, who provided a response, found the process useful, a quarter were ambivalent and a quarter found the process intrusive. This feedback indicates that ongoing communication about evaluation goals is necessary. Several expressed that the logic model and collaborative model provide "valuable guidance" for "what tasks/activities need attention." In addition, others reported being unclear about the role of the evaluation team and appeared to understand that the team should be "focusing on an evaluation of the impact of the project."

Conclusion

Collaborative member engagement is a transformative methodology in aiding administration of the development of large-scale public health informatics systems and tools. While the collaborative model is in alignment with and can compliment project management efforts, and that the model supports the best practices of HIT, there is important clarifying communication about the goals of the collaborative evaluation model that needs to be provided on an ongoing basis.

Acknowledgements

We thank the members of the Utah Cycle III for their engagement with the program evaluation team. This publication is funded by CMS Grants to States to Support Health Insurance Rate Review and Increase Transparency in Health Care Pricing, Cycle III [Grant Number: 1 PRPPR140059-01-00], through a subcontract with Utah Department of Health Office of Health Care Statistics.

References

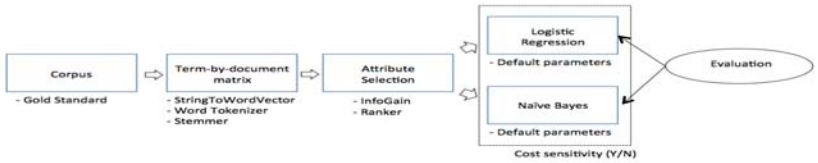
1. Kaplan and Harris-Salamone. 2009. Health IT Success and Failure: Recommendations from Literature and an AMIA Workshop. J Am Med Inform Assoc. May-Jun; 16(3): 291–299
2. McManus, J and Wood-Harper, T. 2007. Understanding the sources of information systems project failure. Management Services, Autumn, pgs. 38-43.
3. Rodriguez-Campos, L and Rincones-Gomez, R. 2013. Collaborative evaluations: Step-by-step. Stanford, CA: Stanford University Press.
4. Utah Department of Health. Utah All Payer Claims Database: Description and Background. http://health.utah.gov/hda/apd/about.php. Accessed March 6, 2014.

Improving Case Management via Statistical Text Mining in a Foster Care Organization

Arturo Castellanos, Alfred Castillo, Monica Chiarini Tremblay PhD
Florida International University Miami, FL

Abstract. Every year more than 800,000 children in the U.S. spend time in foster care. About 35% of these kids are administered psychotropic medication. An increasing ratio of foster care children per caseworker makes it challenging to balance their multiple roles during the lifecycle of a case. Although there are review boards for identifying cases that require special attention, reviewing all the information is time-prohibitive. This task is further complicated by poorly written, or incomplete, case note entries from the overwhelmed case workers. As part of a comprehensive study with a technology-savvy foster care organization, we investigate a challenging nationwide problem of monitoring and auditing children on psychotropic medication; an issue that has led, in some instances, to, unnecessary deaths as a result of over-prescriptions or concomitant medication. Often, the needed information to identify these cases exists in unstructured data (e.g. home-visit notes from three different foster care agencies), but is difficult to disambiguate due to the sheer size and amount of text in these records. We illustrate how Statistical Text Mining (STM) help complement current practices in classifying and prioritizing cases of psychotropic drug use. This would significantly reduce the load for case managers by providing a smaller sample that highlights children showing evidence of psychotropic drug use.

Methodology. In this study we adopt the CRISP-DM framework—briefly described (see Figure 1 for approach)[1]. The steps in CRISP-DM are iterative steps that include 1) understanding the problem and its context, 2) understanding the data, 3) preparing the data, 4) applying machine learning techniques, 5) evaluation and 6) deployment of results. Each step of the framework is often improved by revisiting a previous step. The first step and second steps are the result of a seven-year research relationship with a leading community based foster care organization. The third step, data preparation activities was to solicit the help of nurse case managers to develop an accurate sample to construct a ''gold standard'' dataset with correctly labeled cases of the binary target variable, psychotropic medication use. Since there is health related information contained in the various case notes we had to follow HIPAA guidelines for personal health information (PHI) identifiers. Children and case identifiers were de-identified by generating a random number and using that number for reference of the child/case. The fourth step requires machine learning of unstructured data, which we did using WEKA [2], an open-source data mining tool. First we tokenized the text in the corpus into a large document-by-term matrix. We used tf-idf to enhance potential predictive terms in single documents in the corpus from the different agencies. Words can be left as is or can be reduced into stemmed words, which are not necessarily linguistically valid words (e.g. bruis: bruise, bruises, bruised). Next, we selected important features using the InfoGain evaluator, which evaluates the worth of an attribute by determining the information gain with respect to the target variable; the ranker then ranks the attributes based on the threshold value we set based on feedback from caseworkers. Also, we need to consider the tradeoff between sensitivity and specificity. Sensitivity is the ratio of predicted children on psychotropic medication to the population of children on psychotropic medication, and specificity is the ratio of the population of children on psychotropic medication –the ratio of children predicted not to be on medication to all children not on medication. We included cost sensitivity to systematically force the algorithms to view the false negatives as a cost function.



Summary of Results and Evaluation. We used Logistic Regression and Naïve Bayes machine learning algorithms with various permutations of stemming {no, yes} and cost sensitivity {no, yes} to develop a total of 8 models (4 per algorithm). The performance of the different models is compared using common metrics such as: recall, precision, root mean squared error (RMSE), F-measure, and ROC curve. Based on statistical results, the Logistic model had a better overall fit for STM than Naïve Bayes by .373 in F-measure. The best performing Logistic model resulted in 322 TN, 30 TP, 3 FN, and 3 FP. The best performing Naïve Bayes model resulted in 276 TN, 30 TP, 3 FN, and 49 FP. Selection of a model is ultimately requires consideration of the different implications in practice. There are models that performed with higher precision, and others that performed with higher recall. In the former, clinical case managers have a lower load of documents to revise at the expense of missing potential positive cases. In the latter, they have a higher load of documents to review with the benefit of including more of the positive cases that are misclassified. For example, the "best" performing model (Logistic in bold) was chosen based on F-measure, but with a recall of 0.909 it actually misclassified 3 children that were on psychotropic medication. If this is deemed unacceptable, then the second Logistic model would be chosen, which resulted in a perfect classification of all children using psychotropic medication (recall = 1.000) at the expense of incorrectly classifying kids that do not take psychotropic drugs.

Precision	Recall	F-measure	RMSE
$P = \frac{TP}{TP + FP}$	$R = \frac{TP}{TP + FN}$	$F = \frac{2(P * R)}{P + R}$	$\sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}{n}}$

Table 1. Formulas for Evaluation

	Precision	Recall	F-measure	RMSE	ROC curve	Stemming	Cost sensitivity
Logistic	**0.909**	**0.909**	**0.909**	**0.103**	**0.998**	**No**	**No**
	0.717	1.000	0.835	0.151	0.998	No	Yes
	0.966	0.848	0.903	0.103	0.998	Yes	No
	0.717	1.000	0.835	0.151	0.998	Yes	Yes
Naïve Bayes	0.323	0.970	0.485	0.408	0.970	No	No
	0.267	0.970	0.418	0.468	0.970	No	Yes
	0.378	**0.909**	**0.536**	**0.366**	**0.971**	**Yes**	**No**
	0.327	0.970	0.489	0.409	0.971	Yes	Yes

Future Research and Next Steps. The initial results are promising and show the potential of significantly minimizing the workload of caseworkers in foster care organizations. In addition to using standard metrics, new metrics should be also explored for evaluation (e.g. how much time it saves the case workers).

References

[1] Tremblay, M. C., D. J. Berndt, S. L. Luther, P. R. Foulis and D. D. French (2009). "Identifying fall-related injuries: Text mining the electronic medical record." Information Technology and Management **10**(4): 253-265.
[2] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

Integrating Information from Unstructured Text with Structured Clinical Data from an Electronic Medical Record to Improve Patient Cohort Identification

Victor M. Castro MS [1], Sergey Goryachev MS [1], Christopher D. Herrick MBA [1], Vivian S. Gainer MS[1], Martin Rees BS[1], Shawn N. Murphy MD PhD [1,2]

[1]Partners Healthcare System, Charlestown, MA; [2]Massachusetts General Hospital, Boston, MA

Background - The availability of large-scale electronic medical records (EMR) enables the rapid identification of patient cohorts for research using phenotypes derived from clinically-collected data. The Informatics for Integrating Biology to the Bedside (i2b2) framework provides software tools construct patient queries based on specific inclusion and exclusion criteria[1]. Historically, clinical data available to run these queries originate from diagnosis and procedure billing codes, electronic prescriptions, lab test results, and other structured data sources. Since these data were not collected specifically for research purposes, however, they are often incomplete or lacking in detail. EMRs have a wealth of historical unstructured data in the form of text reports and clinical narratives that have been largely under-utilized to identify patient cohorts. A number of recent efforts have developed natural language processing pipelines to extract information from clinical narratives. In this work, we leverage one such NLP pipeline to extract free text diagnosis and medications from inpatient discharge summaries, map these results to common controlled vocabularies and develop a metadata hierarchy to enable researchers to query the NLP results along-side the structured results using the i2b2 query tool.

Methods – A total of 12,122 discharge summaries for 10,502 patients from two large tertiary care hospitals in Massachusetts were obtained for NLP parsing. The notes were parsed using the UMLSMapper, part of the HiTex NLP pipeline[2]. Text fragments were mapped to either an ICD-9-CM diagnosis or an RxNorm medication. These vocabularies are already in use to query medication and diagnosis data. Negated phrases were tagged as such and included in a separate hierarchy. Mappings were pruned to exclude non-specific mappings (i.e. “evaluation”, “temperature”, “tear”) which occurred often in the text but did not provide much semantic value. In addition, to remove any possible PHI from the results, we pruned mappings that were based on common proper names from the US Census (i.e., “Evan” mapped to Evan’s Syndrome was excluded). The results were loaded into an i2b2 data mart alongside coded diagnosis and prescriptions billing and prescribing data sources allowing researcher to build queries in the i2b2 query tool using both data sources. Finally, we evaluated the performance of the NLP pipeline on a subset of NLP-extracted diagnosis by comparing structured ICD-9 admission and secondary billing codes associated with each inpatient admission. In addition, we used coded medication reconciliation data at admission and discharge available for a subset of discharge summaries to validate NLP-extracted medications.

Results - We successfully extracted 376,716 disorders mapped to 2,244 distinct ICD-9 codes. The most common diagnosis extracted was Essential Hypertension (ICD-9 401.*) identified in 56% of discharge summaries. 19.7% of the mapped text diagnoses were identified as negated by the NLP process. In addition, 405,613 medications were extracted and mapped to 5,753 distinct RxNorm medication preparations. The most common extracted medication was Aspirin (RxCUI 1191). Initial validation results of extracted diagnosis and medications demonstrated a high-level of accuracy. However, a large number of NLP-extracted diagnoses did not have corresponding billing diagnosis indicating the extracted terms provide greater sensitivity at enumerating a patient’s full problem list during an inpatient admission.

Conclusion – Incorporating diagnosis and medication data extracted from unstructured clinical reports with existing structured data in the EMR improves the specificity and sensitivity of identifying patients in an EMR. We have linked these disparate data sources with existing controlled vocabularies to allow researchers to quickly identify relevant criteria and construct queries across both structured and unstructured data sources without requiring access to PHI. Future work will expand the validation of these results to manually annotated gold standards.

References -

1. Murphy SN, et al. Architecture of the open-source clinical research chart from Informatics for Integrating Biology and the Bedside. in AMIA Annual Symposium Proceedings, American Medical Informatics Association; 2007. pp. 548
2. Zeng QT, et al. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Med Inform Decis Mak. 2006;6:30.

Electronic Implementation of Adolescent Health Guidelines for Preventative Care Transformation: Challenges for the Child Health Improvement through Computer Automation (CHICA) System

David Chartash, BEngSc, MHSc[1,2], Tamara M Dugan, BSc[1], Amy Lewis Gilbert, JD, MPH[1,2], Matthew C Aalsma PhD[1,3], Stephen M Downs MD, MS[1,2]
[1]Children's Health Services Research, Indiana University School of Medicine, Indianapolis, IN; [2]Regenstrief Institute, Indianapolis, IN; [3]Adolescent Medicine, Indiana University School of Medicine, Indianapolis, IN

Abstract

Adolescent health screening, particularly when targeted at mental health, substance abuse and sexual health, is a challenge for pediatricians. We describe how the Child Health Improvement through Computer Automation (CHICA) system offers a unique method of deploying guideline logic for a future clinical trial.

Body

Adolescent screening guidelines are an important component of health maintenance. In the specific areas of mental health, substance abuse and sexual health, they are not well followed. Screening is an essential element of the Patient Protection and Affordable Care Act, which mandates coverage of screening according to the American Academy of Pediatrics Bright Futures guidelines. Pediatricians may not be well prepared to care for this population. Few pediatricians provide comprehensive care to adolescent patients. In primary care, adolescent health service delivery is incompleteand of marginal quality as preventive care of adolescents is demonstrably difficult.

This project seeks to improve adolescent health services related to mental health, substance use and sexual health by implementing the *Bright Futures* guidelines. The Child Health Improvement through Computer Automation (CHICA) clinical decision support system (CDSS) has been effective for collecting structured clinical data and deploying pediatric guidelines, including maternal depression screening and improve targeted screening for disease. We are expanding CHICA to accommodate algorithms in *Bright Futures* for depression, sexual risk and substance abuse.

CHICA is unique in its workflow. It includes a tailored selection of questions to ask patients in the waiting room that inform alerts and reminders to the pediatrician that are selected based on the prioritized needs of the patient. CHICA uses clinical context to prioritize value-based decision support. This logic is achieved through the adaptation of the *Arden Syntax for Medical Logic Systems*. CHICA relies on a hybrid procedural/production rule-based implementation of medical logic modules (MLMs). This implementation allows for the greater facilitation of algorithm based guidelines and serial logic, as well as management of the institution specific "curly-brace problem." This allows for the capturing of clinical context in a data dictionary. Further enhancing the implementation is the use of *PRODUCE* and *CONSUME* modes, which allow for the definition, collection and deployment of structured clinical data and decision support logic, using the same MLM.

The deployment of the medical logic modules for adolescent screening is tied to the use of screening algorithms for requiring both sequential receipt and display of clinical data to perform logical operations. Expanding the CHICA system requires further creation of data which are persistent across instances of skip logic and multi-pronged forking algorithms. Given the automated prioritization of questions, the use of a context driven data model to deal with screening will be a core component of future work.

CHICA is a unique CDSS that uses an adaptation of Arden Syntax to deploy guidelines and store and display data incorporating clinical context. As CHICA is expanded to adolescent screening guidelines, it will be evaluated in a cluster randomized trial evaluating quality of care.

Design of an Online Rabies Vaccination Information System

Jason Chase, MA[1], Isabelle Bichindaritz, PhD[1]

[1]State University of New York, Oswego, NY, USA

Introduction

The Rabies Vaccination System (RVS) was developed to replace the traditional data collection process originally maintained in Microsoft Access. Several key issues with the legacy system such as: limited access; lack of efficiency; and data integrity concerns, made it essential for the development of a dynamic database management system.

Methods

Design of the RVS is comprised of several open-source languages such as: MySQL, PHP, FPDF, HTML5, and CSS3. The system is virtually hosted on a secure network maintained by the Oswego County Health Department. Accessing the RVS within the counties intranet permits authorized user's access through Microsoft Active Directory and remote connections outside the secure network through a CISCO Virtual Private Network (VPN). Two participants were selected for stress testing the newly designed system and have been interviewed to provide system performance and functionality recommendations.

Results

The outcome of implementing the RVS has proven to securely give authorized end-users unlimited access to vaccination data efficiently while enforcing data integrity. The traditional method of data collection required copying data twice before securing it into a database. Data collected at clinics on hand written duplication forms were later entered into the Microsoft Access databases following scheduled vaccination clinics. The RVS provides users at vaccination clinics live access to the vaccination system to enter data once freeing up valuable time while protecting the integrity of the clinical data from duplication errors. Projects related to the RVS prove that data collection processes can be made more efficient, secure, and accessible while creating better work-flow for all users. Another consequence of creating the RVS is that with data collection being performed once, valuable hours at the County Health Department would be saved averaging $500,000 in employment costs in a projected 10 year period by replacing the data entry specialist with the Rabies Vaccination System.

Discussion / Conclusion

Enforcing data Integrity is a key element in any data collection process. Errors in data collection are easily made when it is copied from hand written forms, especially when the forms are written from multiple employees with differing handwriting, and then processed into an electronic record keeping systems by an entirely different employee. The Rabies Vaccination System collects data more reliably than the traditional method utilized by the Oswego County Health Department by collecting data one time. This results in reduced data integrity issues.

Acknowledge

The authors would like to thank the Oswego County Health Department and in particular Stephanie Carmody, ABS for their support and encouragement during this work.

Improving Prediction of Type 2 Diabetes Using Genomic Domain Information

Christopher Chen, BS, Michael A. Grasso, MD, PhD
University of Maryland School of Medicine, Baltimore, MD

Abstract

Incorporating genetic information into type 2 diabetes prediction models has so far yielded little improvement, most of which sum mutations in a linear mutation "score". We present a Support Vector Machine model trained with individual SNPs and domain information: the set of information describing a SNP, such as its locus and potential amino acid changes.

Background

Type 2 diabetes is a worldwide health problem with a high heritability rate and multiple sequellae, including nephropathy, retinopathy, and atherosclerosis. Prediction models for type 2 diabetes take clinical measures as input, such as family history, age, sex, BMI, and metabolic syndrome traits[1]. Because type 2 diabetes is a multifactorial disease with a high heritability rate, adding genomic information should improve prediction performance, but so far has only made marginal improvements.

Most previous prediction models have treated SNPs (single nucleotide polymorphisms) equally. However, SNPs are not equivalent and should be weighted independently. Furthermore, it is possible to increase information content via a SNP's domain information, such as its chromosomal location, its location within a gene, potential amino acid changes, and gene network relationships[2]. We aimed to build and test a Support Vector Machine based prediction model incorporating these two changes: domain information and independent SNP weighting.

Methods

We trained and tested our model using the GENEVA Genes and Environment Initiatives in Type 2 Diabetes case-control dataset, which contained data from the all-male Health Professionals Follow-up Study and the all-female Nurses' Health Study. The sample size included 5875 subjects with 3171 controls, with more than 500,000 SNPs per subject. We picked 32 diabetes-related SNPs along with their domain information, gathered from dbSNP and various journal articles. This produced a final dataset with 5875 subjects each with: 32 SNPs, 13 pieces of domain information per SNP, and 25clinical attributes, for a total of 505 attributes per subject.

We compared prediction models trained using Support Vector Machine with each experimental category: (1) genotype + domain + clinical, (2) genotype + clinical, and (3) clinical only. Algorithms and prediction models were evaluated using average Area under the Receiver Operating Characteristic curve (AROC) in a 10-fold cross-validation.

Results

Our results showed a modest improvement in performance between a model built with domain information versus a model built without domain information (AROC improvement of 0.005). The improvement was small, but consistent and meaningful. This shows that the small amount of domain information we added, especially non-redundant domain information, was the driver for the improvement in performance.

References

1. Wilson PW, Meigs JB, Sullivan L, Fox CS, Nathan DM, D'Agostino RB Sr. Prediction of incident diabetes mellitus in middle-aged adults: the Framingham Offspring Study. Arch Intern Med. 2007 May 28;167(10):1068-74.

2. Bochare A, Gangopadhyay A, Yesha Ye, Joshi A, Yesha Ya, Grasso MA, Brady M, Rishe N. Integrating domain knowledge in supervised machine learning to assess the risk of breast cancer. International J of Medical Engineering and Informatics, 2014;6(2):87-99.

"Doctors who ordered this also ordered…" Automated physician order recommendations and outcome predictions by data-mining electronic medical records

Jonathan H. Chen, MD, PhD[1,2], Russ B. Altman, MD, PhD[3,4*]
1 Center for Innovation to Implementation (Ci2i), Veterans Affairs Palo Alto Health Care System, Palo Alto, CA
2 Center for Health Policy / Center for Primary Care and Outcomes Research (CHP/PCOR), Stanford University, Stanford, CA
3 Department of Medicine, Stanford University, Stanford, CA
4 Departments of Bioengineering and Genetics, Stanford University, Stanford
*To whom correspondence should be addressed. E-mail: russ.altman@stanford.edu

Background
National healthcare reforms and incentive programs demand meaningful use of electronic medical records (EMR) to improve patient safety and cost efficiency. This will depend on EMRs integrating clinical decision support (CDS) to drive physician orders (labs, imaging, medications, etc.), the concrete manifestation of clinical decision making. Order sets, risk scores, and similar constructs already help consistency and best-practices, but they are limited by a top-down approach, requiring manual production and limited end-user awareness. A Big Data approach could instead crowd-source clinical expertise from the bottom-up, but it is unknown whether such a data-driven order recommendation system, analogous to Netflix or Amazon.com's "Customer's who bought A also bought B" system, can anticipate real physician orders and clinical outcomes.

Methods
EMR data was extracted from one year of inpatient hospitalizations at Stanford University Hospital (>5.4M structured data items from >18K patients, including physician orders, lab results, and diagnosis codes). Association statistics were counted for all pairs of the ~1,500 most common item types from a random training set of 16,408 patients to derive item-association conditional probability estimates, driving an order recommendation engine. For a separate random test set of 1,903 patients, data occurring within 4 hours of the hospital encounter was used to query for 10 recommended physician orders that were compared against the actual orders occurring within 24 hours. In addition to metrics of recall and precision, inverse frequency weighted variants are introduced to evaluate recommendations that are "specifically relevant" and not just common. The first 24 hours of data for each test patient was also used to query for the conditional probabilities of 30 day mortality and ICU admission within 1 week, which were compared against actual outcomes by ROC analysis.

Results
Compared to a reference benchmark of always recommending the most common items [1], the item-association order recommendation engine improves precision from 26% to 37% ($p<10^{-40}$). Integrating likelihood ratio estimates improves inverse frequency weighted recall from 3% to 17% ($p<10^{-120}$). The framework predicts 30 day mortality and 1 week ICU intervention with ROC AUC (c-statistic) of 0.88 and 0.78, respectively.

Discussion
This data-driven physician order recommendation engine improves upon a reference benchmark in anticipating real physician orders. Different evaluation metrics distinguish common orders from those more specifically relevant to a given clinical context. This same framework predicts clinical outcomes on par with state-of-the-art prognosis scores whose c-statistics range from 0.75-0.90 [2] and 0.69-0.81 [3] for predicting mortality and early ICU admission, respectively.

The key concern with these results is whether their basis on *common* practice patterns represents ideal ones. Prospective trials will be necessary to answer this and many important clinical questions in the pursuit of evidence-based decision making. The inevitable reality of everyday clinical practice however is a perpetual gap of prospective data, resulting in routine dependence upon individual expert opinion and anecdotal experience. Tools as described in this work can still elevate clinical decision making to a *data-driven* process based on the experience of thousands of practitioners from real-world settings in the Big Data era of EMRs.

References
1. Ekstrand, M. D. Collaborative Filtering Recommender Systems. *Found. Trends® Human–Computer Interact.* **4,** 81–173 (2010).
2. Lemeshow, S. & Le Gall, J. R. Modeling the severity of illness of ICU patients. A systems update. *JAMA* **272,** 1049–55 (1994).
3. Renaud, B. *et al.* Risk stratification of early admission to the intensive care unit of patients with no major criteria of severe community-acquired pneumonia: development of an international prediction rule. *Crit. Care* **13,** R54 (2009).

Indoor Location Awareness by Analyzing Ambient WiFi Signals in an Urban Setting – A Feasibility Study

Mike LC Chen, MD, Chia-Wei Lin, MS, Polun Chang, PhD
Institute of Biomedical Informatics, National Yang Ming University, Taipei, Taiwan (ROC)

Abstract

We conducted a pilot study to explore the feasibility of achieving indoor location awareness in a typical residence setting using off-the-shelf Android devices simply by analyzing ambient WiFi signals. We developed a sniffer program on Android platform to record ambient WiFi signals in different rooms, and demonstrated that the location of an Android device can be determined by comparing detected WiFi signals with previously recorded data.

Introduction

Activities of Daily Living (ADLs) is a term commonly used in health care referring to daily self-care activities such as eating, dressing, grooming, toileting, and functional mobility[1]. Clinically, ADLs are evaluated through subjective questionnaires[2]. Since functional mobility is the foundation to most self-care activities[1], a decline in functional mobility may be related to declining ADL scores. A subjective measure of functional mobility may be established by monitoring an individuals movement patterns at home. Previous work on indoor location tracking involved sophisticated and often proprietary equipment, such as infrared sensors, pressure sensors, cameras, and wireless beacons[3]. In this study, we explored the possibility to achieve location awareness at room level precision using off-the-shelf smartphones and just ambient WiFi signals.

Method

The study was conducted in a 3-bedroom apartment residence, on the 15th floor of a 17-story building in Taipei city. A room-by-room WiFi signal sweep was performed using an Android sniffer program developed in-house. The program is designed to repeatedly scan for all detectable WiFi access points' (AP) signals in a given room, and to record their MAC addresses. The signal sweep was repeated every 5 seconds for 50 times in the bedroom, the living room, and the study. The collected data was used to construct a room-by-MAC address probability matrix, depicting the probability of detecting a signal from a specific AP in a given room. Data collected from a separate round of WiFi sweeps were used as test data. The probability of detecting a signal from a given MAC address was looked up from the probability matrix, the results were added up to determine which room the test data was collected.

Result

For building the probability matrix, 2849 AP signals were detected during 150 signal sweeps (50 times in each room), consisting of 73 unique MAC addresses. Some signal was detected constantly across all 3 rooms, while many other signals were detected in only one or two rooms. A room-by-MAC address probability matrix was constructed. Location was determined correctly 13 times out of 15 test cases (Accuracy: 86.7%). One test case in the living room and another in the bedroom were mistakenly interpreted to be in the study.

Conclusion

Location awareness is only a first step towards establishing a quantitative measure of ADLs. Our study demonstrates the possibility of achieving indoor location awareness just by analyzing the MAC addresses of ambient WiFi signals. Our approach does not require proprietary hardware. For simplicity, the signal strength (RSSI) attribute, while also easily obtained through Android API, was not used in this feasibility study. A more sophisticated algorithm taking signal strength into consideration may yield higher accuracy.

References

1. Law M, Letts L. A critical review of scales of activities of daily living. The American Journal of Occupational Therapy 1989;43(8):522-8.
2. Wiener JM, Hanley RJ, Clark R, Van Nostrand JF. Measuring the activities of daily living: Comparisons across national surveys. Journal of Gerontology 1990;45(6):S229-37.
3. Alemdar H, Ersoy C. Wireless sensor networks for healthcare: A survey. Computer Networks 2010, Oct;54(15):2688-710.

A Preliminary Study of Coupling Transfer Learning with Active Learning for Clinical Named Entity Recognition between Two Institutions

Yukun Chen, MS[1]; Yaoyun Zhang, PhD[3]; Qiaozhu Mei, PhD[4]; Thomas A. Lasko, MD, PhD[1]; Joshua C. Denny, MD, MS[1,2]; Hua Xu, PhD[3,1]

Department of [1]Biomedical Informatics and [2]Medicine, Vanderbilt University, Nashville, TN; [3]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX; [4]School of Information and Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI;

Introduction: Named entity recognition (NER) is a common clinical natural language processing (NLP) task used to identify terms of important categories in clinical documents. Supervised machine learning (ML) methods in conjunction with a significant amount of annotation by domain experts can achieve good performance. However, clinical NLP informaticians need to repeat their effort of building the similar ML-based NER models from different institutions for the same task. Active learning (AL) has been reported to reduce the annotation cost while maintaining the desired quality of the model for NER. We investigated whether we could further reduce that cost using transfer learning under a model built previously for a different institution. We call our study transfer active learning (TAL).

Methods: We used a part of the annotated training corpus from 2010 i2b2/VA NLP challenge from two institutions: Beth Israel Deaconess Medical Center (BETH) and Partners Healthcare (PARTNERS). The BETH and PARTNERS corpora contain 73 notes with 8764 sentences and 97 notes with 7551 sentences, respectively. Our goal was to build an NER model for the target institution using active learning with initial samples selected based on a model built from the source institution's data. The model used conditional random fields (CRF) to extract problem, treatment, and lab test entities. We used annotated data from the source domain and unlabeled data from the target domain. The labels of data from target domain can be used only when they are queried. In this preliminary study we focused on methods to select initial samples from the target domain for active learning. We designed two TAL methods and compared them with a pure active learning baseline:

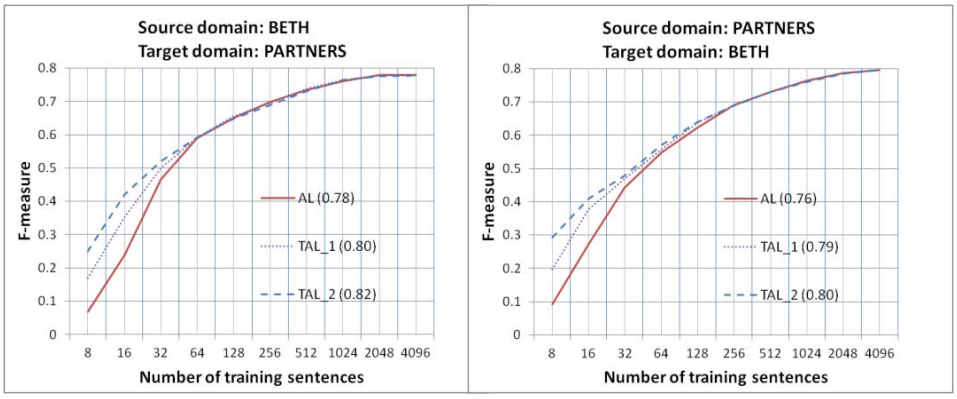
(1) *Baseline (AL)*: a standard active learning process, where initial sentences were selected randomly for labeling to build a CRF, under which the uncertainty of unlabeled sentences was measured by entropy for the next iteration of querying and re-training.

(2) *Full data source model-based uncertainty sampling (TAL_1)*: initial sentences were selected for labeling based on their uncertainty under a model trained using all annotated data in the source domain; we expected this to generate a better initial set of sentences compared to baseline.

(3) *Similar data source model-based uncertainty sampling (TAL_2)*: initial sentences were selected for labeling based on their uncertainty under a model trained using data in the source domain that was most similar to the target domain. The similarity metric used cosine similarity over three types of features: word count, part-of-speech dependencies, and similarity between UMLS concepts. Source-domain sentences were included in the source-domain model if they had a similarity greater than a given threshold to any target-domain sentence.

Evaluation and Results: We ran two TAL experiments: (EXP1) PARTNERS as source domain and BETH as target domain, and (EXP2) BETH as source domain and PARTNERS as target domain. Each experiment used 5-fold cross validation. The learning curve, which plots F-measure as a function of size of training sentences in target domain, is evaluated based on the area under learning curve (ALC) score. Compared to pure active learning, both TAL methods selected more informative sentences for annotation in early training, and the more sophisticated TAL_2 selected better sentences than TAL_1. In later training, where pure active learning had more information available for sentence selection, all three methods approached the same performance. For the ALC scores, TAL_2 (0.82 in EXP1 and 0.80 in EXP2) outperformed TAL_1 (0.80 in EXP1 and 0.79 in EXP2), which is better than AL (0.78 in EXP1 and 0.76 in EXP2). The results were consistent for both experiments.

Discussion: The TAL experiments demonstrate that transferring knowledge from a source domain can benefit data selection for active learning in a target domain. The benefit can be increased by using the most similar subset of data from the source domain instead of the entire set. However, the improvement was limited to an improved starting point of the learning curve. Future work includes investigating methods to increase the slope and asymptote of the learning curve using transfer learning.



Discover Improved-Outcome Evidence for Personalized Treatment from Electronic Health Records (EHR)

Chih-Lin Chi, PhD, MBA[1], Peter Tonellato PhD[2]
[1]University of Minnesota, Minneapolis, MN; [2]Harvard University, Boston, MA

Abstract

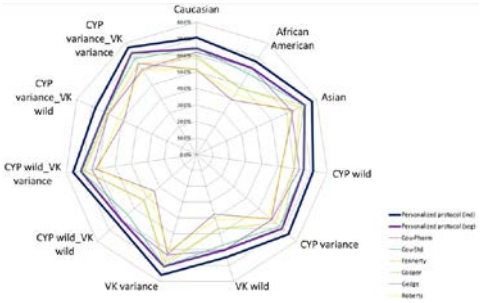
We describe an approach to identify improved-outcome evidence from EHR for optimal personalized treatment. Using personalized warfarin treatment protocol as an example, we identify the protocol option that maximizes outcomes and minimizes risks based on one's clinical and/or genetic characteristics. Results from clinical trial simulation demonstrate that personalized warfarin treatment protocol consistently and significantly improves outcomes from one-fit-all treatment across all patient subgroups.

Introduction

Clinical heterogeneity influences treatment efficacy and, subsequently, results in outcome variations across individuals in one-fit-all treatment settings. On the other hand, there is an opportunity to improve outcomes by currently existing treatments when we understand how clinical heterogeneity influences treatment efficacy and how much difference exists among treatment options. Specifically, we capture the relationship among characteristics, treatment options, and outcomes from electronic health records. With the captured model, we can extract such personalized treatment evidence by examining options that most improve outcome for particular individuals.

Method

The method for evidence discovery consists of three components: prediction, optimization, and production of decision support (DS) rules. We use personalized warfarin treatment as an example to show the method. In this example, we extract such evidence from PharmGKB data, individual-level data with clinical and genetic variables. We demonstrate the three components as follows. (1) Prediction: Clinical trial simulation platform[1] was used as the prediction model. In general, the platform was used to simulate warfarin treatment for a certain protocol over 10 days based on one's clinical and genetic characteristics and drug usage based on a protocol. Finally, the platform was used to predict Time in Therapeutic Range (TTR), which is a surrogate outcome that represents the degree to avoid bleeding and thrombosis when using a protocol for treatment. (2) Optimization: We apply (1) to simulate treatment for six different protocols and predict TTR for each protocol. Due to small solution space, an exhaustive search (an optimization approach) is able to quickly identify the protocol (out of 6 protocols) with the maximum TTR. (3) Production of DS rules: (1) and (2) predict the treatment protocol that maximize TTR outcome for an individual based on personal characteristics. The transparent property of a DS rule, such as the protocol most reduces risks for a certain subgroup of patients, allows easy use of personalized treatment in practical settings. For this component, we first define common characteristic(s) (e.g., African American) of clinical interest and then group patients based on the common characteristic(s). Finally, we identify the protocol that maximize outcome for this subgroup. Discovered evidence was applied to clinical trial simulation setting and we then compare efficacy between personalized protocol and one-fit-all protocol.



Results

Figure shows TTR comparison between two personalized protocols, bold blue (optimize for an individual) and purple (optimize for a subgroup) lines, and six one-fit-all protocols. Compared with one-fit-all protocols, both personalized protocols show consistently and significantly ($P<0.025$) higher TTR (12%~7%) across 11 subgroups.

Conclusion

Without inventing new treatment approach, our research provides another avenue to improve outcomes by identifying personalized treatment option based on one's clinical and/or genetic factors.

References

1. Fusaro, VA, Patil, P, Chi, C-L, Contant, CF, & Tonellato, PJ. A systems approach to designing effective clinical trials using simulations, Circulation, 2013; 127(4):517-526.

Exploring how online communication of cancer patients' symptoms to clinicians affects the changes in family caregiving responsibilities

Ming-Yuan Chih, PhD, MHA[1], Lori L. DuBenske, PhD[2], David H. Gustafson, PhD[2]
[1]University of Kentucky, Lexington, KY; [2]University of Wisconsin-Madison, Madison, WI

Abstract

Online communication of patient symptoms to clinicians may improve family caregiving workload. A latent transition analysis is used to explore this relationship. With the access to an online symptom reporting system that communicates patients' symptoms to clinicians instantly, caregivers are less likely to experience high workload.

Introduction

Family caregivers provide enormous support to those with cancer. As cancer patients' symptoms progress, family caregivers assume more difficult caregiving responsibilities. Communicating patients' symptoms to clinicians online may allow timely support to patients and caregivers, mitigating the rising caregiving responsibilities. A secondary analysis examines the effects of an online reporting system on caregiving workload before and after the intervention.

Methods

In two randomized trials[1], caregiver of advanced cancer patients in the control group were given a web-based system, Comprehensive Health Enhancement Support System (CHESS), designed to support family caregivers and patients.[1] Those in the intervention group were given CHESS plus Clinician Report (CR). CR, a new CHESS service, communicates reported patients' symptoms to clinicians via the Internet.[1] Pretest and 12-month post-intervention survey data from 102 caregivers were analyzed using a 3-step latent transition analysis method.[2] The outcome was measured with the Caregiver Load Scale[3] asking caregivers, "How much time and energy do you spend on the following tasks?" on a 0 (none) to 4 (a lot) scale for these tasks: CLS1: Medical/nursing treatments, CLS2: Personal Care, CLS3: Assistance with walking, moving around, CLS4: Emotional support, CLS5: Monitoring and reporting, CLS6: Providing transportation, CLS7: Managing illness-related finances, CLS8: Additional household tasks, CLS9: Structuring activities, and CLS10: Managing behavior problems.

Results

Two latent classes, high load bearers (HLBs) vs. low load bearers (LLBs), were found at pretest and 12 months (Figure 1). Descriptive statistics showed that male, lung cancer caregivers, whose patients have higher symptom burden, and those with higher physical burden are more likely to be HLBs at pretest—numbers will be presented on the poster. From Table 1, LLBs in CHESS+CR are less likely to become HLBs at 12 months than those in CHESS.

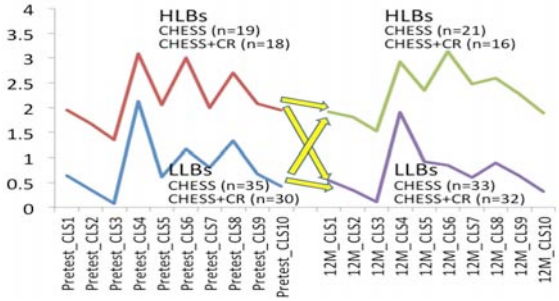


Figure 1. Latent profile plot

Table 1. Transition probabilities

Transition Prob. (Pretest->12M)	CHESS (n=54)	CHESS+CR (n=48)
HLBs -> HLBs	64.4%	62.3%
HLBs -> LLBs	35.6%	37.7%
LLBs -> HLBs	24.8%	15.5%
LLBs -> LLBs	75.2%	84.5%

Conclusion

The lower probability of increasing caregiving load over the 12-month intervention period among CHESS+CR caregivers (15.5%) as compared to CHESS caregivers (24.8%) may be due to the fact that caregivers were more likely to receive timely support because the CR was able to communicate uncontrolled patients' symptoms to clinicians remotely. About 1 in every 3 HLBs at pretest (from both groups) becomes LLBs at 12 months. It is unclear at this point if this is due to the benefits of CHESS overall or to other cancer-related factors.

References

1. Chih M, DuBenske LL, Hawkins RP, Brown RL, Dinauer SK, Cleary JF, et al. Communicating advanced cancer patients' symptoms via the Internet: a pooled analysis of two randomized trials examining caregiver preparedness, physical burden, and negative mood. Palliat Med 2013; 27(6):533–43.
2. Asparouhov T, Muthén B. Auxiliary Variables in Mixture Modeling: Three-Step Approaches Using M plus. Struct Equ Model A Multidiscip J 2014;:1–13.
3. Oberst MT, Thomas SE, Gass KA, et al. Caregiving demands and appraisal of stress among family caregivers. Cancer Nurs 1989;12:209–15.

Cohort-Based Discretization of Continuous Clinical Features to Discover Readmission Risk Factors for Heart Failure Patients

Si-Chi Chin, PhD[1] , Rui Liu, BS[1] , Ankur Teredesai, PhD[1]
[1]Center for Data Science, University of Washington-Tacoma, Tacoma, WA

Abstract

In this paper, we perform retrospective cohort study to validate the use of discretization techniques to discover readmission risk factors for heart failure patients. Insightful and principled visualization techniques may successfully help complex clinical data exploration tasks and aid in the process of knowledge discovery. In this paper, we apply Divide-n-Discover system to visualize and explore clinical data of different cohorts, supporting clinicians to dynamically explore the data and to understand how a given factor may influence the risk of readmission for a given patient. Further, since clinical datasets are nowadays voluminous, and also include large number of continuous attributes, we make use of distributed cloud based infrastructure to scale the dynamic exploration task for the cohort under investigation.

Problem Description

Data integration and understanding is a complex process in healthcare analytics. With the ever-increasing volume of data generated by the Electronic Medical Records (EMR) systems, data analytics in the healthcare domain is facing an overwhelming challenge in recent years. Insightful and principled visualization techniques may successfully help such complex clinical data exploration tasks and aid in the process of knowledge discovery.

Data mining techniques are increasingly used in clinical domains such as predicting risk for a particular disease. However, simply building a better predictive model, traditionally the goal of the data mining community is not enough in these problems. Domain experts are unlikely to trust models that are not understandable or actionable.

To successfully explore the ever-increasing amount of clinical data, we propose using cohort-based discretization of continuous clinical variables, such as age and length of stay, to identify meaningful cut-points for the problem of hospital readmission risk prediction. The proposed cohort-based *Divide'n'Discover* system involves healthcare professionals in the data exploration process and delivers interpretable results. In addition, the cloud based distributed system is designed as a service where both end-users as well as automated applications can submit healthcare data to be discretized.

This poster emphasizes the problem of predicting the risk-of-readmission for heart failure patients within 30-days of discharge, which has received extensive attention among healthcare professionals. However, the system can be extended to a wide range of healthcare problems. Encouraged by the preliminary findings in the prior study, we incorporate a new functionality that allows clinical researchers to dynamically create cohort from a large clinical dataset and visually discover meaning cut-points of the numerical variables of interests.

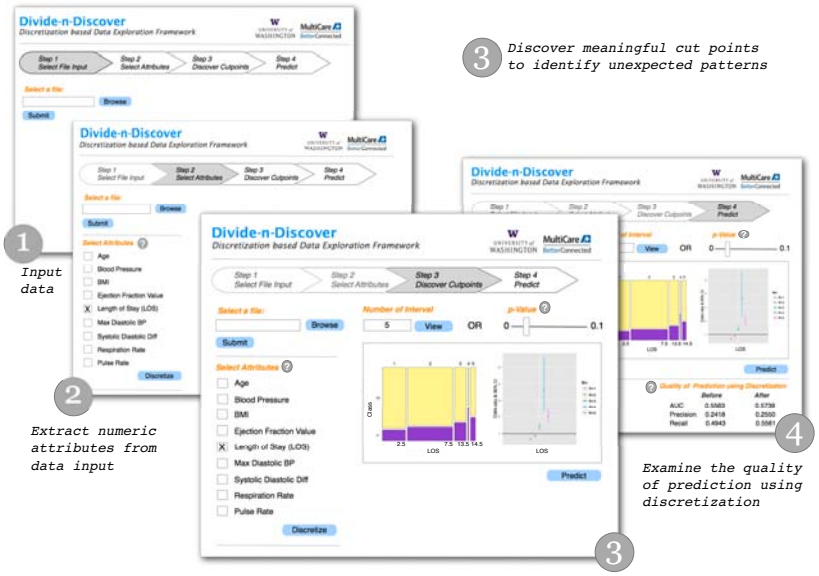


Figure 1. System screenshots of *Divide'n'Discover*.

Utilizing Smartphones to Enhance Urine Strip Test Accessibility

Jason H.D. Cho, M.S., Jacob Guggenheim, B.S., Elizabeth Arcan, B.S., Hannah Friedman, B.S., Shravan Gupta, B.S., Deana C. McDonagh, Ph.D., Joyce K. Thomas, MFA, Bruce R. Schatz, Ph.D.
University of Illinois at Urbana-Champaign, Urbana, IL, USA

Abstract: *Urine strips tests are used by laypersons and hospitals to gauge overall health of patients. From the perspective of laypersons, interpreting urine strip test results is difficult. We conduct preliminary study on accessibility of commercial urine strips that are on the market, identify problems and propose smartphone application to aid interpretation of test results.*

Introduction: Urine strips are used to study the person's health in diagnosing the person's health. These tests are conducted in both clinical settings and consumer settings. The strips test Leukocytes, Nitrite, Urobilinogen, Protein, pH, Haemoglobin, Specific gravity, Ketone, Bilirubin, and Glucose level[1]. These may serve as a good indicator in rapid diagnosis of related diseases[2]. Furthermore, significant number of laypersons buy self-diagnostic tools to gauge their health status. Self testing accounted for $13.8 billion[5] in 2011.

Motivations: We conducted a preliminary user study based on industrial design principles. We noticed two trends. First, asking users to regularly keep diaries of the results of their urine strips seem to have encouraged more health-conscious behavior ("I was very conscious of what I was eating and drinking throughout the day," "This exercise made me more conscious of how much water/liquid I drank each day,"). This type of behavior is also supported in existing health diaries[3] where users also cited improved self-observation as being the motivation for keeping health diaries. Despite these benefits, however, the users had difficulty interpreting the results of the urine strip tests ("I wish I could know what each of the colors of the test strip means"). In order to mitigate these problems, we propose smartphone software to aid interpretation of the urine strips. Naturally, the size of the market and usability, in particular, towards older generation is of interest. 38% of American adults have downloaded a smartphone application in 2011[4]. Even for the older generation, entering diary entries was not a deterrent as long as the software was easy to use[3]. The work further notes that older generations wrote more mobile phone diary entries than the younger generations. This suggests utilizing smartphones to interpret test results is a viable solution.

System Explanation: The goal of the system is to ensure the users do not have to manually distinguish color, and provide intuitive explanation of the results. The users will first take a picture of the urine strip that they have used. The users are then asked to drag the portion of the screen that contains the urine strip. Next, shape recognition detects all the pads used in urine strips. Color recognition is then run, and based on the test results, the system proposes what each pads signifies, along with overall test results evaluation. In order to serve as a user-generated health record, the system also records the test results. We note our program is a standalone software, and does not require any external devices unlike other existing solutions[6]. This allows more laypersons to use the software.

Conclusion: This poster will describe the motivation and system explanation. For our future works, we will evaluate how much the system has improved accessibility in using and interpreting urine test strip results.

References

1. Simerville, JA, Maxted, WC, Pahira, JJ (2005). Urinalysis: a comprehensive review. Am Fam Physician, 71, 6:1153-62.
2. Kim, DY, Kim, JH, Chon, CY, Han, KH, Ahn, SH, Kim, JK, Paik, YH, Lee, KS, Moon, YM (2005). Usefulness of urine strip test in the rapid diagnosis of spontaneous bacterial peritonitis. Liver Int., 25, 6:1197-201.
3. Mattila, E, Pärkkä, J, Hermersdorf, M, Kaasinen, J, Vainio, J, Samposalo, K, Merilahti, J, Kolari, J, Kulju, M, Lappalainen, R, Korhonen, I (2008). Mobile diary for wellness management--results on usage and usability in two user studies. IEEE Trans Inf Technol Biomed, 12, 4:501-12.
4. "Half of Adult Cell Phone Owners Have Apps on Their Phones." Pew Research Centers Internet American Life Project RSS. Web. 11 Mar. 2014.
5. "Point of Care Diagnostics." - HLC043C. Web. 11 Mar. 2014.
6. "New App Turns Your IPhone Into a Mobile Urine Lab." Wired.com. Conde Nast Digital, 24 Feb. 0013. Web. 11 Mar. 2014.

Misclassification of cases by querying modality: comparison of ICD-10 codes with clinical laboratory test results

Soo Yeon Cho, RN, MPH, Eun Kyoung Ahn, RN, MPH, Rae Woong Park, MD, PhD
Department of Biomedical Informatics, Ajou Univ. School of Medicine, Suwon, Korea

Abstract

Misclassification is an important bias in observational studies using electronic health records. This study compared case enrollment by two querying modalities: diagnostic codes and laboratory test results. In total, 7,385 cases of chronic renal failure were enrolled by diagnosis and 17,643 by laboratory test results. The number of enrolled cases was significantly different by querying modalities. In the intersection group (n=4,211), mean difference in the start of observation was significant between the two modalities (711±871 days). The querying modality should be carefully considered and evaluated.

Introduction

Misclassification bias is a major concern when enrolling cases using electronic health records (EHRs).[1] Many observational studies have depended on diagnostic codes alone for their case selection. As clinical laboratory test results are beginning to be available in EHRs, a combination of diagnostic codes and laboratory test results could reduce misclassification bias.[2] The aim of this study was to demonstrate the differences in enrolled cases according to the querying modality used: diagnostic codes versus clinical laboratory test results.

Methods

An EHR of a tertiary teaching hospital accommodating 1,201 patient beds was used for analyses. Chronic renal failure was used as an illustrative disease. Two different methods were used for case definition: the diagnosis group, in which ICD-10 codes of clinical classification software (CCS) for chronic renal failure were used (N18, N18.0, N18.8, N18.9, Z49, Z49.0, and Z49.1); and the laboratory group, in which an estimated glomerular filtration rate (eGFR) of <60 $mL \cdot min^{-1} \cdot 1.73\ m^{-2}$ for >3 months was used. The modification of diet in renal diseases study equation was used to calculate the eGRF. Differences in enrolled cases and the start of observation (index date) were compared between the two case definition methods.

Results

In total, 1,933,284 patients >19 years of age were included from July 1994 to September 2012. About twice as many cases were enrolled by laboratory test results (n = 17,643) than by diagnostic codes (n = 7,385). The proportion of female patients and the mean age were significantly higher in the laboratory group than in the diagnostic code group (50.9% vs. 40.9% and 62.1 vs. 57.3 years, respectively; $p < 0.001$ for both). The number of patients belonging to both groups (intersection group) was 4,211. The mean difference in the index date between the querying modalities in the intersection group was 711 ± 871 days. In this group, 881 patients (20.9%) had the same index date, while 742 patients (17.6%) were enrolled earlier by diagnostic codes and 2,588 (61.5%) patients were enrolled earlier by laboratory test results. In total, 156,271 medications were prescribed between the two index dates. Drugs requiring renal dosing for patients with chronic renal failure, such as amlodipine, furosemide, etc., were included among these medications.

Conclusion

The querying modality significantly affects case enrollment in observational studies using EHR data. Because the clinical characteristics of enrolled cases differ significantly according to the querying modality used, careful consideration and evaluation of the querying modality should be used in observation studies using EHR data.

References

1. Manuel DG, Rosella LC, & Stukel TA. Importance of accurately identifying chronic disease in studies using electronic health records. BMJ 2010;341:c4226.
2. Kadhim-Saleh A, Green M, Williamson T, Hunter D, Birtwhistle R. Validation of the diagnostic algorithms for 5 chronic conditions in the Canadian Primary Care Sentinel Surveillance Network (CPCSSN): a Kingston Practice-based Research Network (PBRN) report. J Am Board Fam Med. 2013;26:159-67.

Crowdsourcing and Development of Health-related Pictographs for Minority Groups by Gaming: A Focus Group Study

Carrie M. Christensen cMFA, Qing Zeng-Treitler PhD, Heather Aiono MEd, Bruce Bray MD, Erica Lake MLS, Marty Malheiro MS, Seneca Perri PhD
University of Utah Department of Biomedical Informatics – Salt Lake City, UT - USA

Introduction

We conducted a focus group with diverse members of Community Faces of Utah (CFU), to understand their health information needs and perspectives, and we have incorporated their feedback into the production of a new crowdsourcing game, Doodle Health. The web-based game will be used in the construction of an online library of health-related pictographs.

Methods

Five minority groups comprise CFU: Best of Africa, Calvary Baptist Church, the Hispanic Health Care Task Force, the Urban Indian Center of Salt Lake, and the National Tongan American Society. We conducted a focus group study to assess the health information needs of the minority group members and to obtain their thoughts regarding development of the crowdsourcing game. The two-hour focus group discussion was transcribed and qualitatively analyzed. We took the results of the focus group and met with Roger Alan Altizer, Jr. and his design team. Dr. Altizer is Director of Game Design and Production with the University of Utah Entertainment Arts Engineering Program. He has surveyed the capability of video games to capture the attention of our minds and bodies, providing unforeseen opportunities for individualized health care. [3]. He facilitated the "Design Box", a brainstorming method by which all stakeholders are encouraged to become involved in the game design process. The mobile web-based technologies used to build the game are easily deployable and can run across all kinds of devices and operating systems. HTML5, JavaScript and CSS were the development tools used, and the backend file system uses a combination of PHP and MySQL. During the game, players are given the choice to draw or guess images and the resulting files are sent to a backend study file for moderation/filtering and analysis.

Results

Adding illustrations to medical texts, patient discharge instructions and other health-related education materials can substantially increase comprehension, recall, and adherence rates [1][2]. Pictographs are not universally understood, however, and interpretation can take on a different color from another cultural perspective. The focus group helped the study team identify points of cultural sensitivity. Health-related images or topics that are part of the vernacular for some groups may be offensive or culturally taboo for others. For instance, topics of a sexual nature have to be approached delicately. A member of the National Tongan American Society shared, "Mere kissing was like, taboo, especially in mixed company when brothers and sisters were there. There's this very respectful kind of a thing where certain things are taboo in discussion, to watch or do."

CFU representatives were interested in gaining information about common diseases, such as diabetes and heart disease. And they wished to promote awareness leading to healthy diet and lifestyle choices for their communities. The health concepts people wanted to see illustrated included: nutrition, Body Mass Index (BMI), the cumulative effects of lifestyle and dietary choices on the human body, and comparisons of healthy food verses unhealthy food (including monetary comparisons). They thought that socioeconomic status was an important factor in health. Information about resources and the health insurance market was requested, in order to overcome economic barriers to quality health care. Technological literacy was identified as an area of concern during the focus group. CFU comprises a broad range of people and age groups that have limited experience and access to the Internet. Everyone in the focus group claimed to be using computers, smart phones, and/or tablets frequently, although the younger generation (under 55) is generally more comfortable using digital technologies. Most people have Internet access at home. If not, libraries and other community centers provide access. The participants expressed an interest in being challenged; yet they wanted the game to be user-friendly. They felt that social networking/interaction was a desirable aspect of gaming, and they provided a list of games popular within their communities.

Discussion

Doodle Health is more than an entertaining pastime. The crowdsourcing game might result in a more meaningful resource library of easily comprehended pictographs. During the participatory design process, minority consumers will be able to create/revise and assess the illustrations to ensure the end product meets their needs. The online library will be made freely available to clinicians to supplement text in patient health education materials. The ultimate goal of this project is to have a positive impact on clinical practices by facilitating and enhancing communication between health care providers and patients.

References

1. Zeng-Treitler Q, Kim H, Hunter M: **Improving Patient Comprehension and Recall of Discharge Instructions by Supplementing Free Texts with Pictographs**. In: *AMIA Fall Symposium Proceeding.* Washington DC; 2008: In press.
2. Austin PE, Matlack R, 2nd, Dunn KA, Kesler C, Brown CK: **Discharge instructions: do illustrations help our patients understand them?** *Annals of emergency medicine* 1995, **25**(3):317-320. Bruggers CS[1],
3. Altizer RA, Kessler RR, Caldwell CB, Coppersmith K, Warner L, Davies B, Paterson W, Wilcken J, D'Ambrosio TA, German ML, Hanson GR, Gershan LA, Korenberg JR, Bulaj G. **Patient-empowered interactive technologies**. So Transl Med, 2012 Sep 19;4 (152)152ps16.

Stage 3 Meaningful Use and Patient-Generated Health Data (PGHD): Outpatient Stakeholder Perspectives on How to Make PGHD Meaningful

Arlene E. Chung, MD, MHA, MMCi[1]; Katherine Treiman, PhD[2]; Carlton Moore, MD, MPH[1]; Christopher M. Shea, PhD[1], Jonathan S. Wald, MD, MPH[2,3]
[1]University of North Carolina at Chapel Hill, Chapel Hill, NC; [2]RTI International, Research Triangle Park, NC; [3]Harvard Medical School, Boston, MA

Abstract: Receiving patient-generated health data (PGHD) is a new measure in the proposed Stage 3 Objectives of the Meaningful Use (MU) Workgroup of the Health IT Policy Committee that would require eligible providers/hospitals to receive provider-requested PGHD, electronically submitted through either structured or semi-structured questionnaires (including patient-reported outcomes) or by secure messaging. A goal of this measure is to enable patients to contribute information to the electronic health record. However, little is known about stakeholders' perceptions of the value of PGHD in outpatient clinical settings.

Objective: To examine key stakeholder perspectives about the utility of the proposed MU Stage 3 objective to receive provider-requested PGHD and explore what workflow or electronic health record (EHR) innovations could help with implementation of this proposed MU objective.

Methods/Evaluation: We conducted semi-structured interviews of physicians, mid-level providers, nutritionists, pharmacists, practice managers/administrators, and nurses from various primary care practices and subspecialty practices at an academic medical center. Interviews focused on several areas: the value of the proposed objective, workflow issues with implementing the objective, EHR innovations needed to meet the objective, and improvement of the objective for implementation. Interviews were transcribed verbatim and then analyzed and coded using qualitative analysis software to identify major themes. Two coders independently reviewed each transcript, and consensus was reached for the predominant themes.

Results: We conducted 45 interviews across five outpatient primary care and subspecialty clinics at University of North Carolina Healthcare System. Overall, participants viewed the prospect of PGHD positively. Many clinics already collected PGHD via screening questionnaires, patient-reported outcomes, blood pressure/blood sugar/weight logs, and patient intake forms in paper format and—in one clinic—electronically. Providers felt that PGHD could be especially useful to track symptoms in patients with chronic conditions or who are actively undergoing treatment regimens such as chemotherapy or radiation. Participants felt the capability to collect these data outside of clinical encounters was valuable since collection between visits and prior to visits could improve efficiency and allow a better view into the patient's health and symptom status longitudinally. Participants did not have concerns about quality of PGHD and did not feel that a certification process was needed for devices used to collect PGHD. However, some had concerns about possible medical liability from receiving PGHD if appropriate follow-up was not provided.

In terms of EHR innovations that could facilitate receiving PGHD, stakeholders felt that various options for modes of collection were necessary to accommodate the heterogeneous needs of patients, including patient portals, in-clinic collection via tablets or kiosks, and interactive voice response systems via telephone. Participants also wanted a way to flag abnormal results and alert providers within the EHR so that appropriate and timely follow-up could be assured. PGHD should be labeled as patient-generated within the EHR so it is distinguished from measured or clinician-entered data from clinical encounters. In addition, patient portal systems should enable the upload of wearable and medical device data automatically via Blue-tooth technology or USB. Meaningful displays of PGHD within the EHR were also thought to facilitate the potential usefulness of PGHD.

Participants felt that the proposed objectives for PGHD were appropriate as a menu item rather than a core item. Moreover, they reported that data standardization across types and sources of PGHD was important both for integration of the data into the EHR and for research.

Conclusions: Overall, outpatient stakeholders had positive views about the proposed Stage 3 objective of the MU Workgroup of the Health IT Policy Committee for providers to receive PGHD; however, workflow considerations and EHR functionalities needed to support the receipt of PGHD are vital for the successful integration and use of this data.

Acknowledgement: *Funding was received from AHRQ through Contract HHSA 290-2010-00024i, Task Order 5. The views expressed are solely those of the authors and do not reflect the official positions of the institutions or organizations with which they are affiliated or the views of the project sponsors. The opinions expressed are solely those of the authors.*

Problem-Oriented Views Provide Cognitive Support to Decrease Drug Errors

Victoria Church, RN, MS, CNS-BC, Kathleen Adams, MPH
Portland VA Medical Center, Portland, OR

Problem Description

Optimizing health information systems is crucial to patient safety and the mission of the Portland Patient Safety Center of Inquiry (PSCI) and the Portland Informatics Center (PIC) at the Portland Veterans Administration Medical Center (PVAMC), a teaching hospital with complex disease processes, evidence-based protocols, and management strategies for a fragile geriatric population of patients. Many disease processes that were once managed by specialists are now expected to be managed by the general practitioner or internist. Additionally, clinicians must contend with fragmented information, functionally limited computer interfaces, and complicated medication distribution systems. The resulting cognitive burden associated with synthesizing volumes of information into problem identification, developing a treatment plan and assessing treatment response jeopardizes clinical outcomes.

Objectives

Reduce data fragmentation and enhance relevant data extraction from the electronic health record (EHR) for glycemic management, intravenous heparin infusion and medication reconciliation (MR) discrepancy detection.

Methods

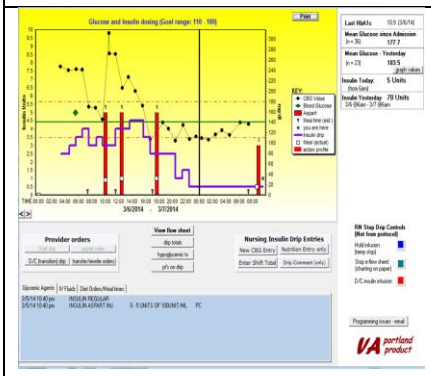
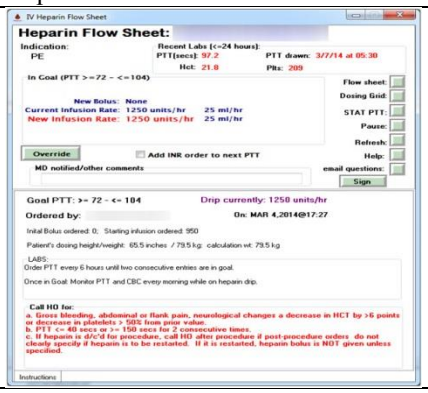
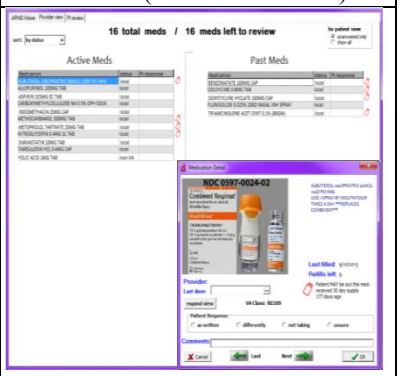
We used discrete multi-modal assessments to create and evaluate the integrated displays. Three different inpatient user-centered analyses identified critical decision support needs. Non-participant observations provided qualitative data on prototype usability and usefulness. Clinical subject matter experts further refined the prototypes using an iterative design process incorporating user feedback. Chart reviews provided quantitative data on length of stay and errors detection.

Results

These tools improve provider efficiency and enable identification of optimal patient management strategies resulting in increased Return on Investment (ROI). Use of the glycemic control tool decreased the average Length of Stay (LOS) for patients from 11.2 days to 9.34 days. Using the 2008 ADA estimate of daily inpatient cost related to diabetes care ($1853), a reduction in LOS of 1.83 days resulted in a savings of $3391per patient. Heparin tool use resulted in 82% reduction in incorrect protocol selection and correct drip rate adjustment; ADEs decreased from 17 to 0 immediately following implementation. Use of the MR tool detected an average of 6.9 discrepancies per admission with 68% of those rated high/very high for severity. Using an avoidable adverse drug event (ADE) rate of 118/year and the average cost developed by Bond of $2378 per ADE, an estimated $280,604/year could be saved.

Conclusion

The informatics tools and processes **that replaced error prone, paper-based processes significantly decreased drug errors, improved operational efficiency,** and **staff performance.**

Glucotron 5000	Heparinizer	APHID 3.0 (Staff view shown)
		

Understanding Primary Care Clinic Patients' Information Needs about their Clinic Visit

Martina A. Clarke, MS[1], Joi L. Moore, PhD[1,2], Linsey M. Steege, PhD[3], Richelle J. Koopman, MD, MS[4], Jeffery L. Belden, MD[4], Shannon M. Canfield, MS[5], Min Kim, PhD[1,6]
[1]Informatics Institute, University of Missouri, Columbia, MO; [2]School of Information Science and Learning Technology, University of Missouri, Columbia, MO; [3]University of Wisconsin-Madison School of Nursing, Madison, WI; [4]Department of Family and Community Medicine, University of Missouri, Columbia, MO; [5]School of Medicine Center for Health Policy, University of Missouri, Columbia, MO; [6]Department of Health Management and Informatics, University of Missouri, Columbia, MO

Abstract

To clearly understand patient information needs, promote informed clinical decision making, and quality patient care, 15 adult patients with acute illness and 14 adult patients with chronic disease were interviewed at a primary care clinic to identify what portions of the clinic visit note they viewed as important. Assessment and Plan were the most common note section identified as important by patients while Review of Systems was the note section identified least frequently as important.

Introduction

Better understanding of information needs is essential for providing patients with updated and relevant information for patient centered care. The purpose of this study was to identify patient information needs after a clinic visit with a primary care physician.

Methods

Two family medicine physicians (JLB and RJK) created fictitious but typical acute and chronic visit note. Sections in a visit note are: Chief Complaint, History of Present Illness, Past Medical History, Review of Systems, Physical Exam, Significant Lab Data, Assessment, and Plan. Fifteen patients with acute illness and 14 patients with chronic diseases participated in semi-structured interviews at the University of Missouri health clinics where patients were asked to identify important information they would need on a paper summary given to them after a visit with their primary physician. Data collection methods included a recorded interview containing 6 questions, Likert scales, Short Test of Functional Literacy in Adults and Patient Readiness to Engage in Health Information Technology survey. Thematic approach was used for qualitative analysis and differences in note section patient ratings of importance were evaluated by t-test.

Results

Table 1 Select examples of themes identified during data analysis and supporting quotes from interviewed patients.

Themes	Supporting quote
The Assessment (p=0.002) and Plan (p=0.045) were the most common note sections identified as important by acute and chronic patients when asked to choose three note sections to include in their end of visit summary.	"Well, again, because I would want to know for sure what the diagnosis was. I'd want to know what the update on it was, and I'd want to know what I need to do, the outcome of what I need to do because, that's the stuff I forget."
Review of Systems (p=0.003) was the note section identified the least by both patient groups when asked; patients viewed this section as unhelpful, stating that they already knew that information because it was about them.	"Cause I've already answered all those. I mean, it's just stuff I said. It's not his interpretation. There's not much to be confused unless he missed something, oh, he forgot to write down that my shoulder hurts, too."
67% of acute patients preferred online access to their information and 50% of chronic patients preferred paper access to their medical record (p=0.465). Acute patients wanted to be able to have access to their information at any time. One chronic patient had no interest in seeing their note.	"I guess just with the ease of the technology and basically that's it, less paper, less, just the ease of technology, basically that's all. It's something I don't necessarily want to leave laying around my house but if I have the availability to it, that would be nice."

Conclusion

This study demonstrates potential improvements to health information that patients receive after a clinic visit to support patient understanding of their plan of action. The results of patients' most commonly identified and least commonly identified note sections are also consistent with our other study findings identifying physician information needs. The results can be used to improve information that may be contained in patient health records (PHR) and patient portals. Further research is warranted to assess how to create accurate and reliable health information sources for patients.

Integrating Neighborhood Food Environment Data with a Comprehensive Community-based Survey Data to Support Population Health

Manuel C. Co Jr., MSN, MS, MPhil, RN[1], Suzanne Bakken, PhD, RN[1,2]
[1]School of Nursing, [2]Dept of Biomedical Informatics, Columbia University, New York, NY

Abstract

Geographic-level data can provide context to understanding the health of a community. This study reports our integrating neighborhood food environment data with our community survey data to enhance our understanding of the influence of place on health in a predominantly Hispanic low-income underserved urban population.

Introduction

Overweight and obesity affects lower-income communities of color living in areas with higher than average access to fast food restaurants and with limited access to healthy foods at reasonable cost such as those offered in supermarkets or other similar retail food outlets.[1,2] To gain a comprehensive understanding of those living in the Washington Heights and Inwood (WaHI) section of Northern Manhattan, this study was undertaken as part of the Washington Heights/Inwood Informatics Infrastructure for Comparative Effectiveness Research (WICER) Project to understand the influence of place on health in a predominantly Hispanic low-income underserved urban population.

Methods

The neighborhood food environment is characterized by integrating external geographic-level zip code data on food stores and farmers markets with geocoded WICER survey data on >5,000 WaHI residents. Food outlets were identified using the North American Industry Classification System (NAICS) definitions obtained from the ReferenceUSA national business database Major Industry Group codes and tabulated in Table 1. Data for the WaHI food environment is geocoded to transform street addresses into mappable data that can be spatially displayed and analyzed.

Result

Eight food outlet types were present in WaHI and total food outlets by zip code ranged from 46 to 79.

	Zip Codes					
NAICS Codes and Food Outlet Description	**10031**	**10032**	**10033**	**10034**	**10040**	**Total**
445110 Supermarkets and other grocery stores	45	64	53	34	42	238
445120 Convenience stores	2	2	3	3	2	12
445210 Meat markets	2	4	8		2	16
445220 Fish and seafood markets	1	2	4	1	2	10
445230 Fruit and vegetable markets		1	1	1	1	4
445291 Baked goods stores	2	2	4	1		9
445299 All other specialty food stores	1	1	1	2		5
447110 Gasoline stations with convenience stores		3	2	4		9
TOTAL	53	79	76	46	49	303

Conclusion

Integrating available external geographic-level data with our comprehensive community-based survey data provides context to our study of place and health of community residents with significant health disparities.

Acknowledgement: WICER (R01HS019853), WICER 4 U (R01HS022961), T32NR007969

References

1. Kumanyika SK. Environmental influences on childhood obesity: Ethnic and cultural influences in context. *Physiology & Behavior.* 2008;94(1):61-70.
2. Taylor WC, Poston WSC, Jones L, Kraft MK. Environmental justice: obesity, physical activity, and healthy eating. *Journal of Physical Activity & Health.* 2006;3:S30.

Electronic Health Record Systems (EHRS) Give Healthcare Providers a False Impression of Compliance with the Privacy and Security Meaningful Use Measure

Alex Cohn MSCIS[1], Andrea Bempong MBA[1], Kathy Fitzgibbon RN[1], Lanis L. Hicks PhD[2], Sam Ross, Theresa Walunas PhD[1], Adam Williams[1], Abel Kho MD MS[1]
[1]Northwestern University, Chicago, IL; [2]University of Missouri, Columbia, MO

Abstract

The HITECH act set up an incentive program for health care providers to implement and meaningfully use electronic health records systems (EHRS). In order to be eligible for incentive payments from the legislation, healthcare providers must meet a series of measures which include a privacy and security focused measure. This measure requires that providers conduct or review a security risk analysis of their organization. Some providers have the false impression that they are compliant with that measure even when no risk analysis has been done.

Introduction

In 2009 the American Reinvestment and Recovery Act included a section called the HITECH act. This legislation set up an incentive program for health care providers to implement and meaningfully use electronic health records systems (EHRS). In order to be eligible for incentive payments from the legislation the healthcare providers must meet a set of core measures and a set of their choice of menu measures. These measures are defined by the final rule for meaningful use released by the Department of Health and Human Services. One of the core measures focuses on privacy and security in health care practices and states the following objective:

"Conduct or review a security risk analysis in accordance with the requirements under 45 CFR 164.308(a)(1) and implement security updates as necessary and correct identified security deficiencies as part of its risk management process."

Meeting this objective is required for all healthcare providers who wish to receive incentive payments under this program. The formal process of conducting a security risk analysis and implementing a formal risk management framework presents a significant challenge to small health care practices and community health centers.

Study Design & Research Methods

The study was performed by distributing a link to an online survey that contained questions regarding EHRS adoption, and compliance with the privacy and security meaningful use measure. This link was emailed to healthcare providers who are members of the health information technology regional extension centers in the states of Illinois and Missouri and were believed to have achieved or were attempting to achieve meaningful use.

A portion of the data from the responses were analyzed to determine the understanding of the requirements for meeting the privacy and security meaningful use measure.

Results

We received 39 total responses (N=39). Of those respondents 37 had either attested that they had achieved meaningful use (MU) or that they planned to achieve MU. Of those 37 respondents 36 (97.3%) stated that they'd already met or planned to meet the privacy and security MU measure. Of those same 36 respondents 21 (58.3%) said that their EHR dashboard reported that they were compliant with the privacy and security meaningful use measure. When asked "What do you plan to do or have done to meet the security measure?" 7 (19.4%) of the previously mentioned 36 respondents indicated that their EHR told them they were checked and compliant with the security measure.

Conclusion

EHRS dashboards may be giving providers a false impression of compliance with the privacy and security meaningful use measure. The measure requires conducting a privacy and security risk analysis which is not recorded or entered into the EHRS. 20% of the respondents believed they were currently compliant based on the EHR dashboard. This survey represented a small sample of providers in two Midwestern states. A larger national study is necessary to confirm these results.

Clinical Informatics Program and Strategy to Support a Large-Scale EHR Implementation

Sarah Collins, RN, PhD[1,2,3], Saverio Maviglia, MD, MS[1,2,3], Perry L. Mar, PhD[1,2,3], Margarita Sordo, PhD[1,2,3], Li Zhou, MD, PhD[1,2,3], Charles Lagor MD, PhD[1], Roberto A. Rocha, MD, PhD[1,2,3]

[1]Clinical Informatics, Partners *e*Care, Partners HealthCare System, Boston, MA; [2]Brigham and Women's Hospital, Boston, MA; [3]Harvard Medical School, Boston, MA

Abstract

A well-organized Clinical Informatics (CI) program is essential for successful system implementations within large institutions. A defined CI strategy and approach can help ensure the institution's mission of providing high quality clinical care, conducting research, engaging in education, and serving the community.

Introduction

The implementation of an EHR system with advanced clinical features may require a cohesive CI strategy to realize improvement in quality and efficiency.[1] Partners *e*Care (P*e*C) is a large-scale effort to replace our existing systems with an integrated vendor-based solution. P*e*C includes a dedicated CI team responsible for all informatics-related activities and services. Current areas of focus include clinical decision support, knowledge management, interoperability (standards), and configuration management, and contributions to research and innovation (e.g., utilization monitoring, evaluation studies).

Methods

A cyclical framework that organizes CI services and activities was developed (Figure 1). The framework describes aims, roles, and deliverables from the CI program to P*e*C project teams. CI responsibilities (rounded boxes) and the responsibilities of implementation teams (text in brackets) are linked and mapped to typical project lifecycle stages.[2]

Results

We have been applying the CI framework with different projects (e.g., specification of decision support interventions, definition of shared data elements, mappings to standard terminologies) to facilitate and measure successful engagement. The framework helps explain roles and responsibilities of CI professionals, and also helps stakeholders understand how they can support CI professionals at different project stages.[3]

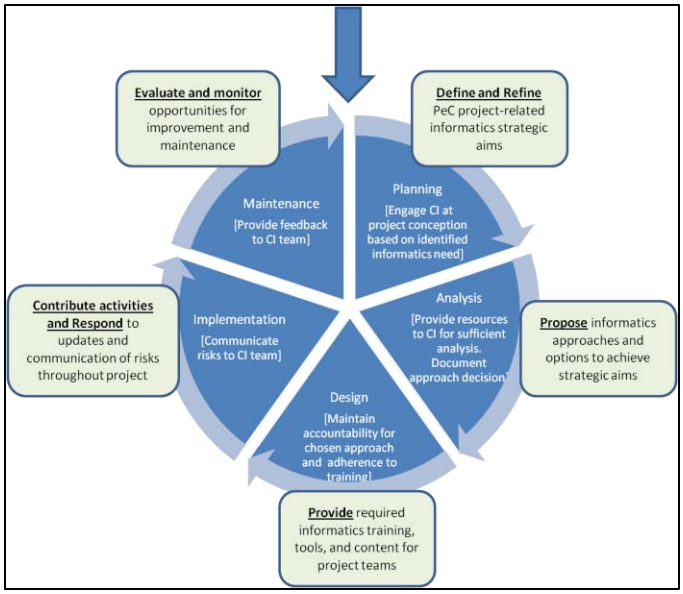


Figure 1 - Clinical Informatics (CI) framework defined for Partners *e*Care (P*e*C)

Conclusion

Ongoing P*e*C efforts underscore the importance of consistently providing CI expertise to help define and support EHR implementation activities. The proposed framework has established an effective mechanism for the CI team to support and collaborate with implementation efforts. Further work should evaluate its generalizabilty to other CI groups. The formal CI strategy promotes informatics best practices with proper utilization of resources, while ensuring that advanced EHR features are supported by evidence and acceptable to clinical users.

References

1. Chaudhry B, Wang J, Wu S, et al. Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. *Ann Intern Med*. 2006;144(10):742–52.
2. McGowan J, Cusack C, Poon E. Formative Evaluation: A Critical Component in EHR Implementation. *J Am Med Inf Assoc*. 2008;15(3):297–301.
3. Hersh W, Wright A. What workforce is needed to implement the health information technology agenda? Analysis from the HIMSS analytics database. *AMIA Annu Symp Proc*. 2008:303–7.

Rapid NLP Development with Leo

Ryan Cornia[1,2], Olga V. Patterson, PhD[1,2], Thomas Ginter[1,2], Scott L. DuVall, PhD[1,2]
[1]VA Salt Lake City Health Care System; [2]University of Utah, Salt Lake City, UT

Introduction
Structured data elements in a medical record poorly represent some aspects of a patient's state of health. Clinical text, on the other hand, can capture the specifics of complex and chronic conditions like the current stage of the disease, the rate of progression, and the ability of the patient to cope with and manage their condition. Natural language processing (NLP) unlocks the information in clinical text. NLP development kits, such as the Apache Unstructured Information Management Architecture (UIMA), provide common application programming interfaces (API) and data models that allow different developer groups to more easily collaborate.[1] UIMA Asynchronous Scaleout (UIMA AS) provides the ability to deploy NLP systems enabling scalable processing and can efficiently analyze large volumes of text. UIMA is a powerful and flexible framework, but in order to accommodate all the different functionality, configuration and deployment of systems built using UIMA, a set of XML descriptor files is required. Creating and maintaining these descriptor files manually can be complicated and time consuming. While UIMA components can be run within UIMA AS with no descriptor changes, additional descriptor files are required to deploy systems based on UIMA AS.
Despite the advantages of using UIMA, NLP system development continues to have a high entry barrier, requiring substantial setup, configuration, and framework specific knowledge. Projects like Apache uimaFit provides programmatic instantiation of UIMA components, but does not yet support UIMA AS deployment.
This poster introduces Leo, a Java-based framework built on UIMA AS that greatly reduces the barrier to entry by providing tools and base classes that facilitate rapid NLP system development and scalable deployment. Leo is named after the Spanish word meaning "I read".

System description
Leo was first developed to provide research teams in the Department of Veterans Affairs an accessible way to take advantage of the scalable processing provided by UIMA AS. The VA Informatics and Computing Infrastructure (VINCI) project led the extraction and aggregation of clinical text from VA medical records and now contains more than 2 billion clinical notes and reports. As many research teams did not have experience with enterprise software architectures, Leo provided libraries that would allow provide complex UIMA AS functionality with a simple API. Creating NLP pipelines, setting parameters, and defining annotation types are done programmatically, eliminating the maintenance required by UIMA descriptor files and allowing development to focus on honing NLP algorithms.

In addition to simplifying configuration and maintenance, Leo was further developed to include reusable, configurable components for finding keywords and phrases, using regular expressions and patterns, building complex rules for identifying and associating annotations within a document, mapping strings of text found in notes to standard terminologies, and developing feature vectors for training and validating machine learning components. Leo provides an extensive set of utilities for accessing documents and their associated annotations, reading and writing from various data sources, and performing annotation functions such as compare, find, and remove.

Leo supports the whole workflow of annotation, iterative NLP development, and validation by seamlessly connecting UIMA functionality with annotation and validation tools and machine learning libraries. The UIMA AS underpinning allows Leo to manage the scale needed for real-time processing and provides remote configuration tools for enabling automatic system optimization. It improves developer efficiency and simplifies NLP development. Leo has been extensively used on almost 100 projects processing millions of clinical notes.

Acknowledgements
This work was supported using resources and facilities at the VA Salt Lake City Health Care System with funding from VA Informatics and Computing Infrastructure (VINCI), VA HSR HIR 08-204 and the University of Utah.

References
[1] Ferrucci, D., & Lally, A. (2004). UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, *10*(3-4), 327–348.

The Texas Advanced Computing Center: A complete scientific discovery environment for biomedical informatics and health science research

Matthew C. Cowperthwaite, PhD[1], James D. Carson, PhD[1], John M. Fonner, PhD[1], Oscar D. Jiao, PhD[1], Jawon Song, PhD[1], Matthew L. Vaughn, PhD[1]
1. Texas Advanced Computing Center, The University of Texas at Austin, Austin, TX.

As the scale and diversity of biomedical data acquired in basic, translational and clinical research continues to grow, more powerful computing resources are needed to support the biomedical informatics community. The Texas Advanced Computing Center (TACC) is a leading academic computing center in the United States. We provide the medical informatics and computational biology communities with the most comprehensive scientific computing environment available, including more than 500,000 processing cores, over 100 petabytes of disk storage, dedicated large-scale visualization systems, and advanced networking and interface technologies. Using our systems, researchers can greatly increase the scale, speed, and throughput of their analyses, as well as conduct big-data studies that are impossible on standard workstation computers.

Life sciences computing at TACC is overseen by a dedicated group of scientists with expertise in genomics, molecular dynamics and docking, and image processing. The life-sciences group deploys and maintains more than 100 software packages of interest to the computational biology and informatics communities, including Matlab, R, Bioconductor, and FreeSurfer. TACC is a leader in the development of software tools for creating powerful, web-enabled scientific gateways such as iPlant and OpenfMRI; TACC also hosts and provide computing resources to the popular Galaxy informatics portal. Through these portal projects, we are able to offer advanced supercomputing technologies to more than 40,000 biologists and biomedical researchers.

TACC is also actively involved in training biomedical scientists to meaningfully incorporate computing technologies into their research programs. We offer roughly 40 courses and workshops via the NSF XSEDE program to more than 1000 researchers each year. As a result of these efforts, we have measured a fivefold increase in utilization of command-line interfaces to advanced computing technologies, to the point where they now comprise over 20% of our usage.

In summary, TACC provides a comprehensive computing environment to enable discoveries in the biomedical informatics space. Our resources are made available to the academic research community via the XSEDE program; industrial partners can work with TACC through our STAR industrial affiliates program.

Elders & Families Rely On Social Networks For Aging-Related Information: Implications For Informaticians

B.H. Crotty MD MPH[1], J. Walker RN[2], J. O'Brien BS[1], L Lipsitz MD[3], M Dierks MD[1], C. Safran MD MS[1]
Division of Clinical Informatics[1], Division of General Medicine & Primary Care[2] Division of Gerontology[3]
Beth Israel Deaconess Medical Center & Harvard Medical School, Boston, MA

Introduction

Aging creates new information and communication needs for elderly patients, as well as for their network of family members and caregivers. Recent data show that both elderly patients and their caregivers are turning to the Internet, but how seniors and families use technology to meet their aging-related information needs has not been fully explored.

Methods

We conducted a qualitative study of the information needs elders and families through a series of eight focus groups. Groups were comprised of either persons over the age of 75, or persons who care for a family member or friend over the age of 75. Groups lasted 90 minutes, and were organized around common experiences, such as managing chronic illness, recent acute care discharge, or assisted living. Groups were transcribed and reviewed using established qualitative methods, including immersion/crystallization.

Results

A total of 16 elderly participants (within 3 groups) and 22 family caregivers (within 5 groups) participated in the in the study. Participants characterized most needs as logistical including housing and transportation, access to basic health information, as well as communication with care providers.

We identified five core themes through our analysis, denoting that (1) seniors and families relied on their social networks for information gathering and appraisal, (2) information gathering requires active advocacy, (3) families often needed more information and communication during times of crisis, especially when occurring outside of business hours, (4) Internet users knew how to look for reputable information to trust, and most were wary of misinformation, and (5) tools and systems require sensitivity around issues of elder autonomy.

We found that seniors and families relied heavily on "word of mouth" and their social networks to find and appraise resources. Many participants supplemented their information gathering by Internet searches to validate what they heard through their social networks. Elderly participants frequently noted that "children do all the legwork," but many desired to be able to protect children from the burden of managing their medical conditions.

Discussion

Our findings are consistent with prior work that demonstrates the importance of social capital and social networks for health. In particular, seniors and families rely on others for information gathering and appraisal, using the Internet secondarily to supplement their research. Family members in particular expressed a desire to connect with one another for support. Participants desired tools to help fill current gaps, including times of 'crisis' or even 'routine' matters outside of business hours.

As the informatics community seeks to improve communication and information management for elderly patients and families, attention should be paid to social strategies and information tools to facilitate support for aging-related needs.

References

1. Miller WL, Crabtree BF. (1992). Primary care research: A multimethod typology and qualitative road map. In: Crabtree BF, Miller WL, eds. Doing Qualitative Research. Newbury Park, CA: Sage Publication.
2. Vedel, I et al, (2013). Health information technologies in geriatrics and gerontology: a mixed systematic review. JAMIA, 20(6), pp.1109–1119.

Override of Age-related Alerts in Older Inpatients: Evaluation of a Clinical Decision Support System

Olivia Dalleur, MPharm, PhD [1,2,3], Diane L Seger, RPh [1,4], Sarah P Slight, MPharm, PhD, PGDip [1,5], Mary G Amato, PharmD, MPH [1,6], Tewodros Eguale, MD, PhD [1,7,8], Karen C Nanji, MD, MPH [4,9], Nivethietha Maniam, BA [1,4], Patricia C Dykes, PhD, RN, FACMI [1,7], Julie M Fiskio, BS [1,4], David W Bates, MD, MSc [1,4,7]

[1]The Center for Patient Safety Research and Practice, Division of General Internal Medicine, Brigham and Women's Hospital, Boston, MA, USA; [2]Louvain Drug Research Institute, Universite catholique de Louvain (UCL), Belgium; [3]Cliniques universitaires Saint-Luc (UCL), Brussels, Belgium; [4]Partners Healthcare, Wellesley, Boston, MA, USA; [5]Pharmacy and Health, The University of Durham School of Medicine, Stockton on Tees, Durham, UK; [6]MCPHS University, Boston, MA; [7]Harvard Medical School, Boston, MA, USA; [8]Department of Medicine, McGill University, Montreal, Quebec, Canada; [9]Department of Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital, Boston, MA, USA

Abstract: *Clinical Decision Support systems (CDSS) can provide physicians with patient-specific recommendations, and suggest suitable treatment alternatives in case of drug-related risk. The objective of this study was to describe the overriding of alerts suggesting a medication substitution in patients aged 65 and older. A large proportion of these overrides occurred in patients aged less than 75 years and involved laxatives, psycholeptics and analgesics.*

Introduction: Several medications are not recommended for older patients.[1] The use of these inappropriate medications in older patients has been associated with adverse events, such as falls or delirium, and should be avoided. Clinical Decision Support systems (CDSS) can improve prescribing appropriateness in older inpatients (>=65years) by suggesting more suitable drugs. The objective of this study was to describe the overriding of substitution-suggesting alerts by prescribers in older inpatients.

Methods: The primary outcome measure was the count of age-related alert overrides over a period of three years (from Jan 2009 to Dec 2011) in a large academic medical center. Staff physicians, residents and non-physicians with prescribing authority were included. Demographic characteristics of patients were described. Drugs involved were categorized according to the Anatomical, Therapeutic and Chemical (ATC) classification system.

Results: A total of 33,141 overrides were observed during that period, equally distributed between 2009, 2010 and 2011. The overrides concerned 12,833 different patients (52% women). Average age was 74 at the time of override. A large proportion (57%) of these overrides occurs in patients less than 75 years old. Half of the alerts overridden concerned drugs used against constipation (bisacodyl rectal 45.0% and bisacodyl oral 5.2%). However, 26% of the overrides for rectal bisacodyl could possibly be considered as appropriate because these were single administrations. Other frequent overrides included psycholeptics (alprazolam 6.5%, diazepam 4.6%), analgesics (oxycodone 7.4%, meperidine 4.2%), psychoanaleptics (fluoxetine 4.0%, amitriptyline 2.1%), antiepileptics (clonazepam 6.2%) and anti-inflammatory drugs (2.6%). The top most frequent categories of drugs overridden did not differ when comparing patients aged less than 75 years old to those aged 75 and over (i.e. laxatives, psycholeptics and analgesics). However, the proportion of overrides of laxatives was higher in patients aged 75 and over (58.2% of all overrides in patients ≥75years vs. 46.8% in patients <75years, $p<0.001$), while psycholeptics and analgesics were detected in a lower proportion in these patients (13.9% vs 15.4%, $p<0.001$ and 9.7% vs. 13.4%, $p<0.001$, respectively).

Conclusion: The override of age-related alerts is an important problem. A large proportion of these overrides occurred in patients aged less than 75. Further work is needed to understand the reasons underlying these alert overrides, identify situations in which the override could be appropriate, and refine the alerts for optimized relevance.

References:

1.American Geriatrics Society updated Beers Criteria for potentially inappropriate medication use in older adults. American Geriatrics Society 2012 Beers Criteria Update Expert Panel. J Am Geriatr Soc. 2012 Apr;60(4):616-31.

This study was funded by grant # U19HS021094 from the Agency for Healthcare Research and Quality (AHRQ)

What women want? Expressing women's voice on contraception

Kavitha Damal, PhD, CCRC, Division of Epidemiology, University of Utah & Salt Lake City Veterans Affairs Medical center, Salt Lake City, UT
Rebecca Morris, MPH, Department of Biomedical Informatics, University of Utah, Salt Lake City, UT
Qing Trietler-Zeng, PhD, Department of Biomedical Informatics, University of Utah & Salt Lake City Veterans Affairs Medical center, Salt Lake City, UT

Background
Women in the modern age have several different contraceptive options to choose from, with each method carrying its own list of risks and benefits. While contraception counseling/ contraception education materials are abundantly availably, they are highly biased towards clinical details. Female values such as peer-experiences, cultural attributes are seldom addressed, leaving behind a knowledge gap that confounds the ability of women to make optimal contraceptive choices. We purport that contraception campaigns will benefit from addressing the attributes that women are keenly interested in knowing about such that a vast majority of unintended pregnancies across the world can be effectively avoided.

Methods and Results
To better understand contraceptive information needs and information seeking through online social media, we gathered 453 messages (165 questions and 288 answers) from three contraception-related online forums (ehealthforum, netmums and ivillage). A majority of all posts (73.5%) referred to the forum users' personal experience, while 38.2% of the questions specifically enquired about others' personal experiences with contraceptives. A review of these forum posts suggests a strong interest among online users to seek and share personal experiences. The most commonly discussed birth control methods include birth control pills, IUD, implant, and the injection. A content analysis was performed on the messages and 10 salient topics emerged as tangible attributes that are of high interest to the forum users, suggesting that information that seems implicit to providers and contraception counselors are still not as clear to patients. Patients crave for information from peers on an interpersonal basis and contraception education needs to necessarily include these patient perspectives in a transparent manner in order to reap more successful outcomes.

In addition, we performed a top model analysis using an adapted Latent Dirichlet allocation (LDA) method. The topic model analysis generated 27 stable topics, i.e. topics that are consistent regardless in multiple runs.

References
1. http://www.ivillage.com/contraception/4-k-28090
2. http://www.netmums.com/coffeehouse/woman-504/sex-contraception-48/
3. http://ehealthforum.com/health/birth_control_options.html
4. Trussell J. Contraceptive failure in the United States. *Contraception.* 2011;83(5):397-404.
5. Frost JJ, Singh S, Finer LB. Factors associated with contraceptive use and nonuse, United States, 2004. *Perspectives on sexual and reproductive health.* Jun 2007;39(2):90-99.
6. Hinyard LJ, Kreuter MW. Using narrative communication as a tool for health behavior change: a conceptual, theoretical, and empirical overview. *Health Education & Behavior.* 2007;34(5):777-792.
7. Turner G, Shepherd J. A method in search of a theory: peer education and health promotion. *Health Education Research.* 1999;14(2):235-247.

Title: Piloting a network of CTS2 terminology service nodes for value sets

de Coronado, Sherri L[1], MS, MBA, Wright, Lawrence W W[1] MA, Fragoso, Gilberto[1], PhD, Stancl, Craig R[2], Solbrig, Harold R[2], MS, Bauer, Herbert[2], Endle, Cory[2], Peterson, Kevin J[2]
1.National Cancer Institute, Bethesda, MD. 2. Mayo Clinic, Rochester, MN.

The biomedical informatics community has moved substantially towards use of common terminologies for annotating and retrieving data, documents and samples. However, the numerous terminologies are developed with different frameworks and models, and they are distributed using a wide variety of formats and channels. Terminology services such as NCBO's BioPortal [1], NCI's Enterprise Vocabulary Services (EVS) [2], NLM's Unified Medical Language System (UMLS) [3], EBI's Ontology Lookup Service [4], and CISMeF's HeTOP [5], provide access to multiple terminologies in a single place; each has its own unifying framework and format, but with different application programming interfaces that impede interoperability between services.

Common Terminology Services Release 2 (CTS2) is a recent Object Management Group (OMG) and Health Level 7 (HL7) standard model and specification for discovering, accessing, distributing and updating terminological resources on the Internet [6, 7]. One potential benefit to wider adoption of CTS2 for terminology services is the inherent ability to create a network of terminology services nodes. There are advantages for both terminology services providers and users. Service providers may still need to provide many of their services to end users who have differing needs through their regular services, but a network of CTS2 nodes would reduce the need for multiple service providers to host terminologies of shared interest. There could be a canonical version of a particular terminology hosted at a particular node accessed by many sets of users. For users, one advantage would be ability to access more terminologies with the same set of service calls, irrespective of the provider. Secondly, they would be more certain that the representation of a given terminology/value set is consistent and is the one intended by the owner, even if the distribution format does not follow other guidelines (e.g. [8]). We report on a pilot project implementing a network of CTS2 terminology service nodes. This effort is starting with value sets, a high value commodity to the community, especially for those needing Meaningful Use value sets in combination with other value sets.

Approach: CTS2 enables the same queries to be used across different implementations of CTS2 nodes with different terminologies, mappings, and/or value sets. This pilot project explored the ease of operationalizing this network of nodes. Three CTS2 value set services were used to test the functionality: (1) The NCI LexEVS 6.1 implementation of value sets, which includes value sets used by FDA and CDISC; (2) The Mayo Clinic implementation of the NLM Value Set Authority value sets for Meaningful Use [9]; and (3) the Mayo TLAMP (Terminology Linux Apache MySQL) CTS2 service that packages several standard terminologies into a standards-based terminology service and exposes HL7 code systems and value sets [10]. Each CTS2 service was originally set up for a different purpose and set of users based on specific requirements. During the prototype phase we analyzed the common functionalities that are currently exposed by each service, and also identified modifications needed in either the way the terminologies and value sets are loaded or how the CTS2 services are implemented. The goal is to provide access across the test nodes to a core set of functionality for accessing specific value sets. We also report progress and lessons learned.

References:
1. http://bioportal.bioontology.org/.
2. http://nciterms.nci.nih.gov .
3. https://uts.nlm.nih.gov/home.html ,
4. http://www.ebi.ac.uk/ontology-lookup/,
5. http://www.hetop.eu/hetop/.
6. http://informatics.mayo.edu/cts2/index.php/Main_Page .
7. http://schema.omg.org/spec/cts2/index.htm.
8. Terminology representation guidelines for biomedical ontologies in the semantic web notations.
Tao C, Pathak J, Solbrig HR, Wei WQ, Chute CG. J Biomed Inform. 2013 Feb;46(1):128-°© 38.
doi:10.1016/j.jbi.2012.09.003. Epub 2012 Sep 28.
9. https://informatics.mayo.edu/vsmc/.
10. http://tlamp.org.

Automatic Content Extraction for Designing a French Clinical Corpus

Louise Deléger, PhD, Cyril Grouin, PhD, Aurélie Névéol, PhD
LIMSI – CNRS UPR 3251, Orsay, France

Abstract

To develop resources and tools facilitating Natural Language Processing for the clinical narrative in French, we plan to select a representative sample from a large corpus of French electronic health records (EHRs). To access the most medically relevant content of EHRs, we develop an automatic system to separate the core medical content from other document sections. The performance of automatic content extraction achieves 0.96 F-measure, on par with human inter-annotator agreement of 0.98.

Introduction

Clinical corpora are necessary to develop resources and tools facilitating Natural Language Processing for the clinical narrative. Our long-term objective is to build a corpus of French clinical notes annotated with entities and relations to enable medical NLP research in French. To this end, we plan to select a representative sample from a large corpus of French electronic health records (EHRs) and to base our selection on the most medically relevant content of the clinical notes. Indeed, clinical notes often include header and footer sections listing the names and contact information of doctors in a healthcare unit. These sections are sometimes lengthy and they are not as helpful to characterize the clinical content of notes as sections describing patient care. Herein, we present an automatic approach to separate the core medical content from other document sections such as headers and footers.

Methods

For this study, we used documents from a corpus of 138,000 clinical notes from a group of French healthcare institutions. To build our system, we established a typology of the sections occurring in clinical notes. We considered 4 high-level section types: (1) a *generic header*, containing contact information for the health care unit in which the note was created. This header is the same for all clinical notes from the same unit; (2) a *specific header*, containing specific information such as the patient's name, birthday, date of admission, etc.; (3) the *main content* of a note; and (4) a *footer*, consisting of the physician's signature, together with greetings if the document is a letter.
We manually annotated 2 samples of randomly selected notes, by marking the beginning of each section (3 annotators participated in the process). Inter-annotator agreement for identifying main content lines was 0.98. Sample 1 (200 notes) was used as a development corpus to design our system, and sample 2 (500 notes) as a test set to evaluate the final system. We trained a conditional random field (CRF) model to identify the sections and extract the main content of notes. We classified each line of a document as belonging to one section type, using the BIO (Beginning-Inside-Outside) format. Features included line length, the first and second tokens of a line, the presence of blank lines before a line, etc. This approach draws on previous work on section identification from clinical notes[1] (e.g., "Past Medical History", "Discharge Medication") and scientific abstracts[2] (e.g., "Background", "Methods").

Results and Conclusions

The accuracy of our system is 0.955. Detailed performance per section type is shown in Table 1. Content is identified with a very high recall (0.989). Performance is very good, so that corpus studies can rely on our method for extracting medical content from clinical notes.

Table 1. Performance of the system

Section type	*Precision*	*Recall*	*F-measure*
Generic header	0.986	0.962	0.974
Specific header	0.975	0.926	0.950
Main content	0.916	0.989	0.951
Footer	0.950	0.901	0.925

References

1. Tepper M, Capurro D, Xia F, Vanderwende I, Yetisgen-Yildiz M. Statistical section segmentation in free-text clinical records. Proc of LREC, 2012.
2. Hirohata K, Okazaki N, Ananiadou S, Ishizuka M. Identifying sections in scientific abstracts using conditional random fields. Proc of IJCNLP 2008.

Data Transformation of Alzheimer's Data

Peehoo Dewan, BS, Naveen Ashish, PhD, Arthur W. Toga, PhD
University of Southern California, Los Angeles, CA

This poster describes a *data transformation engine* for "GAAIN" – the Global Alzheimer's Association Information Network where we are developing a pioneering integrated data access federation for data related to Alzheimer's research. The GAAIN effort is unique in being a first-of-a-kind initiative where Alzheimer's data can be accessed from multiple data networks around the globe in a seamless, integrated and secure manner. This work presents our work on the "GAAIN Data Transformer" – which addresses the specific aspect of transforming data from a GAAIN data *partner* to a common data model and standard developed by GAAIN. The overall GAAIN federation approach is based on a common data model, i.e., where data in each dataset from each data partner is *mapped* to a common data model for integration. The common data model is based on "CDISC" [2] elements and is under development.

For achieving this mapping in a scalable manner for each data partner, we have developed a GAAIN Data Transformer – which is a general-purpose software engine for transforming data partner data. Our paradigm assumes that a) Datasets (from data partners) will typically be available as "exports" in formats such as CSV, Excel or other, b) Data will be transformed to the GAAIN common data model, and c) The data transformer is configured declaratively through "transformation rules". The overall data transformation pipeline and the key modules are illustrated in Figure 1. The following are the key modules and their functionalities:

Database The transformer system includes an in-memory (relational) database, which is the staging database for the (original) data partner data (as is), as well as the cache for the data transformed to the GAAIN model.
(Data) Loader This module reads data from different export formats and loads it into the transformer database. We have written loaders for a variety of common export formats – for instance some data partners may provide all their data in one spreadsheet, others may organize as one (database) table per CSV file etc.,
Transformer Core (Mediator) The core of the transformer is an *Information Mediator* [1]. An information mediator is a general-purpose software system for integration of data from multiple heterogeneous data sources. Data from various sources is mapped to a common "global" model in the mediator. Data transformation is a special use case of the mediator where the mediator global model employed is the GAAIN common model, and the task of the mediator is to transform data from *one* source i.e., the data partner data into the global (GAAIN) data model. The transformation rules are based on *description logic* [1] and are expressive enough to capture *syntactic* as well as *semantic* data transformations.
Transformation Rules The mediator as transformer is powered by *declarative transformation rules* [1] which are logic-based rules that specify how exactly elements in the source data relate to the global GAAIN data model elements.

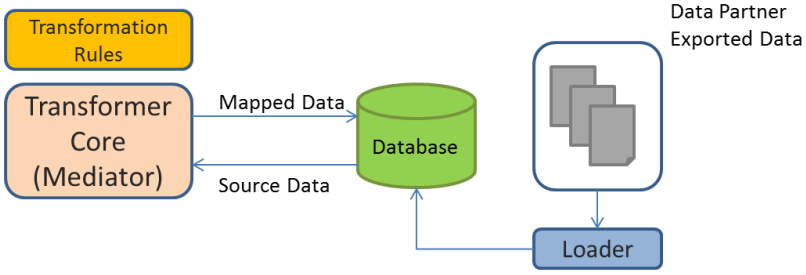


Figure 1. GAAIN Data Transformer

Currently we have developed a working first version of the transformer system and successfully employed it for data transformation from multiple disparate Alzheimer's partner datasets.

References

1. Ashish Naveen, Ambite José Luis, Muslea Maria, Turner Jessica. Neuroscience data integration through mediation: An (F)BIRN case study. Frontiers in Neuroinformatics Vol 4, 2010
2. CDISC: Clinical Data Exchange Standards Consortium (2014) Web: http://www.cdisc.org

A Semantically Enhanced Clinical Rules Management Repository Prototype

Sahithya Dhamodharan, MS[1], Adela Grando, Ph.D[1], Davide Sottara, Ph.D[1]
[1]Department of Biomedical Informatics, Arizona State University, Scottsdale (AZ)

Abstract

We developed the prototype of a Clinical Decision Support (CDS) artifact management platform targeted to a variety of stakeholders, like physicians, knowledge engineers and quality inspectors. The system is built on top of a CDS artifacts repository, allowing the retrieval of artifacts by matching their content using semantic queries. The system partially automates their management by the use of appropriate meta-rules such as "If a clinical rule is in Draft status for more than 90 days, send a reminder to publisher". To this end, we encoded the CDS logic using the national standard Health eDecisions (HeD) format, which supports a variety of metadata annotations describing and summarizing the artifacts and their content. For a chosen subset of those attributes, we defined an ontologic model of their admissible values and types. This model provided the vocabulary and the concepts used by the queries and the rules. We implemented the prototype customizing an open source semantic content management system and tested it simulating different scenarios.

Introduction

Organizations like Partner's Healthcare, the Mayo Clinic and Intermountain Healthcare own dedicated repositories for rule-based CDS artifacts such as alerts. They are manipulated through knowledge management portals that allow users to retrieve the content based on a variety of search options. However, many portals only support basic SQL-like query capabilities. Recently, it has been shown that semantic - rather than purely syntactic - queries simplify the formulation of the search criteria and improve the accuracy of the results by "understanding" the intention and the contextual meaning of the search terms[1]. Semantic-based criteria would allow a variety of stakeholders with different roles to retrieve meaningful sets of rules in a precise and context-specific way. Queries, however, still require an explicit action from the user. In some cases, instead, it is desirable to define policies – and their relative actions - that would apply to CDS rules automatically as they are stored into the repository. To enable these use cases, we defined our tool to support semantic "meta-rules" in addition to semantic queries. For example, imagine a specialist such as a Pulmonologist, who may be interested in artifacts, related to his discipline and may want to be notified whenever a new active artifact is made available in the repository. He would use the portal and make a search for rules that apply to pulmonary diseases. Unlike syntactic queries, which can only filter exact values, semantic queries can traverse taxonomies and relationships. So, the semantic query makes use of the metadata model and retrieves artifacts whose clinical applicability is Asthma, Pertussis or any other type of pulmonary disease. Also, when a new artifact related to pulmonary disease is added to the repository, a meta-rule in the repository like "If an artifact applies to pulmonary diseases and its status is active then alert physician" gets triggered and a notification about the new artifact is sent to the physician.

Methods

First, a literature review to gather the state of the portals helped us identify the search criteria and the use case scenarios for potential stakeholders. Second, we identified a suitable format for the encoding of the CDS rules. We chose the HeD standard because it is based on XML; it is structured and facilitates the automated processing of its content. Moreover, it provides a schema that supports a variety of metadata that was derived from the harmonization of other clinical and non-clinical standards[2]. Third, we selected a subset of the metadata attributes (demographics, clinical focus, provenance and status) and we defined a semantic representation of their domains. For each one, we assembled a simple ontology based on medical and administrative concepts. The ontologies provided the vocabulary to write the queries and the meta-rules, as well as the models used by a semantic inference engine to execute them. Next, we developed a user interface to allow users to define and execute the queries and meta-rules. This interface allows selecting the search criteria using visual widgets such as lists and pop up menus. Finally, we created a prototype implementation as seen in Figure 1, based on standards and open source components. We tested the prototype on the set of rules distributed by HeD as reference examples. The set of rules from HeD in XML format was mapped to OWL as it is semantically more expressive that XML.

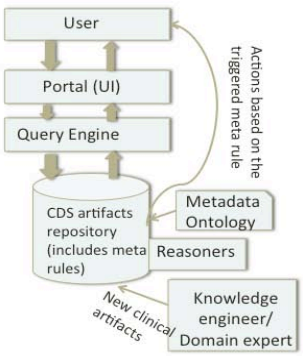


Figure 1: Clinical rules management repository prototype

Conclusion

Our preliminary results show that the use of semantic queries and rules can enhance the usability and usefulness of CDS knowledge repositories. Performance evaluation of the prototype with respect to precision and recall has to be done, once we expand and generalize our approach to include more attributes.

References

1. Kostić P, et al. Semantic Search Engine as Tool for Clinical Decision Support in Register for Acute Coronary Syndrome. *Telfor* 2011;3(1) :66-71

2.S&I Framework HealtheDecisions Initiative: HeD *CDS knowledge artifact implementation* https://code.google.com/p/health-e-decisions/source/browse/trunk/documentation/ (accessed 15 Feb 2014).

v3NLP Marshallers: Providing NLP Workflow Interoperability

Guy Divita[1,2], Brian Ivie[1,2], Qing T Zeng PhD[1,2],
[1]VA Salt Lake City Health Care System, UT; [2]University of Utah, Salt Lake City, UT

Abstract

Many commonly used NLP tools are used within medical community but the field has not yet standardized around one common interoperable plug and play messaging protocol/interlingua that satisfies all (sometimes conflicting) requirements. While interoperability efforts are underway, the necessity for systems to flow from one to another is being addressed within the v3NLP framework via marshallers that transform to and from other formats. V3NLP is a UIMA based framework. V3NLP's marshallers all transform into UIMA's CAS instances before being rendered out to other formats.The v3NLP marshallers are implemented as generic UIMA collection readers and collection writers and could be useful for stand-alone use to transform from one format to another, or within other UIMA based systems. The list below is by no means complete. There are marshallers that are waiting to be developed once there is an explicit need. The LAF/GrAF[1] ISO standard, now used for the American National Corpus Annotation effort, and a compact, terse marshaller that maximizes compression for efficient message transfer and file storage are next to be implanted. There are both general use GATE[2] to UIMA marshallers and some prior work done within v3NLP should a use case arise for such a need. The v3NLP marshallers are components of the v3NLP Framework, which is available via an Apache license and distributed from v3nlp.utah.edu.

Table 1: v3NLP marshallers

Marshaller	From	To	Comment
Text	X	N/A	No use case has come up to render back to text, yet.
UIMA XMI XML			This is UIMA's implementation.
Knowtator[3]/eHOST[4]	X	X	Knowtator and eHOST are useful full featured annotators.
CHIR Common Model	X	X	Used as an interlingua between v3NLP's lightweight Annotation viewer
CSV		X	There has not been a use case thus far to develop a marshaller to read from excel spreadsheets for further NLP processing, but it can be done.
BioC[5]	X	X	Distributed by NCBI.
Jdbc Database	X	X	Additional information about payload fields has to be provided.
Multi-Record Text Files	X	X	Useful to compactly bundle and read from a cohort corpus
VTT	X	X	VTT is a useful, lightweight portable Annotator distributed by NLM.

Acknowledgements

This work is funded by VA VINCI HIR-08-204 and CHIR HIR 08-374.

References

1. Ide N, Suderman K. GrAF: a graph-based format for linguistic annotations. Proceedings of the Linguistic Annotation Workshop; Prague, Czech Republic. 1642060: Association for Computational Linguistics; 2007. p. 1-8.
2. Cunningham H. GATE, a general architecture for text engineering. Computers and the Humanities. 2002;36(2):223-54.
3. Ogren PV. Knowtator: a prot\ég\é plug-in for annotated corpus construction. Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume: demonstrations; New York, New York. 1225791: Association for Computational Linguistics; 2006. p. 273-5.
4. South BR, Shen S, Leng J, Forbush TB, DuVall SL, Chapman WW. A prototype tool set to support machine-assisted annotation. Proceedings of the 2012 Workshop on Biomedical Natural Language Processing; Montreal, Canada. 2391141: Association for Computational Linguistics; 2012. p. 130-9.
5. Comeau DC, Islamaj Dogan R, Ciccarese P, Cohen KB, Krallinger M, Leitner F, et al. BioC: a minimalist approach to interoperability for biomedical text processing. Database : the journal of biological databases and curation. 2013;2013:bat064.

Assessing the Feasibility of Using Electronic Health Records for Community Health Assessments

Brian E. Dixon[a,b,c], P. Joseph Gibson[d], Karen Frederickson Comer[e], Marc Rosenman[b,f]

[a] Indiana University School of Informatics and Computing, Department of BioHealth Informatics
[b] Regenstrief Institute, Center for Biomedical Informatics
[c] Veterans Health Administration, Health Services Research and Development Service
[d] Marion County Public Health Department
[e] Indiana University-Purdue University Indianapolis, Polis Center
[f] Indiana University School of Medicine, Department of Pediatrics
Indianapolis, IN

Abstract

Assessment is a core function of public health. Comprehensive clinical data may enhance community health assessment activities by providing up-to-date, representative information used in the development or application of public health programs and policies. Greater access to clinical data is possible with electronic health record (EHR) systems. Yet public health stakeholders often question the reliability, timeliness, and accuracy of EHR data. We developed a matrix to support assessment of EHR data quality across a range of common community health assessment indicators. The matrix was found to be useful and helped identify target indicators for a single assessment project. We hope the matrix can be used and improved by others to support greater use of EHR data for community health assessment activities within provider organizations as well as health departments.

Introduction

Community health assessments (CHAs) provide information for population health problem and asset identification as well as public health policy and program formulation, implementation, and evaluation. Comprehensive healthcare encounter data may enhance CHAs with up-to-date, representative information for improved development and application of effective public health programs and policies. Traditionally CHAs have been performed using a limited set of information available through public data sets, behavioral surveys, and paper-based disease reporting. Given greater availability of electronic health record (EHR) systems, public health departments might leverage new, electronic data sources to support community assessment processes.

Methods

Through a series of meetings, we iteratively designed a multidimensional matrix to assess the feasibility of using routinely collected EHR data captured by clinical organizations for CHAs. The matrix rates aspects of EHR data quality, such as reliability, timeliness, and accuracy, across a number of common CHA indicators, including disease prevalence, health outcomes measures, and health service performance measures. The matrix is designed to be used by a group of public health stakeholders when defining information needs for a specific use case as outlined in (1). We employed the matrix with and gathered feedback from a group of public health informatics researchers and stakeholders in a metropolitan area to define indicators for a project involving a health information exchange.

Results

Prevalence of chronic diseases as well as several "high-profile" quality and health service performance measures were believed to be feasible to incorporate into CHAs in a "Most Wired" area of the country. However, comprehensive assessment must include integration across currently fragmented clinical and public health data silos. For example, colorectal cancer screening data are available electronically, but only a few providers make these data available to the regional health information exchange restricting analysis to localized areas.

Conclusion

As health departments plan and implement electronic data feeds from clinical organizations, they should consider how to aggregate data across silos for more effective CHA. In addition, health departments should consider how other EHR data sources might be leveraged to improve health planning and policy activities. The matrix was useful in choosing CHA measures which local stakeholder perceived as available through EHRs.

References

1. Dixon BE, Rosenman M, Xia Y, Grannis SJ. A vision for the systematic monitoring and improvement of the quality of electronic health data. Studies in health technology and informatics. 2013;192:884-8.

Pain Assessment Screening Tool and Outcomes Registry (PASTOR)

Nhan Do MD[1], Kim Heermann-Do MHA[2], Rick Barnhill BS[3], Diane Flynn MD[3], Ivan Lesnik MD[6], Eric Shry MD[3], Mary Ann Rubinos MD[3], Terry Newton MD[1], Kevin Galloway RN[1], Karon Cook PhD[4], Richard Gershon PhD[4], Leslie Barker, RN[2]; Chester Buckenmaier MD[5]

[1]Office of the Surgeon General, US Army, Falls Church, VA; [2]Defense Health Agency, Falls Church VA; [3]Madigan Army Medical Center, Tacoma, WA; [4]Northwestern University, Chicago, IL; [5]Defense & Veterans Center for Integrative Pain Management, Rockville, MD; [6]Naval Medical Center San Diego, CA

Abstract

The Military Health System (MHS) is exploring the use of patient reported outcomes (PRO) with Computer Adaptive testing (CAT) to drive consistency in pain management practices. PASTOR was developed to facilitate the integration of PRO within the clinical workflow and information systems of the MHS's Interdisciplinary Pain Management Clinics and Patient Centered Medical Homes (PCMH). Work is in progress to develop study protocols with PASTOR for comparative effectiveness research and expand its use for other clinical domains.

Background

The estimated healthcare cost of chronic pain in 2010 exceeds the combined cost of cancer and diabetes by 30%. Failure to adequately address pain will continue to result in escalating healthcare costs as well as loss of productivity and income. An Army Pain Management Taskforce was chartered in 2009 to make recommendations for a comprehensive pain management strategy, and one of the Taskforce's recommendations was to adopt a clinical information system that provides pain assessment screening with an outcome registry to promote consistency in pain care delivery. Some challenges with implementation of PRO are a lack of information technology infrastructure, lack of guidance for providers on how to interpret and use the results, and lack of clear understanding of when and how often to use PRO in clinical practice. PASTOR is a system based on National Institute of Health (NIH) Patient Reported Outcome Management System (PROMIS) platform to deliver CAT through various information communication modalities and provide decision support for patients and clinical staffs.

System Description

PASTOR was developed at Madigan Army Medical Center in partnership with Northwestern University. Workflow analysis was performed in Madigan's Integrated Pain Management Clinic (IPMC) and PCMH during system design. The IPMC enrolls around 40 patients per week with an intake questionnaire and a baseline PROMIS pain domain instrument. After initial assessment, multiple treatment modalities are offered to the patient and treatment outcomes are monitored with the PROMIS instrument at least monthly or more frequently depending on the treatment modality. Once appropriateness for discharge from the IPMC to the patient's PCMH has been determined, the patients' results are made available for the primary care team through an electronic dashboard developed for the PCMH workflow. The dashboard assists the healthcare team in identifying risks, summarizes key information, follows trends over time, identifies polypharmacy, and evaluates effectiveness of interventions. Results of the PRO instruments are stored in the registry for reporting as well as in a document repository in the electronic health record.

Conclusion

Monitoring patient reported outcome is part of the MHS's strategy to reduce variability in pain management and propagate best practices. We have developed PASTOR to reduce the burden of tracking and maintaining PRO over time through computer adaptive testing and an electronic dashboard that provides decision support for interpreting and managing results within the workflow of both pain specialty care and primary care clinics.

Gathering Information for Situation Models of Syndrome Identification

Kristina Doing-Harris PhD, Charlene Weir PhD
VA Health Care, Salt Lake City UT
University of Utah, Biomedical Informatics Department, Salt Lake City UT

Abstract: *Automatic identification of syndromic illness in EHRs could provide a substantial contribution to clinical decision support systems. Evidence for syndromes is most consistently documented in the narrative notes. Thus, its extraction will require natural language understanding. Cognitive Psychologists recognize that understanding language depends on situation models. The work in this poster begins the process of situation model development by reviewing charts to find descriptions of delirium. Even in a small sample of charts, situational descriptions are complex, overlap mainly on a point of omission, and indicate that a formal diagnosis is associated with worry. These situational findings are associated with the text's author leading us to conclude there are two relevant situational models, one for the text's subject and one for its author. We find that the author's model should include inferences about reasons for noting and coding, ideas of completeness and differential diagnoses.*

Introduction: Syndromic illnesses are some of the most difficult illnesses to diagnose because, by definition, they do not have definitive diagnostic tests. Therefore they are likely to benefit the most from clinical decision support (CDS). However, these CDS systems must be able to identify cases.

Delirium is an example of a syndrome that is difficult to diagnose based on chart review because it is often recorded in narrative notes, but rarely in ICD9-CM codes (paper presented here). This pattern may be due to the complicated patient situation associated with Delirium. While it is easy to recognize when a patient is delirious, the patient may only be given a diagnosis if the care team is worried their care being affected.

Therefore, the construction of a system to automatically identify syndromic illnesses will require natural language understanding. Cognitive Psychologists have recognized since 1983 that text based models of language comprehension are inadequate. The consensus is that situation models are combined with text-based representations.[1] Situation models are thought to consist of instantiated schemas, where different schemas apply to different aspects of the current situation. Schemas are prototypical instances of canonical situational aspects. The classic example is eating at a restaurant. A restaurant schema would include the activities "waiting for a table," "being seated," roles "waiter," "chef," and objects "cutlery," "napkin." Schemas allow the integration of world knowledge (in the form of expectations) with information from the current situation. In order to identify the scope of the Situation models and schemas for Delirium, we have undertaken a chart review to find the situations in which a patient is diagnosed with Delirium.

Table 1. Evidence of Delirium from three patient records.
Admitting diagnosis: altered mental status; Cognitive: Impaired vision, impaired hearing LOC: Calm, Confused Oriented to (person/place/time/situation): Person
…presents with altered mental status and report of several recent GLF's. Cognition: Patient is oriented to self only. He states that the month is July and could not offer days of the week, date or year. He shows inattention with inability to spell WORLD backward. DSM-IV DIAGNOSIS: Axis I: 1. Delirium 2/2 multiple etiologies 2. Dementia NOS (likely mixed)…
Admitted for altered mental status, failure to thrive in a setting of UTI. GENERAL – alert, oriented to person and time, NAD A gradual progress in what seems to be a chronic dementia Failure to thrive, gradually progressive weakness, confusion

Methods: The first author examined patient data, with IRB and DART approval, in the VINCI workspace. The patients were selected from the SLC VA population: over 65, inpatient stay in the last year on the acute medicine or telemetry wards. In this study 30 patients' notes for a single stay were read. 8 patients' notes were examined in detail to identify the diagnostic situation.

Results: Table 1 shows the evidence of delirium from three patient records. While both patients one and three were admitted for altered mental status, only patient two had an ICD9_CM code for Delirium, ground level falls (GLFs) and a Psychiatry referral with full work up. Both patients two and three have also been diagnosed with dementia. All three notes mention altered mental status and patient orientation, notable by the omissions of two or more aspects. The only patient with a formal diagnosis had an associated fall risk.

Conclusion: In reading the notes it became apparent that there were two situation models at work. In order to build a model of the patient's situation, it was necessary to construct a model of the note author's situation. Most importantly the note author's model must address the discrepancy between narrative descriptions and ICD9_CM codes. In the example notes, we postulated that the worry over fall risk motivated the formal diagnosis. Secondarily, it must address the author's notion of completeness in order to detect deliberate (as opposed to accidental) omissions. In addition, it should include the author's conception of the distinction (or not) between delirium and dementia.

1. Zwaan RA, Radvansky GA. Situation models in language comprehension and memory. Psychol Bull. 1998 Mar;123(2):162–85.

Individualizing Information Presented in Quality Dashboards: Preliminary Study

Dawn W. Dowding, PhD, RN[1,2], Yolanda Barrón, MS[2], Sylvia Ames, BA[2]
[1]Columbia University, New York, NY; [2]Visiting Nursing Service of New York, New York, NY

Abstract

Information regarding quality measures are often presented to staff in the form of dashboards using visualization techniques. The Visiting Nursing Service of New York (VNSNY) has over 150 such dashboards/interactive reports that can be accessed by administrative and clinical staff. This preliminary study explored the pattern of use of the dashboards over 1 year by users across VNSNY. The data will be used to develop strategies for individualizing the information provided in dashboards to users.

Introduction

Dashboards integrate and visually display information (often focusing on quality measures) to inform decisions about the delivery of patient care. [1] The Visiting Nursing Service of New York (VNSNY) currently provides over 150 customized dashboards/interactive reports (visualized displays of data) for administrative and clinical staff on a variety of quality and outcome measures. The purpose of this study was to explore patterns of dashboard use, with the aim of developing dashboards individualized to user information needs.

Methodology

Log files which list the individual user, which dashboard they accessed and date of access were obtained from the period 1st December 2012-30th November 2013. Data were analyzed to provide descriptive statistics. Heat maps (constructed in Microsoft Excel) were used to explore patterns in dashboard use.

Results

Over the 12 month period 1206 unique users accessed the dashboards, with a total of 30,029 hits. The role of the user and the department where they worked influenced the dashboards that were accessed most frequently (Figures 1 and 2). Users working in a clinical role were more likely to access the information contained in 'quality scorecards', which provide data on a variety of quality outcomes related to processes and outcomes of care. In contrast users with an administrative role were more likely to access information related to a patient's risk of rehospitalization.

	Quality Scorecards	Scorecard - by COC & CM Main Page	Scorecard by COC & CM	Scorecard - by Team	Scorecard - Patient Detail	Scorecard - Four Quadrant Dashboard	Scorecard - Measures	Main Landing Page	Patient List by Risk Score for the Hospital	Scorecard - High Level Summary (YTD)	Clinical Initiative Dashboard by Dx	VNSNY Glossary Dashboard
Nurse	26.09	23.73	23.7	6.47	6.48	1.76	1.98	1.65	0	0.53	0.04	0.15
Clinical/Provider Services Manager	22.19	13.79	13.42	11.93	6.77	4.7	4.26	2.35	2.04	1.06	0.1	0.07
Accounts/Business/Admin	11.34	3.13	2.95	4.49	0.55	4.97	3.9	3.65	19.29	1.69	8.91	0.59
Social Work	25.66	22.37	22.09	8.86	5.86	2.22	3.36	2.14	0	0.57	0.07	0.14
Staff Development/Education	24.55	11.88	11.38	14.33	7.07	7	9.08	2.38	0	1.89	0.23	0.15
Rehabilitation	25.84	8.77	8.45	15.31	4.47	9.89	8.93	1.12	0	3.83	0.48	0
Quality Improvement/Innovation	22.94	6.72	6.34	12.83	5.04	8.64	8.88	3.21	0	1.53	0.8	0.09
Research/Outcomes/Analyst	19.07	1.82	1.64	5.91	2.92	8.39	7.94	5.38	0	1.92	2.83	1.09
Technology/IT	25.68	5.41	5.41	2.7	0	6.76	6.76	4.05	0	6.76	0	12.16
Member/Customer Service	23.64	10.91	7.27	9.09	3.64	5.45	7.27	0	0	3.64	0	1.82
Other Clinical	26.63	7.61	7.61	10.87	3.26	17.39	10.33	0.54	0	2.72	0	0

	Quality Scorecards	Scorecard - by COC & CM Main Page	Scorecard by COC & CM	Scorecard - by Team	Scorecard - Patient Detail	Scorecard - Four Quadrant Dashboard	Scorecard - Measures	Main Landing Page	Patient List by Risk Score for the Hospital	Nursing Utilization Dashboard
Acute Care/CHHA	20.3	14.09	13.93	8.75	3.69	3.61	2.76	3.38	1.83	8.92
CHOICE	25.09	18.27	17.95	10.72	7.69	4.19	5.16	1.89	0	0.02
Children and Family Services	20.85	12.32	11.85	9.48	0.95	9.48	4.27	0.47	0	8.06
Congregate Care	24.26	15.65	15.02	12.99	2.5	3.91	1.41	1.25	0	7.2
Executive	7.37	2.11	2.11	2.11	0	6.32	5.26	0	0	2.11
Hospice/Palliative Care	35.07	4.48	4.48	7.21	0.5	3.23	1.49	3.73	0	2.49
IT	28.57	5.19	5.19	3.9	0	6.49	0	5.19	0	0
LTHHC	27.78	14.44	13.33	13.33	0	8.89	3.33	0	0	1.11
Operations (Supporting care programs)	7.79	2.22	2.05	2.62	0	2.16	2.5	2.79	28.82	1.14
Other Clinical Programs	23.53	12.16	12.55	10.59	2.35	10.2	5.49	3.53	0	2.35
Other Executive Support Functions/Programs	4.26	0.32	0.16	1.1	0	2.05	1.74	5.99	22.87	0.79
Quality Management Services/Education	22.88	7.41	7.13	12.88	4.62	8.87	6.48	3.4	0.08	2.11
Rehabilitation	26.12	1.26	1.08	14.54	6.1	15.8	13.55	1.71	0	0.36
Research Center	13.82	1.63	1.42	3.05	0.2	5.28	5.69	6.91	0	2.24

Figures 1 and 2: Heat map patterns of Dashboard Use by Role of the User and Department

Conclusions

Our analysis indicates that there is considerable variation in the information accessed by individuals depending on their role in the organization and/or the department where they work. There are some dashboards (e.g. the Quality Scorecard) that have utility for individuals across roles and departments, and others which are only useful for some users or areas of the organization. We will be using this information to develop strategies for individualizing the information provided in dashboards, to the needs of users. We will also be exploring how to present this information to users at the time and location of their decision making.

References

1. Daley K, Richardson J, James I, Chambers A, Corbett D. Clinical dashboard: use in older adult mental health wards. The Psychiatrist 2013;**37**:85-88

Using Laboratory Data for Prediction of 30-Day Hospital Readmission of Pediatric Seizure Patients

Amie J. Draper, BS[1], Ye Ye, MS[1], Victor M. Ruiz, BS[1], Christina Patterson, MD[2], Andrew Urbach, MD[2], Fereshteh Palmer, RN[2], Shouqiang Wang[2], Mark Somboonna[2], Fuchiang Tsui, PhD[1]
[1]Real-time Outbreak and Disease Surveillance Laboratory, Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA
[2]Children's Hospital of Pittsburgh of UPMC, Pittsburgh, PA

Abstract: *Laboratory data may provide valuable clinical information related to a patient's risk of readmission. This study assessed the contribution of laboratory data in predicting readmission risk for pediatric seizure patients. We extracted basic summary features from laboratory data and selectively incorporated them into a baseline prediction model to determine if readmission prediction accuracy could be improved. We found that the added features from laboratory data increased the model's AUROC from 0.55 to 0.63 with borderline significance (p-value=0.023).*

Introduction: Hospitals have been challenged to reduce pediatric readmission rates by 20% (recently estimated at 6.5%)[1]. Pediatric seizure is the fourth highest contributor to readmission risk[1]. To the best of our knowledge, prior studies have not identified readmission risk factors specific to pediatric seizure and have incorporated little to no laboratory data, which may provide information about a patient's condition[1-3]. We hypothesize that identification of patients at risk of readmission can be improved by utilizing laboratory test results. This study focuses specifically on pediatric seizure patients, but our hypothesis can be further tested on patients with other diseases.

Methods: We retrieved all in-patient electronic health records (EHRs) for visits to the Children's Hospital of Pittsburgh of UPMC during 2007-2012. Of these visits, 3,286 were seizure-specific based on discharge ICD-9 codes and 2,417 had laboratory test values for eight different test types (e.g. sodium). Abnormal/normal tests were identified according to age-specified normal ranges. We extracted basic features to summarize laboratory data for each patient, which included the percent of abnormal tests, the most recent test value, and the test type weight (Number of tests for a given test type/Number of all tests). The data were split into training (2007-2011; 234 readmissions, 1733 controls) and testing (2012; 75 readmissions, 375 controls) sets. A Naïve Bayes classifier using only patient age as a feature was our baseline prediction model. We used information gain ratio with 10-fold cross validation to rank each feature in the training dataset. Features were added individually in order of average merit to the baseline model and were kept if the accuracy of the model improved when using 10-fold cross validation on the training data. The final model was retrained and the testing data was used to compute the area under the receiver operating characteristic curves (AUROCs) of the baseline model and the new model, which were then compared.

Results: We found three laboratory features that improved the prediction accuracy of the baseline model: 1) percentage of abnormal blood nitrogen urea (BUN) tests, 2) the most recent potassium (K) test value, and 3) the most recent chloride (Cl) test value. The AUROCs of the baseline model and the model with the added features were 0.55 and 0.63, respectively (p-value=0.023).

Discussion and Conclusion: Clinically, derangement of homeostasis can potentially cause seizure[4]. The three automatically selected laboratory test types (K, Cl, BUN) reflect homeostatic conditions. The accuracy of a readmission predictive model for pediatric seizure patients increased when incorporating information from laboratory data, which implies that laboratory data may be useful to healthcare providers identifying patients with risk of being readmitted. Future work will include adding more clinical information for readmission risk assessment and examining additional patient populations.

References

1. Berry J, Toomey S, Zaslavsky A, Jha A, Nakamura M, Klein D, et al. Pediatric readmission prevalence and variability across hospitals. JAMA. 2013; 309 (4): 372—380
2. Sobota A, Graham D, Neufeld E, Heeney M. Thirty-day readmission rates following hospitalization for pediatric sickle cell crisis at freestanding children's hospitals: Risk factors and hospital variation. Pediatric Blood & Cancer. 2012; 58 (1): 61—65
3. Berry J, Hall D, Kuo D, Cohen E, Agrawal R, Feudtner C, et al. Hospital utilization and characteristics of patients experiencing recurrent readmissions within children's hospitals. JAMA. 2011; 305 (7): 682—690
4. Delanty N, Vaughan C, French J. Medical causes of seizures. The Lancet. 1998; 352 (9125): 383—390.

Title: Data for Patient-Centered Outcomes Research on Care Processes, Transitions, and Coordination

Theme: Clinical Research Informatics

Authors: Prashila Dullabh, Lauren Hovey, Michael Latterner, Samantha Zenlea, Petry Ubri

Introduction and Research Objective

There is a growing effort by clinical care settings, public health organizations, and researchers to share information in a way that enhances clinical decision-making and patient care in real time and contributes to a learning health care system.[1] This paper addresses important challenges and opportunities to patient-centered outcomes research (PCOR) on care coordination and transitions. Specifically, we focus on the data infrastructure necessary to support this research, including: 1) Actions needed to identify, share, and harmonize data across sources; 2) Essential data infrastructure at the individual and population health levels (e.g., data standards, security); and 3) Existing linkages, data sources, and services that could be leveraged.

Methods

This qualitative study included; 1) an environmental scan and literature review to identify key resources; 2) Semi-structured telephone interviews with 13 experts on data and research on care transitions and coordination; and 3) An in-person advisory work group to discuss a strategic roadmap for resolving data-related issues.

Principal Findings

Essential needs and areas for improvement include: standard measures to capture care coordination activities; strengthening the use of common data elements and standards, filling gaps in existing standards; mechanisms to collect, share, and incorporate patient-generated health information into PCOR; information exchange and interoperability among researchers, clinicians, and care coordinators; linking relevant data sources; and developing mechanisms that support quality improvement activities.

Conclusions

Here, we present a summary of essential needs, improvements, and mechanisms to support PCOR on care coordination and transitions, as well as opportunities for short term, strategic investments that alleviate common challenges and advance the field. In spite of their different vantage points, experts converged on a number of important opportunities to bolster data sharing, infrastructure, policies, and other accelerants of PCOR in this area. We also identify use cases for data on care transitions and coordination that would greatly benefit both clinical providers and researchers conducting downstream research to improve patient outcomes.

[1] The Learning Healthcare System, as conceived by the Institute of Medicine, refers to a vision of the healthcare system in which "science, informatics, incentives, and culture are aligned for continuous improvement and innovation, with best practices seamlessly embedded in the delivery process and new knowledge captured as an integral by-product of the delivery experience."

Check it with Chex: A Validation Tool for Iterative NLP Development

Scott L. DuVall, PhD[1,2], Ryan C. Cornia[1,2], Tyler B. Forbush, RCP, Corinne H. Halls, MS[1,2], Olga V. Patterson, PhD[1,2]
[1]VA Salt Lake City Health Care System; [2]University of Utah, Salt Lake City, UT

Introduction

Natural language processing (NLP) system development typically follows an iterative software development life cycle. Rapid error analysis and performance evaluation are essential steps in timely and successful system implementation. In many cases, an annotated dataset to support system development may not exist, or may be small due to cost and other factors, yet the developer must evaluate in order to iterate. This common scenario often results in highly skilled team members performing tasks that lie outside their skillset. Developers often take on the role of clinical experts in the iteration cycle because it has been challenging to display system output and collect feedback from clinical professionals in a system that is easy to set up and use. Likewise, clinical professionals struggle to provide feedback and clinical perspective in a format that can be easily implemented in the system design.

While multiple tools exist for chart abstraction (such as Knowtator or eHOST), none of these tools are optimized for system output review. To meet the need to support a rapid annotation, we developed a system, called Chex, that provides an easy, user-friendly way to validate system output for accuracy.

Chex was designed to enable developers to easily output system annotations to a database, and clinical reviewers and research team members to easily set up the desired workflow and rapidly validate the system output. Multiple human reviewers can validate on a project, if assigned by the administrator, allowing scale-out of reviewers to meet the needs of the developer and project. Each reviewer is provided annotations to validate as they complete them, eliminating the time needed for batching and managing reviewers' pace and workload. The results of the validation are written back to the database in a format that is directly consumable by the system developer or machine learning components, all resulting in much faster evaluation, and subsequent iteration cycles.

System description

Chex is a web-based application that can be configured for instance level or classification and extraction NLP system output. NLP system annotations and any associated features are displayed in a user-defined context window from the document. In cases where more context is needed, a single click will pop up the full document text with the annotation information. The interface can be customized by the administrator. In order to simplify management of annotations tasks, Chex is integrated with Business Process Management 2.0 for defining validation workflows, and utilizes a common database schema for reading and writing annotations. Chex uses the open source product Activiti (http://activiti.org/) for its workflow implementation, allowing for flexibility in how processes are defined and used, and following existing process definition standards. Chex is database platform independent, and can draw text and annotation data from most major databases.

Chex has two levels of access – administrator and user. Administrators have the ability to create new validation projects, create user accounts, assign specific users to annotation projects, and review summary reports. User level access is given to clinical specialists that serve as human reviewers. Annotation project definition includes the following elements: database connection information, annotation schema definition, a list of assigned reviewers, and annotation instructions for the chart reviewers. Once a validation project is created and users are assigned, each user has the ability to specify the number of annotations to draw for review and enter validation decision for each presented validation annotations. Each reviewer can be assigned to multiple validation projects at the same time, and each validation project can have multiple assigned reviewers.

Once NLP system output is reviewed and validated for accuracy, Chex summary reporting functionality can be used to review annotations at a project level or drill down to an individual annotation level.

Conclusion

NLP system development greatly benefits from an intermediary error checking. By supporting manual chart validation, Chex improves the iterative development process by providing timely error analysis, thus improving the quality of an NLP system.

Acknowledgements

This work was supported using resources and facilities at the VA Salt Lake City Health Care System with funding from VA Informatics and Computing Infrastructure (VINCI), VA HSR HIR 08-204 and the University of Utah.

Informational Needs During Intensive Care Unit Physician Handovers: A Multicenter Survey

Lewis Eisen, MD[1], Irene Yip[1], Pierre Kory, MD, MPA[2], Brian Gross, M.Sc., BSEE, RRT, SMIEEE [3], Brian Pickering, MD[4], Vitaly Herasevich, MD, PhD[4], Michelle Gong, MD[1]

1 - Montefiore Medical Center, Albert Einstein College of Medicine, Bronx, NY, 2 - Beth Israel Medical Center, Mount Sinai School of Medicine, NY, NY, 3- Philips Healthcare, Andover, MA, 4 - Mayo Clinic, Rochester, MN

Problem: The Joint Commission has identified communication errors during patient handovers as a potential source of patient harm. According to the Joint Commission handovers should 1) Identify relevant people related to patient care, 2) Provide clinical information, 3) Communicate issues specific to the patient, 4) Present expected trajectory, problems and solutions and 5) Assign accountability for monitoring and checking on the patient.
Current status: No electronic medical record (EMR) tool exists that cover all aspects of Joint Commission suggestions for adequate handovers even though clinical data relevant to handovers may be found in EMRs. In the data heavy environment of the intensive care unit (ICU), it is not clear what information intensivists think is important to communicate directly during handovers in the ICU and what data can be abstracted from the EMR.
Potential solution: AWARE (Ambient Warning And Response Evaluation) is an EMR vendor independent and real-time electronic dashboard for extracting and presenting high value clinical information at the bedside which could be modified to present important information for handovers [1].
Objective: To understand information needs during handover process in the ICU.
Results: A survey was administered to a convenience sample of 103 attending intensivists in 4 United States academic medical centers. The ICUs at all participating sites have electronic medical records(EMR) and one site already uses AWARE in the ICUs. Overall response rate was 41%. Most respondents hold primary board certification in internal medicine (66%) or anesthesia (32%). Respondents predominantly worked in medical (41%), surgical (17%), cardiothoracic (17%), burn (3%) or mixed medical and surgical (22%) ICUs. Most respondents work in large ICUs: 15% work in ICUS with >25 beds, 27% with 21-25 beds, 30% with 16-20 beds, and 28% with 10-15 beds. 94% had night intensivists. The respondents reported spending a median of 5.5 minutes signing out each patient after each shift and a median of 6.3 minutes signing out each patient at the end of a week.
The following handover survey items were deemed important by greater than 90% of respondents and are readily available in most EMR: Patient name, source of infection, additional organ dysfunction, chief complaint, mental status, hemodynamic status, code status, risk of deterioration, operation type, pertinent intraoperative issues, pending tasks, pending consultations, prioritization of pending tasks, opportunity for questions.
The following survey items were deemed important by greater than 90% of respondents and are not currently available in most EMR: poor patient response to specific therapies, variances from standard patient presentation, anticipated problems, recommended solutions to problems, contingency planning. The one site where AWARE is available in the ICUs was significantly more likely to identify items as not important to communicate in handovers as the information is already easily available to the clinician in the EMR and AWARE (median 8.5 interquartile range 3-14 vs. median 1 interquartile range 0-2) $p<0.001$)
Conclusions: This survey has identified several handover items deemed important by an overwhelming majority of intensivists. ICUs that currently use AWARE are more likely to state that information commonly included in handover is already easily accessible to the clinician. Using these survey results, a handover tool with AWARE has potential to be adapted to fulfill all Joint Commission suggestions and items deemed important by intensivists in our survey.
References:
1. Pickering, B. W., Herasevich, V., Ahmed, A., & Gajic, O. (2010). Novel Representation of Clinical Information in the ICU: Developing User Interfaces which Reduce Information Overload. Applied Clinical Informatics, 1(2), 116–31

Pilot Assessment of a Caregiver Decision Support Mobile Health (mHealth) Application for Food Allergy and Anaphylaxis in the School Environment

Christina Eldredge MD MS[1], Golam Mushih Tanimul Ahsan[2], Asriani Chiu MD[1], Brenda White Ed.S.[3], Taylor Atchison BS[1], Leslie Patterson PhD MS[1], Sheikh Iqbal Ahamed PhD[2]
[1]Medical College of Wisconsin, Milwaukee, WI; [2]Marquette University, Milwaukee, WI; [3]Archdiocese of Milwaukee Office for Schools, Milwaukee, WI

Abstract: *Food allergy is an increasing common student health condition in the school environment which requires prompt use of an epinephrine auto-injector in the event of accidental exposure. A mobile health (mHealth) caregiver decision support application for electronic food allergy action plans was piloted with teachers in the school environment using case-based scenarios. Initial teacher assessment of this application revealed the need for more student specific caregiver instructions and more emphasis on epinephrine auto-injector instructions.*

Learning Objective 1: Understand the need for caregiver decision support applications to help manage children with potentially life-threatening food allergy.
Learning Objective 2: Understand the importance of usability testing in mHealth application development.

Background: Food allergy is a chronic health condition and frequent cause of anaphylaxis (a severe potentially life-threatening allergic reaction) in children.[1] The prevalence of food allergy is rising among school-aged children with approximately 1-2 children in each class living with this condition.[2] Caregiver recognition of anaphylactic symptoms and immediate treatment with an epinephrine auto-injector is critical for child survival during an anaphylaxis episode as these incidents often occur in community settings (e.g. schools). National guidelines recommend the use of paper emergency action plans. However, many schools do not have on-site nurse support and students have frequent caregiver transitions (e.g. field trips, aftercare) which can lead to potential points of failure in school emergency health response. Therefore, there is a need for electronic exchange of student-centered food allergy action plans with interactive caregiver decision support, specific for use in the school environment.
Methods: An interactive electronic food allergy emergency action plan (eEAP) was developed by a multidisciplinary team for an iPad, iPhone, or Android device. eEAPs entered by school healthcare providers can be viewed by school staff securely at remote sites. The eEAP decision tree is based on evidence-based medicine and nationally accepted food allergy emergency action plans. Five teachers were recruited for pilot assessment of the system's accuracy and usability using two case scenarios developed in collaboration with an Allergist. The user's interaction was observed by two study members and documented via voice recordings. After testing, two post-survey tests were completed, one by the user and another by an observer.
Results: Four teachers participated in study as of July 2014. Teachers rated the app high on a 1-5 scale in both ease of use (4.74/5) and willingness to adopt the app at school (4.25/5). Three out of four participants did not correctly complete the first case scenario, all due to errors in epinephrine administration. However, all participants correctly made the decision whether to inject epinephrine. The second case scenario was completed successfully by all participants. Several usability and design issues were identified including the need for more "personalized" student eEAPs, improved student search mechanisms to decrease time to access of the eEAP, and improvement of the epinephrine auto-injector instruction screens.
Conclusion:
Initial usability assessment of this application revealed the teachers' need for more student specific caregiver instructions and more emphasis on epinephrine auto-injector instructions. System redesign is addressing these issues and lessons learned based on user input will be presented in November.

References

1. Boyce JA, Assa'ad A, Burks AW, et al. Guidelines for the diagnosis and management of food allergy in the United States: report of the NIAID-sponsored expert panel. *J Allergy Clin Immunol.* Dec 2010;126(6 Suppl):S1-58.
2. Gupta RS, Springston EE, Warrier MR, et al. The prevalence, severity, and distribution of childhood food allergy in the United States. Pediatrics. Jul 2011;128(1):e9-17.

The Need for a Nimble Decision Support Tool for Implementing Clinical Pathways in Oncology

Aymen Elfiky, MD, MPH(1,2); Adam Wright, PhD(2); Julie Bryar, MPH(1); Joseph Jacobson, MD (1); David Bates, MD, MSc (2); Edwin Rodgers (3); Kathleen Lokay(3); David Jackman, MD(1)
(1) Dana-Farber Cancer Institute, Boston, MA; (2)Brigham and Women's Hospital, Boston, MA; (3) D3 Oncology Solutions, LP, Pittsburgh, PA.

Problem Addressed:

Given the drive toward value-based health care, clinical pathways (CPs) are being used to plan, deliver, and review care. As an evidence-based clinical decision support (CDS) tool, CPs nonetheless require tailored implementation to reflect nuanced organizational infrastructure and care preferences. Concurrently, as a quality improvement tool, CPs help define anticipated care decisions and outcomes. As such, CPs content is continuously re-evaluated in light of evolving clinical evidence and practice metrics.

Within the context of a rapidly evolving oncology landscape, the challenges to implementing CPs within a tertiary cancer center include keeping pace with changing clinical content, quantifying levels of evidence, integration of a clinical trial portfolio, maximizing utilization of provider CDS to minimize unwarranted variability, and analyzing variability to identify improvement opportunities.

Purpose of System:

The Via Oncology Pathways portal is a nimble software platform that actualizes consensus-based, multidisciplinary clinical content at point of care through an interactive decision making tool. In parallel to provision of CDS, it also plays an important role in clinical practice improvement. Specifically, the portal is being piloted initially within the Genitourinary and Thoracic Oncology disease centers as part of a cancer care reform strategy at Dana-Farber Cancer Institute to coordinate an agile and iterative process of:

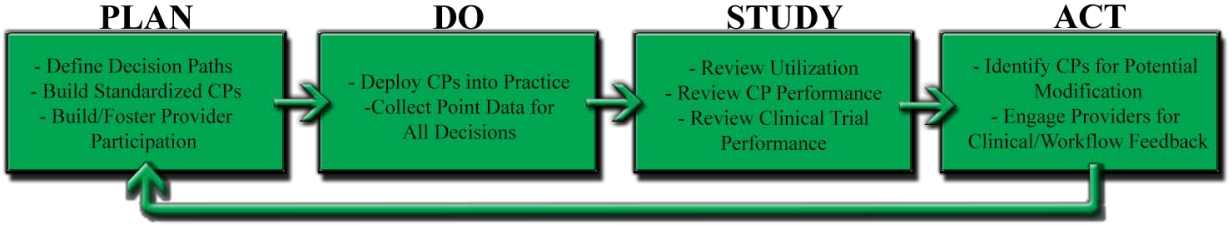
(i) clinical content development:
- defining distinct decision paths – branching, evidence and evaluation criteria within branches
- quantifying levels of evidence
- building discreet, standardized CPs into an algorithm
- incorporation of local preferences including clinical trials

(ii) delivery of algorithm in a user interface:
- deploy an XML-based "Disease Version" reflecting above local preferences
- employ a runtime environment to display and capture clinical info discreetly
- interface with practice management system to present patients respective to the provider
- consistently collect all point data for "on-" and "off-pathway" decisions

(iii) assessment of platform use:
- review utilization/login by providers
- review CPs performance in capturing "on-" and "off-pathway" rates
- review of clinical trial performance i.e. presentation of active portfolio, number of patients enrolled in trials, number of screening failures

(iv) process of Via platform modification:
- identification of specific CPs for potential modification based on "on-"/"off-pathways" rates
- continued engagement providers in workflow feedback and CPs updates

The above process can be conceptualized within context of a quality improvement framework as follows:



Future Directions:

These data provide the foundation for other analytics which allow for total quality management and improving both clinical and research programs. Via Pathways platform performance and feasibility will help dictate decision by the cancer center to scale the platform's use in the other disease centers.

eHealth Strategy transforming health care delivery in Tanzania

Mturi Elias[1], Jovitus Mukela[2], Hermes Rulagirwa[2], Prof Andrew Swai[2], Happy Danford[2], Sri Perera[3], Niamh Darcy[1] • [1]RTI International, NC, USA; [2]MOHSW, Dar Es Salaam, Tanzania; [3]CDC, Dar Es Salaam, Tanzania

Tanzania national eHealth Strategy 2013-2018

1. Background

- There is increased interest by national and local governments, partners, and private institutions to invest in global and national eHealth initiatives..
- eHealth can transform health care delivery by enabling information access and supporting health care operations, management and decision-making. The successful implementation of eHealth requires strategies that are aligned with the national respective health priorities.
- There are several models or frameworks for developing a national eHealth strategy, often developed by technologists and thus are complex to understand and use by health system actors; Often there is a missing causal link between health sector priorities and how technology can be applied to these priorities

2. Objectives

- Develop a pragmatic eHealth strategy framework that is guided by health sector priorities and also easy to implement
- Demonstrate feasibility by applying this framework to develop the national Tanzania eHealth Strategy 2013-2018

3. Methods

- Review existing strategy and eHealth strategy methodologies. This includes the
  - WHO eHealth Strategy Toolkit[1], Business Motivation Model[2] and Ishikawa Fishbone Diagram Strategy development model
- Review existing national eHealth Strategies including Canada[3], Australia[4]
- Complete a practical application of this eHealth strategy framework in Tanzania over 12 months (Sep 2012 to Sep 2013)
- Final launch of the Tanzanian eHealth Strategy 2013-2018 on 30 September, 2013 and national eHealth Steering Committee inaugurated on the same date

4. Results

- The eHealth Strategy was guided by the Tanzania Health Sector Strategic Plan (HSSP) III (2009-2015), which identifies health sector priorities.
- The initial background work occurred over a period of 2 years, culminating in a 1 week multi stakeholder workshop (health sector and ICT/eHealth experts) in September 2012 that used this framework to develop the eHealth Strategy
- Continued Ministry of Health and Social Welfare (MOHSW) and stakeholder review resulted in a final eHealth Strategy publication in September 2013.
- The following figures illustrate the 1) eHealth Strategy Development Framework, 2) mapping of the HSSP III (2009-2015) health sector priorities to the eHealth Strategic Objectives and 3) the four eHealth Pillars and their related Strategic Objectives

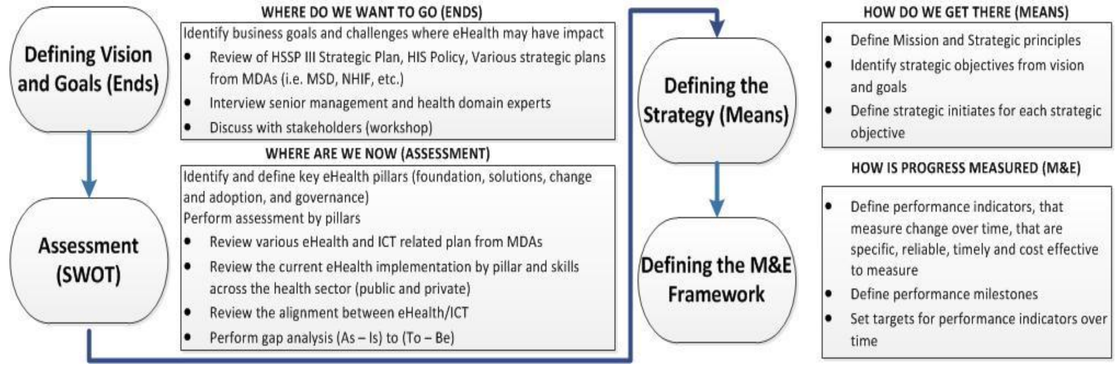


Figure 1: eHealth Strategy Development Framework

Health Sector (HSSP III) Strategies	eHealth Strategic Objectives																	
	SO1	SO2	SO3	SO4	SO5	SO6	SO7	SO8	SO9	SO10	SO11	SO12	SO13	SO14	SO15	SO16	SO17	SO18
District Health services	√			√					√	√	√							
Referral Hospital Services				√					√	√								
Central-Level Support																		
Human					√						√							

Figure 2: Linking eHealth Strategic Objectives to Health Sector (HSSP III) Strategies

Foundations	Solutions
SO1. Enhance ICT infrastructure and services to improve communication and information sharing across the health systems and at all levels	SO4. Enable electronic financial management to ensure effective collection, allocation, and use of health financial resources at all levels in accordance with health plan priorities
Change and Adoption	SO5. Strengthen an electronic HR system to improve planning and management of health professionals at all levels
SO15. Establish a comprehensive change and adoption strategy to promote and enforce the development and use of eHealth solutions for both public and private institutions at all levels	SO6. Enable an electronic logistics and supplies system to ensure adequate quality and quantities of health commodities are always available at the point of service to meet patient demand
Governance	SO7. Enable electronic delivery and interventions of health services to reduce child mortality ; maternal mortality ; and the burden of HIV/AIDS, TB, malaria, and non-Communicable diseases
SO16. Establish an eHealth governance structure and mechanism to ensure effective management and	SO8. Strengthen an electronic health management information system (HMIS) to support evidence-based health care and decision making

Figure 3: eHealth Strategy Pillars and Strategic Objectives

5. Summary

The Tanzania eHealth Strategy 2013 – 2018 was launched on 30 September, 2013.

The national eHealth Steering Committee was inaugurated on 30 September, 2013 and has met several times.

6. eHealth Strategy Vision & Mission

Vision: By 2018, eHealth will enable a safe, high-quality, equitable, efficient and sustainable health system for all citizens by using ICT to enhance planning, managing, and delivering health services.

Mission: To support the transformation of the Tanzanian healthcare system by leveraging ICT to improve the health and social welfare of all citizens.

7. Limitations

Known limitations of the eHealth Strategy framework

- This framework has only been applied in Tanzania
- The eHealth Strategy implementation is under way but is not yet complete

References

1. World Health Organization and International Telecommunication Unit, National eHealth Strategy Toolkit. 2012. Available at: http://www.itu.int/pub/D-STR-E_HEALTH.05-2012
2. The Business Motivation Model, The Business Rules Group, 2007, Release 1.3
3. eHealth Strategic Framework. British Columbia eHealth Steering Committee. November 2005. Available from http://www.health.gov.bc.ca/library/publications/year/2005/ehealth_framework.pdf.
4. National eHealth Strategies. International Society for Telemedicine and eHealth. Available from http://www.isfteh.org/media/category/national_ehealth_strategies.

Acknowledgments

The related project has been supported by the U.S. President's Emergency Plan for AIDS Relief (PEPFAR) through the U.S. Centers for Disease Control and Prevention (CDC). Support has also been provided by the Embassy of the Kingdom of the Netherlands, the Ministry of Health and Social Welfare (MOHSW) Tanzania and the World health Organization (WHO).

More Information

***Presenting author**: Niamh Darcy
919.485.2610
ndarcy@rti.org

RTI International
3040 East Cornwallis Road, Research Triangle Park, NC 27709-2194

Point-of-Care Knowledge-Based Resource Needs of Clinicians: A Survey from a Large Academic Medical Center

MA. Ellsworth, MD[1], JM. Homan, MA[1], J. Cimino, MD[2.3], SG. Peters, MD[1], BW. Pickering, MD[1], V. Herasevich, MD, PhD[1]

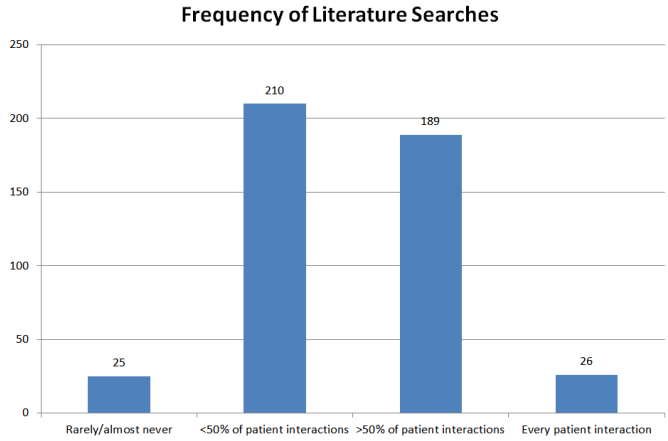
1 - Mayo Clinic, Rochester, MN, 2- Laboratory for Informatics Development, NIH Clinical Center, Bethesda, MD, 3 - Department of Biomedical Informatics, Columbia University, New York, NY

Abstract: *A single center survey demonstrated that medical providers frequently perform literature searches to answer clinical questions related to recent patient interactions. There is a distinct preference for using synthesized information sources mainly for the diagnosis and therapy domains. Adapting point-of-care information tools to match the preferences of the users potentially increases their usage rates and overcomes barriers often cited for not performing knowledge searches.*

Introduction: The development of point-of-care context-sensitive information retrieval tools, termed infobuttons, link local clinical information systems to electronic knowledge resources to aid in clinically related knowledge searches. Adapting infobuttons to the clinician's needs are necessary to provide resources that are tailored to their specific preferences in an effort to increase the sensitivity of their use and limit information overload. We conducted a survey at our institution to better understand the literature searching tendencies of clinical providers in order to best create adaptive infobuttons suited to their needs.

Methods: A survey regarding literature searching preferences was sent to 1862 unique clinical providers throughout the Mayo Clinic. The survey consisted of 25 items asking respondents to select which clinical scenarios most often prompt literature searches as well as identify their most preferred knowledge resources. Demographic data regarding the respondent's clinical role and area of patient care were also included. The clinical role or practice area for individuals offered an invitation to respond could not be determined (it could only be determined in returned surveys) limiting the ability to determine response rate by role.

Results: A total of 450 completed surveys were returned and analyzed (24% response rate). The number of respondents varied according to clinical role and patient care area (outpatient – 40%, OR – 24%, hospital floor – 21%). 48% of respondents perform literature searches for more than half of their patient interactions with 91% of all searches occurring either before or within 3 hours of the patient interaction. When a search is performed 57% of respondents prefer synthesized (i.e. UpToDate) information sources as compared to only 13% who prefer original research (i.e. PubMed). The two most common clinical domains that prompt literature searches are therapy (80%) and diagnosis (46%) with only 13% of respondents frequently searching for prevention related clinical answers. UpToDate was the most selected preferred reference in all 4 clinical domains with 68% of respondents selecting this reference in both the therapy and diagnosis domains. The 2 next most popular sources for therapy questions were MEDLINE and Micromedex with Google and MEDLINE following for diagnosis related questions. Additionally, more than 30% would ask a colleague and/or reference AskMayoExpert when faced with a therapy or diagnosis question. 82% of knowledge searches are performed on a workstation or office computer with just 10% occurring on a mobile device or at home. Finally, the number one identified barrier to an effective literature search was lack of time.



Discussion: Providers in our survey demonstrate a large need to answer clinical questions on a regular basis, especially in the diagnosis and therapy domains. Most of these searches occur in the patient care setting within a very short time from the patient interaction using synthesized (i.e. UpToDate) knowledge sources. Reasons for these preferences could not be gleaned from this survey design. However, these findings underscore the need for the creation of point-of-care tools that can guide clinicians to clinical answers quickly and efficiently. Adapting infobuttons that include only sources for specific domains that are preferred by the user would best accomplish this and provide an environment where quick literature searching is accomplished while at the same time avoiding the complication of information overload.

Development of an Efficient, General Purpose TCP/IP Listener in Perl to Capture HL7 Data for Real-Time Clinical Decision Support

Richard H. Epstein, MD, CPHIMS;[1] Michael Perino, BS;[2] Jerry Magrann, MS[2]
[1]Jefferson Medical College and [2]Thomas Jefferson University Hospital, Philadelphia, PA

Abstract
We developed a general purpose Perl script to receive HL7 data from an interface engine (via a TCP/IP socket) and store it in a relational database. Our method bypasses limitations of EHR databases optimized for transactional throughput rather than query performance, thus making the data accessible for real-time clinical decision support. At least 1295 transactions per minute can be processed with perfect accuracy.

Introduction
To implement automated clinical decision support (CDS), algorithms require real-time access to electronic health record (EHR) data, much of which originates outside the EHR and is transmitted using the HL7 standard. Interface engines are deployed to receive such data and populate the EHR database. However, these databases are often designed primarily for transactional throughput, rather than for query performance, a CDS requirement. Additionally, important HL7 data are frequently transmitted but not stored in the EHR database. The purpose of this project was to develop an efficient, general purpose method using open source tools to capture HL7 data and to store that information in a relational database, making it available for CDS. We assumed that organizations would already have an HL7 compatible interface engine. The use case was a need for real-time monitoring of transactions in the perioperative period involving controlled substances ("narcotics") executed by anesthesia providers on a distributed network of approximately 50 Pyxis® (CareFusion) automated drug delivery cabinets (ADDCs).

Methods
The open source language Perl (www.perl.org) was used to develop a short "listener" script to which our interface engine (Cloverleaf®, Infor) connected over a persistent TCP/IP socket. The open source IDE platform Padre (www.padre.perlide.org) was used for software development and testing on a network server hosting a Windows 7 virtual machine (VM). The CPAN library *dbi* (search.cpan.org) was used to enable database connections. The dbi:ODBC:driver = {SQL Server Native Client 10.0} was loaded *via* the complimentary installation of SQL Server Express 2008 R2 (Microsoft) on the VM.

HL7 vending transactions are filtered by the interface engine according to the physical location of the ADCC and scheduled drug class of the transaction. Fields from the MSH, PID, FT1, and ZPM segments (as identified in the Pyxis HL7 reference manual) are concatenated into a pipe delimited string for each transaction using Tool control language (Tcl), embedded in the Cloverleaf Interface, and then sent to the perl listener using the TCP/IP socket. Data are received to a buffer by the listener, logged locally to a text file on the VM (for debugging), and written to a SQL 2008 R2 database, located on a separate server. An acknowledgment (ACK) is sent back to the interface engine and the socket address and port refreshed, resulting in transmission and processing of the next transaction message in the queue.

Performance metrics were obtained using a custom perl script on the VM that simulated performance of the interface engine. Processing times were logged for 100 groups of 100 transaction messages, each transmitted immediately after each ACK was received. The mean processing time ± standard error were determined using the method of batch means for the N=100 groups. Accuracy was assessed by comparing the transmitted messages to the values populated in the SQL database.

Results
Processing time was 46.3 ± 0.44 msec per transaction. This permits a throughput of at least 1295 transactions per minute. Of the 10,000 messages transmitted during the test, 100% were inserted correctly into the database.

Conclusions
The developed system met our criteria of efficiency, accuracy, and use of open source software. For our use case, the highest frequency of transactions over a 2 month interval in any given minute was 11. Thus, our process has a capacity at least 100x greater than anticipated peak demand on a relatively low performance Windows 7 VM. Our process can be applied to any HL7 transactions that can be concatenated into a single, delimited string. Perl's strong support of text parsing and readily available libraries for construction of TCP/IP sockets and database connectors make this language an excellent choice for such use.

Process Improvements from Implementing an Electronic Checklist and Rounds Choreography to the Intensive Care Unit

Aysen Erdogan MD, Sumanjit Kaur MD, Lisbeth Y. Garcia Arguello MD, Hart L, John C. O'Horo MD, Ronaldo A. Sevilla Berrios MD, Adil Ahmed MBBS, Vitaly Herasevich MD, PhD, Brian Pickering MBBCH, Ognjen Gajic MD

Multidisciplinary Epidemiology and Translational Research in Intensive Care (METRIC), Mayo Clinic, Rochester, MN

INTRODUCTION: The volume of data generated in the intensive care unit (ICU) environment provides ample opportunity for errors, both from commission and omission. We have developed an electronic, context-sensitive rounding tool that algorithmically selects which relevant best practice goals to be addressed for each patient on a daily basis. This was integrated into a novel electronic platform, the Ambient Warning and Response Evaluation (AWARE) system. It prescribed a rounding choreography termed ProcessAWARE. In this preliminary study we sought to test the adoption and effectiveness of this intervention on the processes of care in the ICU.

METHODS: In a tertiary medical center medical ICU, 50 random patient-days were selected from April 2013 for appraisal of adherence to each of 16 best practice items included in the rounding tool that were reliably assessable. This was compared to 50 random patient-days from April 2011, the last year in which no checklist component of AWARE was available. Best practice adherence in the two time frames was compared.

RESULTS: Cases and controls were similar in respect of demographic characteristics. APACHE III scores at 1 hour were higher in 2013 (46 vs 64, p <0.01). Changes in processes outcomes presented in table.

Table. Process outcomes on April 2011 (Pre-AWARE) and April 2013 (Post AWARE implementation)

Variable	Pre-AWARE	AWARE with Checklist	P value
Sedation break	5/10 (50%)	19/21 (90%)	0.02*
Address Delirium	3/25 (12%)	2/25 (8%)	1.00
Addressed Pain	5/12 (42%)	2/7 (29%)	0.65
Review necessity of vasoactive medications	9/9(100%)	12/15 (80%)	0.27
Review cardiac medications	27/29 (93%)	23/24 (95%)	1.00
Appropriate use of stress ulcer prophylaxis	38/50 (76%)	40/50 (80%)	0.23
Review necessity of antibiotics	41/42 (98%)	44/47 (94%)	0.62
Address Infectious Source Control	5/20 (25%)	0/20 (0%)	0.05*
DVT prophylaxis appropriate	45/50 (90%)	48/50 (96%)	0.43
Central line necessity reviewed	11/22 (50%)	14/24 (58%)	0.77
Fluid balance goal established	29/50 (58%)	34/50 (68%)	0.41
Electrolyte Review	35/50 (70%)	38/50 (76%)	0.65
Urinary catheter necessity reviewed	16/29 (55%)	18/40 (45%)	0.47
Glucose reviewed	23/50 (46%)	19/50 (38%)	0.54
Physical Therapy goals reviewed	5/50 (10%)	12/50 (24%)	0.10
Goals of Care addressed?	37/50 (74%)	46/50 (92%)	0.03*

CONCLUSIONS: The initial implementation of a context-sensitive electronic rounding tool was associated with improvement in addressing goals of care and sedation practice. Clinician training and further refinement on the automated algorithmic function are needed to improve the adoption of the structured rounding tool.

De-duplicating Distributed Research Cohorts using Health Information Exchange Identity Services

Jon-David Ethington, MHS, PA-C[1,2], Shan He, PhD[1], Jerry M. Westberg[1,2], Darren K. Mann[1], Sidney N. Thornton, PhD[1,2],

[1]Intermountain Healthcare, Salt Lake City, UT; [2]University of Utah, Salt Lake City, UT

Abstract

Health Information Exchange (HIE) identification services can be repurposed to de-duplicate distributed research cohorts. An investigator can increase the statistical power of a research question by extending cohorts across organizational boundaries. The clever reuse of HIE identification services may reduce an organization's participation risk in inter-organizational studies by systematically preventing inadvertent disclosure of Personally Identifiable Information (PII), and avoiding the need for a pre-correlated, static master subject index.

Background

Distributed analytics may benefit from enhanced healthcare data interoperability. By using existing query tools, a research question can be applied to both internal and external cohorts, thus providing an investigator with increased statistical power[1]. De-duplication of the distributed cohort, however, is required. A constraint of de-duplication processes is that either PII must be disclosed for the purpose of resolving duplicates, or the researcher must subscribe to a federated service typically relying on static, pre-correlated repositories of identifiers. Since healthcare data interoperability also relies on patient identification across organizations, it is reasonable to adapt HIE identification services for research purposes. Such a strategy provides on-the-fly cohort identification without the need to maintain a costly, pre-correlated, static master subject index across organizations. Furthermore, HIE identification services such as developed by the Care Connectivity Consortium (CCC) provide dynamically-generated, non-PHI disclosing correlations back to the participating organization[2].

Approach

The existing patient identity management of the CCC forms the basis for the cohort de-duplication service. Demographic information for patients from distributed sources identified by the cohort query tools are sent to the CCC identification service's batch matching interface. The CCC matching engine links patient records that are considered belonging to the same entity through querying an existing identity correlation table built from previous operational transactions, its built-in probabilistic matching algorithm, and external authoritative sources such as commercial identity services or state department of transportation databases. For any unresolved identifiers, a level of certainty is calculated. For each study question, the service provides a list of de-duplicated, study-specific identifiers, and the linked organizations to each identifier without disclosing the organization-specific PII.

Results

This service provides to the investigator a de-identified, coded list of potential subjects, and the organization of origin for those subjects. For both resolved and unresolved identities, a level of certainty is supplied. The service also provides a throughput for data collection by communication with the utilized query tool without storing data.

Conclusion

It is possible to apply HIE identification services to research purposes and reduce organizational risk in participating in inter-organizational studies by systematically preventing inadvertent disclosure of PII. The service can be used with existing query tools without the need for additional master subject indices across member organizations.

References

1. Standards & interoperability (S&I) framework- query health. http://wiki.siframework.org/Query+Health (accessed March 12, 2014).
2. Thornton SN, Westberg JM, Gurr GE, Westberg LJ, Mann DK, Rasmussen DN. Sharing qualitative matching parameters among master patient indices. AMIA Symposium Proceedings 2013.

Virtual Learning Environment: a Proposal for Teaching Emergency Care for Nursing Students through WebQuest

Yolanda Dora M. Évora, RN, PhD[1], Marta Cristiane A. Pereira, RN, PhD[1], Andrea Bernardes, RN, PhD[1], Carmen Silvia Gabriel, RN, PhD[1]
[1]University of São Paulo College of Nursing, Ribeirão Preto, São Paulo, Brazil.

Abstract

The aim of this study was to develop a virtual learning environment (VLE) for teaching emergency care for nursing students, using tools of Information and Communication Technology. The design, based on Computer-Mediated Education learning, have the constructivist approach. The WebQuest methodology presents the following components: Introduction, Task, Process, Evaluation, Conclusion and Credits. It was available in Moodle platform, enabling safe use of the internet to find information.

Introduction

Due to the increased number of cardiovascular incidents, traffic accidents, violence and insufficient structuring of the health care network, the need to intensify educational activities in these scenarios arises such a strategy to qualify the patient care.
The aim of this study was to develop a virtual learning environment (VLE) for teaching emergency care for nursing students, using tools of Information and Communication Technology.

Method

This is a methodological development research involving the creation of a Virtual Learning Environment by using the WebQuest as a teaching methodology and pedagogical strategy. The design, based on Distance Education learning Computer-Mediated, have the constructivist approach. The WebQuest was developed in Power Point (Microsoft ®) following the structure: Introduction, Task, Process, Evaluation, Conclusion and Credits. It is available in Moodle platform, providing for the students, links that allow access to the internet pages or websites.

Results

The VLE involved with the development of the WebQuest presents the following components[(1)]: Introduction: short text that introduces the subject and anticipates which activities the students should conduct. Task: Describes what is expected of students in the end and what tools they should use for their training. Process: steps that students have to follow for developing the task. Conclusion: summarizes the issues explored in the WebQuest, the goals achieved and supposedly indicates how the student can continue to study the issue. Evaluation: Rubric evaluation for use in self-assessment, peer assessment and teacher assessment. Credits: Present the sources of all materials used by the teacher in the preparation and development of the WebQuest (hypertext, photos and educational videos).This WebQuest is simple, objective and easy to use, and its potential is extensive, allowing the nursing student to have a dynamic learning, access to large amounts of information, perception of information, construction and application of knowledge, and activity involving research, reading, collaboration and interaction.

Conclusion

The use of WebQuest strategy as a virtual learning environment applied to the teaching of emergency care is a viable and innovative possibility. The teacher becomes a facilitator and helps students to construct knowledge. This research presented a pedagogical proposal that seeks to integrate theory and practice, from the appreciation of the knowledge and context of real situations experienced in daily life, based on resources of the computer and information technology. Continuing the study seems relevant evaluating the effectiveness of virtual learning environment for the students of undergraduate courses in the health areas. This challenge demands more work and also the evaluation by experts in Emergency care.

Reference

1. Dodge B. What is a WebQuest? 2007. Available: http://webquest.org/index.php [10 oct. 2012].

iDECIDE: A Mobile Application for Pre-Meal Insulin Dosing Using an Evidence Based Equation to Account for Patient Preferences

Akram Farhadi, MS[1], Buffy Lloyd, B.S[1], Danielle Groat, B.S.[1], Jelena Mirkovic, Ph.D. [2], Curtiss B. Cook, MD[3] , Adela Grando, Ph.D.[1]

[1]Arizona State University Department of Biomedical Informatics, Arizona;

[2]Oslo University Hospital HF, Oslo, Norway; [3] Mayo Clinic, Division of Endocrinology, Arizona

Abstract

Type 1 diabetes (T1D) requires the patient to conduct frequent self-monitoring of blood glucose and dosing of insulin. Evidence has shown that patients are more compliant with their diabetes management when they incorporate personal preferences[1]. We have developed a mobile application prototype, iDECIDE, to further personalize pre-meal insulin dosing by incorporating current evidence related to two variables that influence prandial glucose level: alcohol and exercise.

Introduction

Insulin pumps are medical devices that deliver continuous insulin. Current insulin pumps do not incorporate the latest medical evidence to further personalize *pre-meal insulin dosing* based on individual's *preferences* for *exercise* and *alcohol* intake[2, 3]. Evidence shows that these personal preferences have a short-term impact on *glucose measurements*, which in turn affects insulin dosing. The proposed solution is to incorporate these parameters into the current equation that calculate and *recommend insulin dosing* to help achieve *target glucose*. In contrast, there are numerous mobile applications for diabetes management that allow users to track carbohydrate intake, exercise, medications and insulin dosage. iDECIDE differs from these current mobile applications in that it is evidence based, a criteria largely missing in current mobile applications**.**

Methods

First, we conducted a literature review to gather the latest guidelines and evidence on insulin dosing for T1D patients to expand the current insulin dosing equation to include exercise and alcohol intake. Second, we created three prototypical T1D patient case scenarios with different pre-meal preferences. The scenarios were based on the American Diabetes Association guidelines and the opinion of domain experts. Third, we use the Ontology Web Language (OWL) to model the domain knowledge (Fig. 1 depicts main OWL classes and relationships).

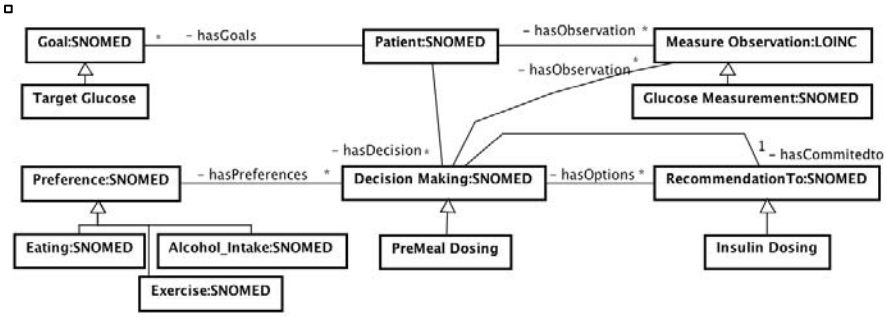


Figure 1: Knowledge Representation

Finally, we deployed a mobile application prototype (Fig. 2) that uses the new dosing equation to suggest insulin amounts based on patient's glucose reading and pre-meal preferences for carbs, alcohol and exercise.

Figure 2: iDECIDE Mobile Application

Future work

We have submitted an IRB for approval to conduct a preliminary retrospective calibration of iDECIDE evidence-based formula. We will collect data from 20 diabetes patients on alcohol and exercise preferences, and data generated from their insulin pumps. We will compare insulin recommendations from iDECIDE against insulin pumps, as recorded by study participants. Future work will also incorporate patient SMART goals (Specific, Measurable, Attainable, Realistic and Timely) related to fitness and nutrition to further help patients achieve a healthy lifestyle.

References

1. *A* shared *treatment decision-making approach between patients with conditions and their clinicians: The case of diabetes.* **Montori, Victor M, Gafni, Amiram and Charles, Cathy.** 1, 2006, Health Expectations, Vol. 9, pp. 25-36.

2. *Patient handout: Alcohol and diabetes.* **Phillips, Pat J., Carpetis, Melissa and Stanton, Connnie.** 6, 2010, MedicineToday, Vol. 11, pp. 73-75.

3. **Scheiner, Gary.** Exercise and pump therapy. [book auth.] Karen M. Bolderman. *Putting your patients on the pump.* Alexandria : American Diabetes Association, 2013.

Exploring the Use of SemRep Predications to Help Identify Secondary Drug Targets for Personalized Cancer Therapy

Safa Fathiamini, MD[1], Amber Johnson, PhD[2], Vijaykumar Holla, PhD[2], Ann Bailey PhD[2], Jia Zeng, PhD[2] , Lauren Brusco, PhD[2], Funda Meric-Bernstam, MD[2], Elmer V. Bernstam, MD[1], Trevor Cohen, MBChB, PhD[1]

[1]The University of Texas School of Biomedical Informatics at Houston, TX

[2]The University of Texas MD Anderson Cancer Center, Houston, TX

Abstract

We report some preliminary findings related to our efforts to utilize SemRep predications as a means to identify associations between molecular aberrations and targeted cancer therapies in the literature. Though SemRep by design emphasizes precision over recall, we found extracted predications to be an informative source of gene-gene relationships. Utilizing newer versions of the UMLS improved the ability of SemRep to recognize agents of interest.

Introduction

Genetic information used in personalized cancer therapy is complex, drug responses vary from patient to patient, and no updated comprehensive resource describing these variations exists. Therefore, researchers manually catalog recent discoveries published in journals to support clinical decisions. Notably, therapies can be targeted to the patient's specific molecular aberration, or to a secondary (downstream) target. Our goal was to automatically identify drugs that target specific genetic alterations through a *secondary target downstream of the altered gene*.

Methods

We used a gold standard developed by a panel of experts at the MD Anderson Cancer Center Sheikh Khalifa Bin Zayed Al Nahyan Institute for Personalized Cancer Therapy (IPCT). The gold standard consisted of one main gene (GeneA), two intermediate genes/proteins (GeneB) that interact with the main genes and 7 drugs (DrugA) that inhibit GeneB or GeneA. In search for relationships, we used this pattern: "DrugA inhibits/interact_with GeneB" (Pattern1). We used SemMedDB[1] to identify those Medline abstracts (28,839 abstracts, Set1) that contained any mention of GeneA or GeneB. SemMedDB is a predication database provided by SemRep team which by default uses MetaMap/UMLS 2006 as its backbone, and contains more than 70 million predications extracted from more than 20 million citations. To find Pattern1 relationships with the most recent versions of UMLS/MetaMap, we modified the UMLS 2013AB data files and used them to create an up to date version of MetaMap, and used them as the backbone for our locally run SemRep (SemRepLocal). We then ran SemRep2 on Set1 to create a database with 2,346,486 predications.

Results

With respect to identification of genes, we were able to identify matches for 100% of the gold standard GeneA and GeneB concept exemplars in SemMedDB. Using Pattern1 to identify DrugA, the precision and recall were 0.3% and 29% with SemMedDB, and 1% and 100% with SemRepLocal, respectively.

Conclusions

When newer versions of MetaMap/UMLS are used, SemRep is able to identify more cancer related drugs. The still low precision with SemRepLocal may be due to the fact that our gold standard only includes those drugs that are clinically available and relevant, while our system retrieves all concepts with pharmaceutically-related UMLS semantic types. Further review of the results in order to identify sub-categories of drugs and exclude the ones that are of no known clinical benefit may lead to the development of methods to improve precision, such as the application of more sophisticated filters. Though further evaluation is required before definitive judgment can be made about the practical value of such search methods, a clear finding of our current efforts is that the modified UMLS 2013 data files we have employed are required for recognition of many key cancer-related drugs. This is a prerequisite to the application of SemRep to the task of identifying pertinent indirect targets for personalized cancer therapy.

Acknowledgement

This research was supported by NIH grant U01 CA180964 (PIs Bernstam and Meric-Bernstam) and the MD Anderson Cancer Center Sheikh Khalifa Bin Zayed Al Nahyan Institute for Personalized Cancer Therapy. The authors would like to thank Halil Kilicoglu and Thomas Rindflesch for their assistance with modifications to SemRep.

References

1. Kilicoglu H, Shin D, Fiszman M, Rosemblat G, Rindflesch TC. SemMedDB: a PubMed-scale repository of biomedical semantic predications. Bioinformatics. 2012;28(23):3158–60.

Effect of Informatics Intervention on Compliance with Surgical Quality Metric

Vitali Fedosov, MD, PhD[1]; Ing C. Tiong[1]; Brian W. Pickering, MD[1]; Vitaly Herasevich, MD, PhD[1]
Multidisciplinary Epidemiology and Translational Research in Intensive Care (METRIC),
Mayo Clinic, Rochester, MN, USA

Abstract: Carefully built and appropriately implemented automated notification system of health care providers on the essential clinical parameters can improve the quality of care and impact reimbursement[1]. We have developed the automatic notification panel on the glucose level in critically ill patients presented in an intensive care unit (ICU) within 18-24 hours post an anesthesia time (SCIP-4 metric). We utilized the existing Ambient Warning And Response Evaluation (AWARE) middleware database tracking patients from the operating room and surgical ICU. The trial implementation of the developed tool was conducted by two nurses. The pilot study testing automatic notification has demonstrated significant (> 30%) improvement in compliance with the glucose control guidelines. With appropriated implementation the automated notification panel for a tighter glucose control in post cardiac surgery patients has demonstrated positive effect on quality of care.

Background: The tight glycemic control has been associated with reduction in an early mortality (death in ICU), incidence of complications, and length of stay in an intensive care unit (ICU). This impacts an overall quality of patient care and cost of hospitalization. Compliance with the glycemic control guidelines developed by the surgical care improvement project (SCIP) of Centers for Medicare and Medicaid Services (CMS) is suboptimal. **Aim:** To develop the automated electronic medical records (EMR) tool for improving compliance with the SCIP-4 metric.

Methods: The automated notification panel for ICU health care providers was developed as an application within the existing AWARE Clinical Desktop Application. The study took place at Mayo Clinic, Saint Marys Hospital, Rochester, MN. The hospital operating room (OR) DataMart database was employed as a primary data source to track the patient path in a hospital and extract the data on the glucose level and critical event history. The measurement of outcomes was a percent of compliance with SCIP guidelines over the 4 months period of time. The measurement of quality of care was the glucose level less than 180 mg/dL in two 6 hour periods within the 18-24 hours after the anesthesia end time (AET). Patients who were dismissed from the ICU within 24 hours of the anesthesia end time, deceased, left against medical advice, required cardiac pulmonary resuscitation, or required further surgeries were excluded from the study. The User Interface contained the set of information boxes: i) monitoring windows for the patient location; ii) insulin interventions; iii) laboratory results with trends with a spark line and optional full size visualization; and iv) timer for 12 to 18, and 18 to 24 hour periods after the AET. Notifications were designed to have specified colors according to the status hierarchy (‘warning’ - yellow, ’fail’ - red, ’pass’ - green, and ’excluded’ - grey), and AET date (Figure 1).

Results: Two nurses in a surgical ICU had access to the beta-version of the automated glucose status notification panel for 4 months. The glycemic control data were analyzed for 448 pre-implementation and 433 post-implementation patients. The number of noncompliance cases declined from 48 to 31. Compliance with the glycemic control was 10.92% and 7.57% pre- and post- panel incorporation respectively (p=0.06). The analysis of cases of noncompliance with guidelines identified the 3 main reasons for failures: i) spike increase in a glucose level immediately at the end of the 18-24 hour period; ii) failure to obtain glucose measurements; and iii) poorly controlled diabetes mellitus at the baseline. The analysis also demonstrated that compliance with the tight glucose control 12-18 hours after the AET was associated with a better glucose control 18-24 hours after the AET.

Conclusions: The automated glucose level notification tool facilitated the better glucose control, especially, in the most complex ICU patients.

. Lack of statistical significance in improvement of metric compliance is likely due to a small number of testing providers. Improvement of the tool implementation process is required.

References: S.Eslami,A.Abu-Hanna,N.de Keizer,R.Bosman,P.Spronk,E.de Jonge,M.Schultz, Implementing glucose control in intensive care: a multicenter trial using statistical process control, Intensive Care Medicine, 36 (2010), 1556-1565

GeneAnswers: Integrated Translational Interpretation of

Gang Feng, PhD[1], Pan Du, PhD[2], Xishu Wang, MS[3], Jing Wen, MS[4], Tian Xia, PhD[5], Warren A. Kibbe, PhD[6], Simon M. Lin, MD[7]
[1]Northwestern University, Chicago; [2]Genentech, San Francisco, CA; [3]International Trading Group, Chicago; [4]Greenline Financial Technologies, Chicago; [5]Huangzhong University of Science and Technology, Wuhan, China; [6]National Cancer Institute, Bethesda, MD; [7]Marshfield Clinic, Marshfield, WI

Description

It is not enough for most researchers to identify one or multi- groups of interesting genes or units from current clinical and translational studies. People expect more advanced concept analysis, such as functions and pathways, for groups of given genes. Our team has developed a new R-compatible Bioconductor package, GeneAnswers, to automatically present potential correlations between interested genes and specified concepts (functions, pathways, diseases, etc) based on statistical test. Besides Gene Ontology, KEGG pathways database, EBI Reactome pathway database and caBio pathway database that is more specific and up-to-date resource integrating NCI-Nature curated pathways, Biocarta and Reactome. Besides this, Disease Ontology, which is derived from NLM Unified Medical Language System (UMLS) and developed by our group and collaborators, is also supported by the current version GeneAnswers. Customized concepts from biological studies or clinical information can also allow users to perform own concepts enrichment tests in GeneAnswers and identify the potential relationship between given genes and the novel findings in biology and medicine. The package GeneAnswers can not only visualize the network between potential concepts and interested genes as well as gene interactions, but also uniquely combine optional data matrix (such as gene molecular profiles) and relative concepts together. With the support of Entrez eUtils, people can find the links connecting the given genes to customer-specified keywords so that possible biomarkers could be identified for validation. Moreover, homologous gene mapping function makes it possible for piloting experiments based on animal models to apply the achievements on human being. Furthermore, computer-generated multiconcepts-genes table is introduced to integrate concepts analysis for correlated groups of genes (such as time course study). For efficient visualization of our knowledge, a "wordle"-like strategy is also developed and integrated in GeneAnswers, which includes text-mining based know information preprocess. The package GeneAnswers, combining Gene Ontology, varies of pathway databases and NLM based Disease Ontology with gene interaction information as well as customized annotation support, is truly helpful for people to gain an insight into clinical diagnosis and remedy at molecular and genetic levels.

Using Electronic Health Record Access to Infer Physician Follow-up After Handoffs

Stephanie Feudjio Feupe Msc; Robert El-Kareh, MD, MS, MPH
University of California, San Diego

Abstract

Introduction

Physicians benefit from feedback of patient outcomes to help calibrate their diagnostic and therapeutic decision-making and learn from their errors[1]. This benefit may be especially pronounced for physicians in training, however we know that many resident physicians do not get adequate feedback of patient outcomes.[1,2] Fragmentation in care leads to lack of a longitudinal view of the effectiveness of initial management plans. Despite these barriers, an electronic health record (EHR) can enable more efficient follow up of patient outcomes. We sought to use EHR data to determine how often emergency medicine (EM) and internal medicine (IM) residents follow up on their hospitalized patients after handing their care off to other providers and to describe factors related to this follow-up.

Methods

We obtained IRB approval for our project and obtained consent from 22 EM and 40 IM residents to study their chart access histories. We identified patients who were hospitalized at UC San Diego Heath System between December 16, 2012 through April 30, 2013 and cared for by at least one of the consenting physicians. We identified handoffs using the treatment team assignment data in our EHR. We linked these data with resident rotation assignments. We measured the earliest time that the physician handing off the patient's care to another provider subsequently re-accessed that patient's electronic chart. We generated descriptive statistics to evaluate the post-handoff chart access relative to physician, patient and encounter characteristics. We performed a preliminary logistic regression analysis of IM resident re-access controlling for patient age and gender, level of care that the patient required , resident year of training, rotation assignment, and whether handoffs occurred prior to discharge, on a weekend and during off hours.

Results

We analyzed records for 6075 handoffs involving 2406 patients during the study period. Overall, residents re-accessed patient charts within 14 days following handoffs in 111/3581 (3%) cases for EM residents and 1394/2494 (56%) cases for IM residents. For IM cases, we found that senior residents were significantly less likely to re-access charts compared with junior residents and interns with an odd ratio (OR 0.5795% CI 0.47-0.67, $p<0.001$). Compared with daytime residents on ward rotations, re-access of charts occurred significantly less frequently for EM residents (OR 0.03, 95% CI 0.02-0.05, $p<0.001$) and IM residents on night admitting rotations (OR 0.23, 95% CI 0.17-0.32, $p<0.001$). In addition, re-access was significantly more likely if the handoff occurred prior to the patient's hospital discharge (OR 2.99, 95% CI 2.45-3.64, $p<0.001$).

Discussion

We found that EM residents and IM residents who admitted patients overnight were least likely to re-access patient charts within 14 days following their handoffs of patient care. For many patients, the initial diagnostic and therapeutic decision-making is performed by residents on these rotations. Therefore, these residents may benefit the most from feedback of the results of their decisions. These results suggest the need to create a more reliable system for providing patient outcome feedback to physicians in training.

References

[1] Schiff GD. Minimizing diagnostic error: the importance of follow-up and feedback. Am J Med. 2008;121(5 Suppl):S38-42.

[2] Lavoie CF, Plint AC, Clifford TJ, Gaboury I. "I never hear what happens, even if they die": a survey of emergency physicians about outcome feedback. Cjem. 2009;11(6):523-8.

Understanding the Dispensary Workflow at the Birmingham Free Clinic: Responding to Challenges with Informatics Interventions

Arielle Fisher[1], Gerald Douglas[1], PhD
[1]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA

Abstract
Objective: *Provide formal documentation in the form of work models of the routine pharmaceutical workflow at the Birmingham Free Clinic (BFC).* ***Methods:*** *We used a contextual inquiry methodology to document the current workflow and facilitate the identification of critical aspects of intervention design specific to the user.* ***Results:*** *A total of three pharmacists were observed and interviewed at the BFC. Work processes that may benefit from the introduction of informatics interventions were identified. Maintaining medication stock levels, medication dispensing, patient counseling, and effective use of the EMR for patient prescription information were identified as candidate areas for informatics interventions.*

Introduction
The Birmingham Free Clinic in Pittsburgh, PA is an organization designed to serve uninsured and other medically vulnerable groups through the use of a volunteer group of health care providers. This clinic offers a variety of services including primary care, chronic disease management, physical exams, and medication access and management through an in-house dispensary. The introduction of an electronic medical record (EMR) has improved several aspects of workflow in the free clinic, such as the ability to better document patient visits and archive prescriptions. However, many of the pharmacists' everyday tasks regarding the dispensary and medication management have become more challenging since the EMR was introduced. While the implementation of the EMR is intended to increase patient safety and quality of care, benefits cannot be fully realized due to the inability of the EMR to support workflows between the clinical service and the dispensary. We hypothesize that inefficiencies that exist in the current dispensary workflow at the BFC may potentially be alleviated by a series of informatics interventions. In order to identify workflow inefficiencies and propose interventions, a detailed understanding of the free clinic workflow and the needs of the key user, the pharmacist, is necessary.

Methods
Contextual inquiry is a qualitative, user-centered, social method designed to identify and understand users' needs and is used for collecting, interpreting, and aggregating in-detail aspects of work [1]. Pharmacists were observed and interviewed at the BFC according to contextual inquiry guidelines. Notes describing the overall workflow in the BFC, common breakdowns and inefficiencies, photos of physical artifacts, and detailed information regarding specific dispensary processes were collected during observation sessions. Five graphical models were produced to aid data visualization including sequence, flow, artifact, physical and cultural models.

Results
A total of three pharmacists were observed and interviewed at the BFC. Notes from a total of three observation sessions lasting approximately 3-4 hours each were documented and analyzed. The current dispensary workflow at the BFC is labor intensive and lacks efficiency when integrating with the clinical service. Observations identified processes that may benefit from the introduction of informatics interventions including: maintaining medication stock levels, medication dispensing, patient counseling, and effective use of the EMR for patient prescription information.

Future Work
We will identify a rank order of workflow inefficiencies as perceived by three pharmacists who have experience working in this setting and define informatics interventions designed to address the highest-ranking inefficiencies.

References

1. Holtzblatt K, Beyer H. Contextual design: defining customer-centered systems. San Diego: Academic Press; 1998.

Title: Using HIE Data to Calculate Quality Measures for Public Health Surveillance

Authors: Elaine Fontaine, BS, Rhode Island Quality Institute, Providence RI 02908
Jonathan Leviss, MD, Thundermist Health Center, West Warwick, RI

Problem/Purpose: Disparate data systems across stakeholders in the healthcare system result in siloed views of quality metrics, focusing on either insured subpopulations or Meaningful Use measures reported by the provider community. Payers rely on claims for case identification and measure achievement for insured patients. Providers are dependent on their electronic medical record to identify both denominators and numerators. Public health agencies have access to limited data from payers and providers, making it difficult to evaluate overall compliance with agreed upon standards of care across a broad population. The goal of this project is to explore the differences in rates calculated using standard claims and EMR data and to begin to assess the suitability of health information exchange data as an alternative approach to assess statewide performance on quality metrics.

Methods: Data collected from Current Care, Rhode Island's Health Information Exchange, is being analyzed to calculate a statewide Chlamydia Screening Rate for women 16-24. Upon enrollment in Current Care, patients are uniquely identified using QuadraMed's master patient index software; thereafter clinical data automatically flows into the HIE from 90% of Rhode Island labs, all Rhode Island adult acute care hospitals, 90% of Rhode Island pharmacies and over 40 doctor's offices . At the time that this preliminary analysis was conducted, there were approximately 270,000 patients enrolled in Current Care. Using data from patient enrollments and continuity of care documents from primary care offices, a denominator was generated based on the patient's age and gender, regardless of payer or primary care physician relationship. A numerator was generated based on evidence of laboratory testing for Chlamydia. These rates are compared to health plan reported rates and rates reported by practices participating in Rhode Island's Patient Centered Medical Home project, the Chronic Care Sustainability Initiative (CSI), which relies on electronic medical record data for denominator and numerator identification.

Results:

	Current Care	Rhode Island Commercial and Medicaid Plans [(1)]	CSI Practices Reporting with EMR Data
Measurement Period	CY2013	CY 2006	CY2013
Denominator	11,303	12,729	1,600
Numerator	3,663	---	957
Rate	34.2%	44.5%	59.8%

Conclusion: These results suggest that the overall screening rate for patients in the Rhode Island Current Care population may be lower than reported by other sources, including health plans. Key items to consider as we advance this analysis include selection bias of Current Care enrollees, differential screening rates for uninsured patients, and/or completeness of HIE data.

1. Centers for Disease Control. (2010).*Chlamydia Screening Percentages Reported by Commercial and Medicaid Plans by State and Year* [Data file]. Retrieved from http://www.cdc.gov/std/chlamydia/female-enrollees-00-08.htm

Do Consensus Abstracts Agree with Meta-Analyses or Systematic Reviews?

Paul Fontelo, MD, MPH, Raymond F. Sarmiento, MD, Raymonde C. Uy, MD, MBA, Fang Liu, MS
National Library of Medicine, Bethesda, MD

Abstract

We reviewed whether conclusions from systematic reviews and meta-analyses agreed with those derived from Consensus Abstracts, a tool for finding relevant, validating abstracts and bottom-line (TBL) summaries. A crowd-sourced evaluation found that only 66% were concordant. However, a review of discordant results showed that only 2/33 reviews were considered marginally discordant, subject to interpretation. The most common reasons of disagreement were: unresolved medical controversies, incorrect search methods, inappropriate selection of abstracts, and erroneous conclusions.

Introduction

'Consensus Abstracts**'** (CA) is a Web and mobile application that finds related, validating abstracts that may be useful for clinical decision making.[1] For clinicians in low resource settings interested in evidence-based medicine, CAs may be more useful. Using simulated cases, a recent study among resident physicians in a developing country showed that the accuracy of clinical decisions improved equally using abstracts alone or full text articles.[2] Moreover, even those who stated that accumulated knowledge or "stock knowledge" alone were adequate, sought evidence. Whether or not consensus abstracts reach the same conclusions as systematic reviews or meta-analyses was raised recently.[3] Although the CA's algorithm is designed to include systematic studies (Cochrane reviews, systematic reviews, and meta-analysis), we wanted to find out whether Consensus Abstracts based solely on original publications (i.e., systematic studies excluded) reach the same conclusions as more systematic studies. If CAs will be a key source of evidence, it is essential that conclusions derived agree with synthesized evidence sources.

Methods

Through a Web interface, reviewers selected a clinical topic then chose one or both CA search tools, i.e., PICO or *ask*MEDLINE, to search and retrieve citations from PubMed. Filters were added to remove citations without abstracts, Cochrane reviews, systemic reviews, and meta-analysis to simulate instances where no systematic studies were found and the clinician relying solely on abstracts and bottom line summaries. Reviewers selected as many citations as considered essential then compared the conclusion derived from them with those from preselected systemic reviews or meta-analysis articles from high-impact journals (Lancet, NEJM, JAMA, etc.). All selected citations, and conclusions were stored in MySQL tables for review by the authors. All discordant topics were reviewed and the reasons for discrepancies were determined. All the clinical topics were also searched with the regular CA search tool to establish if a PubMed/MEDLINE search would retrieve systematic level studies.

Methods and Discussion

Although there were only 18 topics searched, some were reviewed more than once totaling 33 reviews. Thirteen of 33 (39%) reviews were discordant. An analysis showed that the most common reasons for discordance were: incorrect formulation of clinical questions, incorrect search terms, inappropriate selection of abstracts, erroneous conclusions and unresolved medical controversies. Each of the 18 topics retrieved between 1-12 systematic studies.

Conclusion and Lessons Learned

Poor search and appraisal skills accounted for most disagreements. Search retrievals can vary widely depending on terms used and the formulation of clinical questions. In PICO, searches that are too specific limit retrievals and may even retrieve none. Clinicians need to be more familiar with "Related Items" in PubMed because it is useful for finding relevant articles. It is evident from this study that searching medical literature is an art and skill that need to be taught and developed among healthcare providers so they can practice evidence-based medicine better.

References

1. **Consensus** abstracts for evidence-based medicine. Evid Based Med 2011;16:36–8.
2. **A comparison** of the accuracy of clinical decisions based on full-text articles and on journal abstracts alone: a study among residents in a tertiary care hospital. Evid Based Med 2013;18:48–53.
3. **EBM** apps that help you search for answers to your clinical questions. Evid Based Med; 2014 Feb 14.

Title: Using Mental Models To Teach Public Health Informatics Evaluation In An Applied Training Fellowship

Authors/Degree: Laura H. Franzke, PhD, MPH; Herman Tolentino, MD; and Sridhar Papagari Sangareddy, MS
Affiliation: Centers for Disease Control and Prevention, Public Health Informatics Fellowship Program**,** Atlanta, Georgia

Abstract (50-75 words): Mental models allows for a shared language and conceptual representation of evaluation concepts when teaching informatics trainees. Mental models representing form and function of a public health information system (IS) assists in visualizing concepts and learning. Specifically, dissection of an IS into form and function facilitates problem formulation and application of methods, tools, and techniques to the evaluation problem. Use of two mental models in an applied fellowship demonstrates their utility for teaching IS evaluations.

Description: Public Health Informatics trainees come from very diverse backgrounds. Mental models allow use of a common language and shared conceptual representations of evaluation problems. Living (e.g., human body) and non-living (e.g., computers) systems can be understood in terms of form (how its components are assembled together) and function (how the components work together to achieve specific tasks). Evaluations of information systems including those used in public health (PH) can be framed according to an understanding of their form and function. Two mental models to aid evaluation of IS in PH are based on the Onion Ring Model (ORM) and the Information Value Cycle (IVC). Dissection of an IS into form and function facilitates problem formulation and application of methods, tools, and techniques to the evaluation problem. Use of these frameworks in an applied fellowship demonstrates their utility for teaching IS evaluations.

References

1. Heeks R, Bathnagar S. Understanding success and failure in information age reform. In: Heeks R, ed. Reinventing Government in the Information Age: International Practice in IT-Enabled Public Sector Reform. Abingdon, Oxon: Routledge; 1999: 49–74.
2. Taylor RS. Value-added processes in the information life cycle. Journal of the American Society for Information Science 1982;33:341–6.

Building the Clinical Personalized Pragmatic Predictions of Outcomes (Clinical3PO) Pipeline Within U.S. Department of Veterans Affairs (VA)

Lewis J. Frey, PhD[1], Leslie Lenert, MD[3], Augie Turano, PhD[4], Wim Cardoen, PhD[5], Sean Igo, MS[2], Richard Bradshaw, MS[2], Scott Duvall, PhD[6]
[1]Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC; [2]Biomedical Informatics, University of Utah, Salt Lake City, UT; [3]Department of Medicine, Medical University of South Carolina, Charleston, SC; [4]Department of Veterans Affairs, Pittsburgh, PA; [5]Center for High Performance Computing, University of Utah, Salt Lake City, UT; [6]Internal Medicine, University of Utah, Salt Lake City, UT

Problem

The U.S. Department of Veterans Affairs (VA) Informatics and Computing Infrastructure (VINCI) has 20 million patient records within its electronic health records system. The vital sign measurements alone consist of over a billion measurements. While the growth of data offers significant benefits to research, it does add an additional load to the VINCI computing capabilities. The adoption of a big data system enables the use of commodity hardware to perform analysis with greater sample sizes and better management services while reducing wait times for the production and analysis of datasets.

Introduction

This poster presents the design and deployment of the Clinical3PO environment within the VA. Leveraging their experience working within the VA technology infrastructure, the authors provide vital information on connecting to the vast amount of data collected on VA patients to improve health outcomes and advance data mining. Clinical3PO pipeline is an open-source big data system that enables large-scale analysis of medical data, particularly the VA data.

While maintaining veteran's privacy and security, data from all the Veterans Hospitals are organized and integrated within a central data warehouse and made available for research through VINCI. This includes patient care encounters/visits; Unified Medical Language System tagging codes derived from Natural Language Processing (NLP) for both clinical and administrative data elements; vitals; laboratory tests; clinical notes; procedures; medications; and a host of other medical measurements collected from patients enrolled at veteran hospitals. VINCI has been growing at a substantial pace providing researchers and clinicians access to vast amounts of medical data in a convenient and secure environment. VINCI has 20 million patient records within its EHR system. The preponderance of records are encoded from 2002 to present and includes approximately 12 million patients with large amounts of laboratory data. The vital sign measurements alone consist of over a billion measurements. Currently VINCI supports 2696 users and over 1278 projects.

Given the existence of an integrated system that has compiled, curated and tagged health data on millions of patients across United States, the opportunity exists to deploy novel and extant data mining approaches to improve predictive capabilities and outcomes within the Veterans Affairs (VA) system.

Methods

To minimize the burden on limited VA research resources, new code development takes place on an external test bed that simulated the VA environment. The external test bed is encoded on the Amazon Web Services (AWS) along with a non-VA publicly available intensive care unit dataset mapped to the Observational Medical Outcomes Partnership (OMOP) data model. After initial testing, code is migrated to the VA system through a remote login connection. The current VA Hadoop platform is Horton DataWorks 2.0 and consists of 5 Dell machines containing 2 processors with 4 cores each and 144 Gigabyte of RAM, running the Red Hat 6.0 operating system. There is one name node and 4 data nodes in the current environment with a total of 40 Terabytes of storage and the systems are connected using 1 Gigabit Ethernet.

Conclusions

Using external test bed approach enables rapid code development without directly interfacing with the VA hardware environment, thus, providing a flexible environment to explore software configurations and algorithms without using limited VA research resources.

A Layered context model: a basis for customized treatment – a GDM patient case study

Adi Fux[1], MA; Mor Peleg[1], PhD; Pnina Soffer[1], PhD; Mercedes Rigla[2], MD

[1]Department of Information systems, University of Haifa, Israel

[2]Endocrinology and Nutrition Dept., CSPT, Hospital de Sabadell, Sabadell, Spain

Abstract. Clinical guidelines (CPGs) provide evidence-based treatment options based on patient's clinical context. A guideline-based Clinical Decision Support System (DSS) matches a patient's clinical data with the clinical knowledge represented as a computer-interpretable guideline (CIG) to provide patient-specific recommendations. Physicians try to provide personalized treatment recommendations by different methods such as adjusting the recommendations to suit the patient's personal situations, to abide with his personal constraints, such as availability level of family support, and his personal preferences regarding side effects of medications. To support such personalization, CIGs would need to be customized to address not only the clinical context but to include also personal context and preferences. Yet it is important to stress that the (evidence-based) clinical context should remain the primary contextual aspect that determines the treatment options.

Our aim is to develop a decision model that will help physicians define the treatment recommendations based on the specific clinical context and personalize the recommendations, using other contextual aspects.

Background. DSS crosses the patient's data from the medical record with the CIG's medical knowledge to provide patient-specific recommendations. The medical staff manually adjusts the treatment recommendations based on the preferences, personal context or additional considerations, such as comorbidities. Our goal is to support physicians in this adjustment task by defining a layered context model that will refer to different context layers hierarchically and provide personalized treatment recommendations.

Research Question. How can we develop the layered context model to present personalized recommendations, based on clinical and personal patient's context?

Methods. We have been developing a layered context model using PRO*forma* CIG. Each layer uses the same candidates and decisions. Layer 1 includes primary arguments e.g. BG trend, BG level 1 hour after meal etc., to define the primary layer recommendations. Layer 2 includes secondary arguments e.g. routine level, support level etc., to define the secondary layer recommendations. The net results of recommendations from both layers define the specific recommendation. To demonstrate the approach we have used an example of Gestational Diabetes (GDM) and personal context to personalize the patient's treatment recommendations.

Results. The results of the personalized treatment recommendations are presented in Table 1. The first layer provides the clinical-based treatment recommendations. The personal context layer changes the number of measurements per day, exercise level, and consequently the expected blood glucose level (highlighted in the table) to provide specific recommendations for patient routine and support changes.

Table 1.layered context model

Personal Context	Affected objects	
Routine diet or schedule AND NOT Good metabolic control	Fasting BG	70-100 ml/dl
	Expected BG Threshold	70-120 ml/dl
	Measurements	3-4 times per day, every day
	Physical activity	mild to moderate exercise (1-6 MET), 0.5-1 hour each time, 3-4 times a week
Semi routine diet or schedule AND NOT Good metabolic control	Fasting BG	70-100 ml/dl
	Expected BG Threshold	70 - 140 ml/dl
	Measurements	2-3 times per day, every day
	Physical activity	mild exercise (1-3 MET), 0.5-1 hour each time, 3-4 times a week

Layer 1:

decision :: 'BG_measurement_per_day' ;
  candidate :: 'three_times_per_day' ;
    argument :: confirming,**BG_trend_state** = improve_and_below_threshold attributes
  recommendation ::netsupport(BG_measurement_per_day, three_times_per_day) >= 1;
  candidate :: 'four_times_per_day' ;
    argument :: against,**BG_trend_state** = stable_below_threshold attributes
  recommendation ::netsupport(BG_measurement_per_day, four_times_per_day) >= 1;
end decision.

Layer 2:

decision :: 'BG_measurement_per_day' ;
  candidate :: 'three_times_per_day' ;
    argument :: excluding,patient_routine_level = routine AND petient_support_level = frequent attributes
  recommendation ::netsupport(BG_measurement_per_day, three_times_per_day) >= 1;
  candidate :: 'four_times_per_day' ;
    argument :: confirming,patient_routine_level = routine AND petient_support_level = frequent attributes
  recommendation ::netsupport(BG_measurement_per_day, four_times_per_day) >= 1;
end decision.

Discussion. Gómez et al. [1] use a first layer to define the specific options that will be used in the second layer, without affecting them. In our model the first layer defines the treatment recommendations and a set of valid goals and options. In the second layer we will reason about the effect of the secondary considerations to provide specific treatment recommendations. The effect of the second layer may change the goals or the options and provide other valid options, more suitable to the secondary considerations.

Acknowledgement. This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 287811.

References

Gómez, E. J., Cáceres, C., López, D., & Del Pozo, F. (2002). A web-based self-monitoring system for people living with HIV/AIDS. *Computer Methods and Programs in Biomedicine*, *69*(1), 75–86.

Configuring Health Information Exchange Identity and Consent Services for Operational Use in a Changing HIE Landscape

Jason Gagner, MBA, Darren Mann,
Shan He, PhD, Neil Web, Greg Gurr, Sidney N. Thornton, PhD,
Intermountain Healthcare, Salt Lake City, UT

Abstract

The Health Information Exchange (HIE) landscape continues to evolve. The proximal roadmap for interoperability includes compatible network to network connectivity including vendor-operated exchange networks. Data services for HIE identity and consent management originally designed for central network connectivity have been refactored to allow multiple component architecture configuration options to support the diverse network requirements and organization-specific operational constraints. Logical processing has been added to the services to capture and track the transactional value added by the HIE services.

Introduction

During early 2011 Intermountain Healthcare entered into an HIE collaborative effort with the Care Connectivity Consortium focused on using open source tools that would allow for the beginning developments of a patient identification servicers that could be utilized across the country and consumed by any provider regardless of their EMR configuration. One of the goals for the collaborative effort was to simplify the process and reduce the overhead needed to manage patient matching and consent from external care providers. With the recent emergence of the exchange partnerships such as eHealth Exchange, competitive EHR vendor exchange networks, and Carequality, the immediate HIE reality will require interconnected networks rather than a single dominant network. Conservative organizational data security constraints have also limited the widespread adoption of open services for identity management outside the entity's firewall. Additionally, participating exchange organizations require return-on-investment transparency to justify expanded participation.

Methods

The components supporting identity and consent services have been reconfigured to work in three primary workflows so as to broaden the ability of adoption by provider EMR's regardless of geographical location or vendor: 1) Distributed services behind organizational firewalls; 2) Sponsored services behind a sponsor's firewall; and 3) In-parallel services to augment EHR-vendor HIE services. Logical processing and transactional reports have been configured to monitor the impact of the HIE services beyond the capability and constraints of the host network.

Results

Through the restructuring of the component architecture, the implementation and consumption of the HIE services for identity and consent management have increased from one organization using central network services in 2013 to early 2014 implementations of 2 organizations with distributed services, 3 additional organizations implementing the sponsored configuration;1 organization using the in-parallel configuration; and more under active consideration. Moreover, configuration challenges with onboarding to the e-Health Exchange Network have been resolved by shifting from central services to distributed configurations. Data collected from the return-on-investment logic are being included in multiple studies determining trait sensitivity and identification efficiency.

Conclusions

To increase the number of participating networks and organizations, HIE services for patient identification and consent have been adapted for flexible, distributed configurations. Logical processing can support transactional return-on-investment reports.

References

Care Connectivity Consortium: http://www.careconnectivity.org/
HealtheWay: http://healthewayinc.org/
Thornton S, Westberg J, Gurr G, Westberg L, Mann D, Rasmussen D. Sharing qualitative matching parameters among Master Patient Indices. AMIA Symposium 2013

Development of the Medication Image Library (MIL)

Robert Gale, RpH, MBA

Introduction

A Medication Image Library (MIL) has been developed by the Veterans Affairs (VA) Consolidated Mail Outpatient Pharmacy (CMOP) to establish and maintain a library containing pharmacist verified, standardized, high quality images of all medications and products dispensed by the VA to our Veterans. Included in the library are prescription medications, over the counter (OTC) medications, medical devices, medical supplies, nutritional agents and all medications and products dispensed by the VA to our Veterans. MIL launched in June 2011 with the creation of the first standardized images using existing VA medication and product pictures. Currently the library contains over 46,000 images representing over 17,000 VA medications and products. MIL has photographed over 10,000 VA medications and products for the library.

Background

The Medication Image Library (MIL) was started to support the VetLink self-service Kiosks that were developed in response to the VA T21 Transformation Initiative, "Enhance the Veterans Experience and Access to Health Care". MIL will provide medication and product images for the VetLink Medication and Allergy Review Module. MIL's database and images can also be used to assist with Veteran education, Veteran medication reconciliation and the processing of Veteran prescriptions.

Objective

1) Improve Veteran learning through visual medical references. 2) Aid to Veterans and their health care team in medication reconciliation to help reduce adverse drug events 3) Standardization of medication descriptor and imaging procedures, tool to standardize imaging and descriptor naming across the VA. 4) Use of high quality images to improve the accuracy of processing and verifying Veteran prescriptions.

Process

85% of the medications and products that the Veterans receive are filled through the seven CMOPs. MIL is located within the Great Lakes (GL) CMOP, Hines, IL, allowing it to have access to a large variety of VA medications and products. VA medications and products not stocked at GL-CMOP are obtained from the other six CMOP's and also from visiting Hines VAMC, Jesse Brown VAMC and the Lovell Federal Health Care Center, which are all located in the Chicago, IL area. This close relationship with the CMOP's and VA outpatient pharmacies, allows MIL to stay up-to-date when new VA medications and products enter the VA system or when style changes occur with existing medications and products. Currently MIL contains 99.5% of the medication and products used by the CMOP's and 95% of the medications and products used by the VAMC to fill Veteran prescriptions.

System for 'Intent-to-Treat' Analysis in a Real-world Setting

Purav B. Gandhi, MBBS, MBA[1], Amarinder S. Sidhu, BE, MBA[1]
[1]ConvergeHEALTH by Deloitte, Newton, MA

Abstract

Real-world datasets have multiple challenges such as inconsistencies in quality and lack of quality. An 'Intent-to-Treat' Analysis was performed on such dataset sourced from electronic medical records and other secondary sources, to compare superiority of one medication over another in a retrospective fashion, for comparative effectiveness research.

Introduction

With growing amount of data due to EMR adoption, better technology to gather and harness data, and increased computational power to analyze the same, researchers have increased access to real-world evidence for comparative effectiveness purposes.

There are a number of challenges with regards to analyzing the data and generating reliable insights due to inconsistencies in the quality, lack of continuity, etc. There are multiple challenges associated with aggregating, normalizing and loading this data into a platform for analysis.

Method

We, at ConvergeHEALTH, have developed an active data surveillance system to monitor clinical effectiveness and drug safety related information for treatment medications available in the marketplace using secondary information captured from electronic medical records and other sources of real-world evidence. We leveraged this platform to double de-identify data from a healthcare system, and load it into a normalized structured data warehouse. Next, we configured a visualization application to load two select patient cohorts on different medications.

Boundary condition for the cohorts were defined based on the 1) diagnosis codes, 2) prescription of drugs to be compared, 3) index date based on the first prescription, 4) washout periods to select a treatment naïve patient cohort, 5) observation and reference period for measurement of the outcomes, and 6) propensity matching algorithm to balance cohorts. A certain set of pre-defined outcomes relevant to the disease area were tracked to measure superiority of one medication over another.

Results

'Intent-to-Treat Analysis (ITT)' evaluating superiority of one medication over another was performed. This was achieved by tracking multiple health outcomes relevant to the disease area over the period of entire observation period in the visualization application.

Discussion

Experiments of this nature bring into question various aspects of an ITT analysis, and its applicability for retrospective research on a real-world dataset given they are not as controlled and regulated; with the expected dropout percentage being much higher than clinical trials. However, eventually the goal of comparative effectiveness research would be to have an analysis that can account for non-compliance due to multiple reasons while evaluating the safety and effectiveness of a medication.

Conclusion

'Intent-to-Treat Analysis' provides a view in what is the expected performance of a medication in a real-world setting after accounting for all the patient factors that drive non-compliance.

Future Plans

We intend to compare the results of this analysis with various published observational research studies to identify the key similarities and differences.

Development of a Community Care Information System: A Case Study in Singapore

Alex Gavino, MD[1], Na Liu, PhD[1], Deepa Rengarajan[1], Kharel John Rebada[1], Liming Bai, MS[1], Eugene Shum, MBBS, MPH[2], Alfred Wu, MS[1]
[1]SMU-TCS iCity Lab, Singapore Management University, Singapore; [2]Eastern Health Alliance, Singapore

Abstract

We present a preliminary case study for the development of a community care information system (CCIS) for a regional health system in Singapore. The CCIS aims to increase the productivity of the medical-social care team during their home visits to clients who are enrolled in their community care program. Using the action design research approach, the iCity Lab proposed a design method for the CCIS and developed an information system that addresses the various processes of the program and the dynamic user requirements.

Introduction

SMU-TCS iCity Lab is a research collaboration between Singapore Management University (SMU) and the Tata Consulting Services (TCS). Our aim is to study future trends and directions in urban development and develop appropriate information and communications technology (ICT) solutions that bring better quality of life to people. We developed the City Process Management (CPM) framework[1] to guide us in creating innovative solutions for future cities. We focused on healthcare and ageing to demonstrate and prove the CPM framework in various sectors of the city. In Singapore, we are facing new challenges due to an ageing population, growing incidence of chronic diseases such as diabetes, and decreasing size of households. This trend changes the needs of a city resident and hence affects services which the city should offer to its residents.

We partnered with a regional health system in Singapore to extend our assistance to their community care program for clients with high care needs, vulnerable elderly persons and at-risk residents. In this program, the community care team (CCT) conducts an initial home visit to assess the social and medical needs of the resident. The CCT then creates a personalized intervention plan based on the initial assessment. Follow-up home visits and calls are made to monitor the status of the client. All encounters with the clients are documented using paper-based forms and notebooks. Our lab designed a community care information system (CCIS) to help the care team in their record-keeping and retrieval of client data.

Methodology

We used an action design research method which leads to the building of innovative information technology artifacts in an organizational context and learning from the intervention while addressing a problematic situation. Consecutive meetings with the CCT allowed us to understand their workflow and needs. After the initial user requirements gathering, our software development team created an initial version of the CCIS within 2 weeks and demonstrated the solution to the CCT. Comments and recommendations by the CCT were noted and incorporated in the next iteration. This process was repeated over a total of 3 months in order to better address the requirements of the end users.

Results and Discussion

Based on the user requirements, the following capabilities were needed in the CCIS: collect patient information and assessment data through digital forms, create client goals and intervention plan, remind the CCT of tasks for the clients, track the client encounters through an occurrence log, and provide a dashboard for reports. We are currently alpha-testing the CCIS. Further enhancements will be made once the CCT provides us with feedback based on their use of the system in the field.

References

1. Teo CS, Wu A, Bhandarkar H, Chaudhuri RC, Venkatachari SR. Intelligent Cities: A City Process Management Approach. SMU-TCS iCity Lab White Paper. 2013.
2. Sein MK, Henfridsson O, Purao S, Rossi M, Lindgren R. Action Design Research. Management Information Systems Quarterly. 2011, 35(1), pp.37-56.

Machine Learning Made Easy with Sherlock

Thomas Ginter[1,2], Olga V. Patterson, PhD[1,2], Ryan C. Cornia[1,2], Scott L. DuVall, PhD[1,2]

[1]VA Salt Lake City Health Care System; [2]University of Utah, Salt Lake City, UT

Introduction

Natural language processing (NLP) tasks commonly employ various machine learning (ML) algorithms to create statistical language models for document classification, information extraction, and an array of other purposes. A wide variety of ML algorithms and accompanying frameworks have been developed. Each framework comes with its own feature vector representation and associated objects to support the unique workflow and features that are provided. This diversity of implementations introduces a barrier for NLP system developers when attempting to incorporate a ML step into an NLP system. Each implementation, being unique to the algorithm it supports, is rarely reusable across NLP projects. Additionally, implementations of a feature vector that will work for one ML framework will need to be recoded to work for another, even if the feature set does not change.

This poster introduces the Sherlock ML framework that provides a common infrastructure optimized for text processing with ML algorithms regardless of the ML framework used..

System description

Sherlock is an extensible ML framework based on and compatible with the Apache Unstructured Information Management Architecture (UIMA).[1] Sherlock simplifies ML algorithm integration into UIMA-based NLP pipelines and takes full advantage of the scalability and flexibility of UIMA AS by providing modular and reusable building blocks for a ML workflow. The modules utilized in both training and prediction workflows include feature vector generation, training, and prediction.

Standardizing a feature vector representation is the most important factor in seeking to create a reproducible and translatable system across ML frameworks. Sherlock provides a standardized feature vector representation that can be stored for later use and shared across multiple ML frameworks in different pipelines or in the same pipeline. While this representation has been standardized, Sherlock does not impose any limitations on the scale or methods of calculating feature vectors on the given inputs. With Sherlock, developers only need to engineer a desired feature vector once to use any of the ML frameworks. We developed a feature vector implementation for bag-of-words approach.

Sherlock feature vectors are consumed for prediction and training by ML framework-specific implementations of a vector translator interface defined in Sherlock. Vector translators can be written to allow nearly any ML framework to interact with Sherlock. We have developed vector translators for LIBSVM and Mallet CRF implementations of ML algorithms.[2,3] Once a Sherlock vector translator has been created for a specific ML framework, it is reusable across projects and feature vectors without the need for customized reprogramming.

To use Sherlock, an NLP pipeline is created that can extract the features of interest and generate a Sherlock feature vector. The training module stores a collection of these feature vectors and uses a vector translator to perform the training in the specific ML framework using a training set or k-fold validation. With a trained model, the same NLP pipeline is used and the feature vector is used in the prediction model for classification.

With Sherlock, NLP developers can easily incorporate ML with reusable feature vectors and interchange ML frameworks.

Acknowledgements

This work was supported using resources and facilities at the VA Salt Lake City Health Care System with funding from VA Informatics and Computing Infrastructure (VINCI), VA HSR HIR 08-204 and the University of Utah.

References

1. Ferrucci D, Lally A. UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Nat Lang Eng*. 2004;10(3-4):327–348.
2. Chang C-C, Lin C-J. {LIBSVM}: A library for support vector machines. *ACM Trans Intell Syst Technol*. 2011;2(3):27:1–27:27.
3. McCallum AK. MALLET: A Machine Learning for Language Toolkit. 2002. Available at: http://mallet.cs.umass.edu.

Impact Of Implementation Efforts on AWARE Checklist Compliance

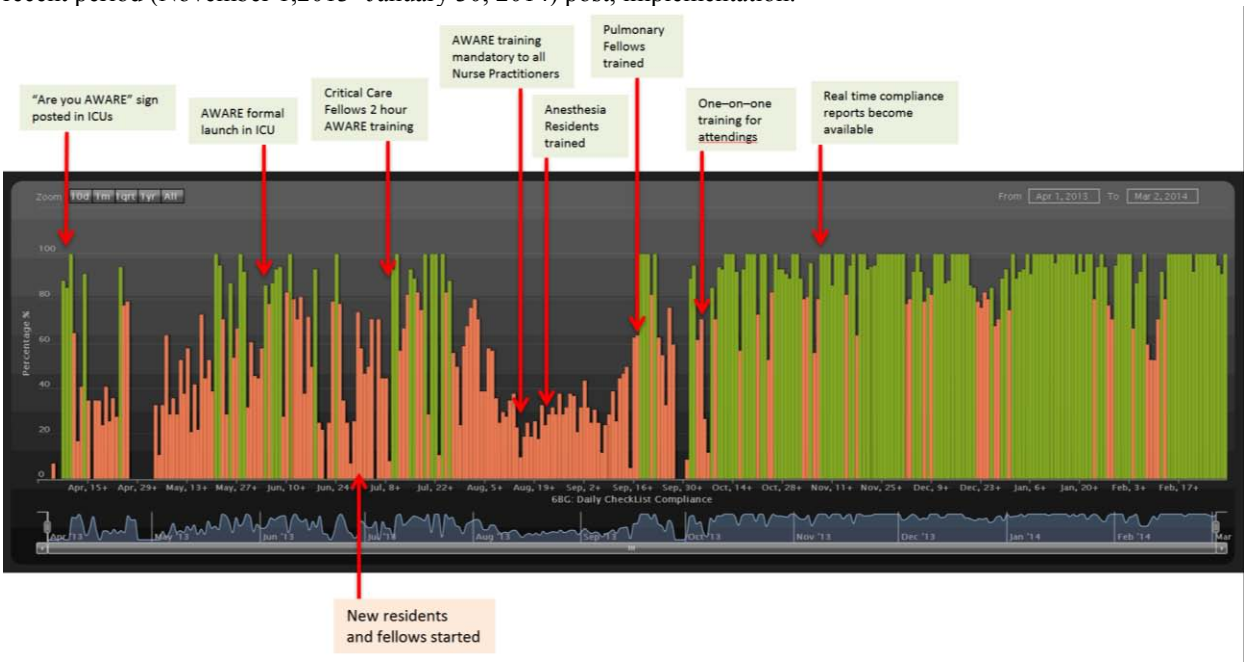
J. Giri MD, J. O'Horo MD, MPH, R. Sevilla Berrios MD, M. Resner, V. Fedosov MD, PhD, V. Herasevich MD, PhD, O. Gajic MD, B. Pickering MD,
Mayo Clinic, Rochester, MN

Abstract: An electronic checklist in AWARE, a novel Electronic Medical Record (EMR) viewer, was formally introduced in four Intensive Care Units (ICU) at the Mayo Clinic in April 2013. Initial adoption was low, leading to a variety of implementation efforts aimed at improving checklist adoption. The highest and most sustainable compliance of the checklist was observed 7 months post implementation efforts.

Introduction: AWARE is a data integration and decision support tool developed for the ICUs at the Mayo Clinic. In a pilot study, AWARE was shown to decrease the amount of time spent in data gathering by the providers, thereby decreasing workload [1]. The implementation of any new electronic system for clinical care is critically impacted by organizational, workflow and technological factors [2]. Training, information technology support, practice size, and choice of EMR were cited as some of the significant organizational factors affecting implementation [3]. We recently implemented a new feature on the AWARE platform, a best practice checklist aimed at improving care in the ICU.

Methods: The novel EMR viewer AWARE was designed with a context sensitive best practice checklist for ICU patients. Prior to the formal checklist launch in April 2013, we of designed educational materials, sought various committee and leadership approvals to formalize training, and trained physician providers at various levels. The impact of the various implementation efforts was closely monitored in the medical ICU. Recognizing challenges throughout the implementation process and low compliance led to using several plan-do-study-act (PDSA) cycles. Levels of checklist use, as well as each PDSA cycle change, are presented in the figure The rates of checklist compliance were used as the outcome of interest.

Results: The checklist compliance rates steadily progressed from 28% in the initial period (June 1, 2013- August, 30, 2013) to 64% in the late period (September 1, 2013 – November 30, 2013) and eventually above 78%, in the recent period (November 1,2013- January 30, 2014) post, implementation.



Conclusion: In order to achieve consistent compliance with the checklist, multiple interventions were required. The significant interventions included training and education, familiarity with electronic tool and involving leadership at every stage of intervention. We now have to focus on finding methods to sustain this impact and export the model to other units

References:

1. Vitaly Herasevich, et al. nformation Technology Can Reduce Time Spent on Data Gathering Activities in the ICU. HIMSS Physicians symposium, 2013.
2. Campion, T.R., Jr., et al., Implementing unique device identification in electronic health record systems: organizational, workflow, and technological challenges. Medical care, 2014. 52(1): p. 26-31.
3. McGinn, C.A., et al., Comparison of user groups' perspectives of barriers and facilitators to implementing electronic health records: a systematic review. BMC medicine, 2011. 9: p. 46

Data Driven Approach to Vital Sign Parameters at Lucile Packard Children's Hospital Stanford (LPCH)

Veena V. Goel M.D.[1], Sarah F. Poole[2], Terry S. Platchek, M.D.[1], Christopher A. Longhurst M.D., M.S[1,2], Paul J. Sharek, M.D., M.P.H.[1], Jonathan P. Palma, M.D., M.S.[1,2]
[1]Lucile Packard Children's Hospital Stanford, Palo Alto, CA;
[2]Stanford University School of Medicine, Palo Alto, CA

Background
Multiple reference ranges exist for age-stratified pediatric vital signs. These vital sign parameters, although widely accepted in both inpatient and outpatient pediatric clinical settings, result in a large number of out-of-range heart rate (HR) and respiratory rate (RR) measurements among hospitalized children, which contributes to alarm fatigue. We aim to establish safe data-driven HR and RR parameters for hospitalized patients at Lucile Packard Children's Hospital (LPCH) Stanford, and evaluate the proportion of out-of-range LPCH vital signs that result.

Methods
Vital sign data (HR and RR) were extracted from the LPCH clinical data warehouse for all non-ICU hospitalized patients <18 years old during calendar year 2013. The data were stratified according to patient age, and a balanced random sampling of the data from each of the 7,202 unique patients yielded 62,508 vital signs measurements. Using these measurements, data-driven 5th and 95th percentile ranges for HR and RR were established and compared to the ranges published by Bonafide et al.[1] using a similar method. A validation cohort of 2,287 pediatric patients at LPCH from January to May 2014 was constructed. The LPCH data-driven HR and RR limits were applied to all 82,959 HR and RR measurements in this cohort to find the proportion of out-of-range results. This was then compared to the proportion of out-of-range results using the National Institutes of Health (NIH) 2004 norms, which currently define alarm limits by LPCH policy. Finally, HR and RR data were analyzed in 148 patients who had a rapid response team (RRT) or code event from March 2013 to March 2014 to determine how many patients had out-of-range measurements using the current parameters compared to the proposed parameters. In this population of patients, a true positive was defined as having at least one out-of-range vital sign during the 12-hour period leading up to the RRT or code event. Comparison was performed against randomly selected 12-hour periods in an age-matched control cohort of non-RRT/code patients.

Results
The HR and RR percentile tables developed based on LPCH patients were comparable to the curves developed by Bonafide et al. Most notably, the 95th percentile values for both HR and RR were significantly higher than current LPCH upper limits. Using calculated LPCH 5th and 95th percentile limits, 55.7% fewer measurements would have been considered out-of-range when applied to all HR and RR measurements in the validation cohort. Analysis of HR and RR data in the 12 hours leading up to RRT and code events demonstrated that 136 (94.4%) of the 144 patients who previously had out-of-range measurements were still identified as abnormal. Two additional patients not previously flagged as having any vital sign abnormalities were identified as having low HR values by the new limits. Eight patients previously identified as having high RR measurements were excluded by the new limits. Manual review of these patients' charts, however, showed that all 8 of the RRT/code events that would be within the new RR norms were not called exclusively for the reason of abnormal respiratory rate.

Conclusion
A large proportion of HR and RR values among children at LPCH are out of range according to current vital sign reference ranges. Our historical data suggest that adopting data-driven values for HR parameters will safely decrease the proportion of out-of-range measurements, and should decrease alarm fatigue by increasing the specificity of monitor alarms without decreasing sensitivity to RRT and code events. With regards to data-driven RR parameters, there was a slight drop in sensitivity towards detecting RRT and code events, and therefore future work to further tune these limits could be considered. We recommend this as the first step towards a more customized approach to inpatient bedside alarm fatigue reduction, consistent with a learning healthcare system philosophy.

References

1. Bonafide CP, Broday PW, Keren R, Conway PH, Marsolo K, Daymont C. Development of heart and respiratory rate percentile curves for hospitalized children. Pediatrics 2013: 131(4)e1150-7.

A Qualitative Analysis of Patients' Behavior in an Online Health Community: Dynamics of Self-regulated Health Goal Achievement

Nadee R. Goonawardene, BSc[1], Sharon S L. Tan, PhD[1]
[1]Department of Information Systems, National University of Singapore, Singapore

Abstract

With the increased popularity of online social media platforms in healthcare, medical information sharing has become democratic and patient controlled. As a result, patients are more empowered and equipped with necessary information to proactively manage their health. Our research examines the factors that facilitate or inhibit the success of patients' self-health management efforts initiated in an online healthcare social network.

Introduction: Online social networks based on common health concerns are becoming increasingly popular. Virtual healthcare communities built on top of Health 2.0, have provided people with common interests a channel to gather virtually and share experiences, discuss treatment issues or guide and encourage each other to achieve individual healthcare targets. More importantly, for people who face challenges in developing self-regulatory behavior that is essential for self-health management, online health community participation may be useful in developing and motivating such behaviors. The objective of this study is to explore and evaluate how online health community participation can affect self-regulated health behaviors. In particular, we would like to explore the question: "How could Health 2.0 help individuals set and achieve their health management goals in developing self-regulated health behaviors? What are the dynamic relationships between goal setting, online social support and goal achievement?"

Methodology: This study adopts a qualitative research methodology to address the identified research questions. Data was gathered from Dailystrength.com, which is a popular community based healthcare website comprising of 500+ special interest groups. Data from the special interest group of obesity was chosen for the analysis. Obesity requires a substantial amount of self-health management, including life style modifications and self-monitoring. Therefore, we believe that such patients possess the required level of motivation in-order to set goals and continue to work towards achieving it. Individually created goals (i.e. lose weight goals) from 1026 profiles were gathered and a total of 786 completed or given-up goals were chosen for the analysis. Information such as: goal descriptions, progress updates, comments, moods and the anticipated end dates of goals were extracted. In the first phase, 'open coding' was carried out to extract important themes and variables. Then, 'selective coding' of all the data was carried out based on the core variables found at the open coding stage.

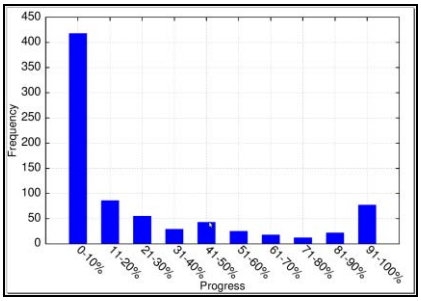


Figure 1. Last updated goal progresses

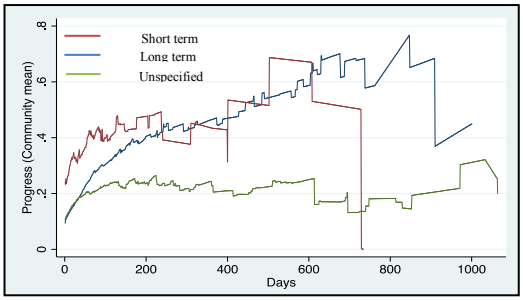


Figure 2. Community means distribution for three types of goals

Preliminary results: As illustrated in Figure 1, the highest number of goals has ended with little progress. By charting the goal progression for all the goals, we found six goal achievement patterns. A follow up investigation is then carried out to validate whether certain goal progression patterns are more likely to result in goal completion and how goal characteristics, mood and types of social support interact to influence progression and subsequent success in goal achievement. Based on the anticipated completion date, goals were divided into three categories: Long-term, short-term and goals with no deadline. Figure 2 shows the mean progression of these three categories across the community. Our data is consistent with the extant research on goal achievement whereby short-term goals mobilize efforts sooner and guide individuals to achieve goals faster. In contrast, long-term goals make it easy to postpone efforts; therefore the goal progress rate could be slower compared to short-term goals. The successful completion of our research can provide recommendations for healthcare practitioners in designing efficient online health management programs.

Acknowledgement: This study is supported by the MOE AcRF grant no. T1- 251RES1111 awarded to the second author.

Longitudinal Tracking of Pain Phenotypes in Electronic Health Records Using an SVM

Clare T. Grasso[1], Anupam Joshi, PhD[1]
[1]University of Maryland Baltimore County, Baltimore, MD

Abstract

Mining electronic health records (EHR) holds great potential to increase patient safety and quality of care. Most natural language processing (NLP) systems for extracting data from medical text rely on the underlying grammatical structure of the text. However, text from EHRs is oftentimes grammatically incorrect or incomplete, contains many nonstandard abbreviations, and is interspersed with other semi-structured text. This research focuses on a new approach that does not rely on the underlying grammar.

Introduction

As computer-readable clinical health data becomes increasingly available through the use of electronic health records (EHRs), the potential to mine this data for health safety, quality of care, surveillance, and clinical decision support (CDS) systems is very great[1]. Most natural language processing (NLP) systems developed for extracting medical concepts rely on the grammatical structure of the formal text[2]. However, these systems can have difficulty with the weak grammatical structure of clinical narratives. For example, the free-form clinical text entered by physicians, nurses, and other health professionals is filled with sentence fragments, missing punctuation, many nonstandard abbreviations, and is interspersed with structured and semi-structured text such as lab results and questionnaires. Instead, we developed a machine-learning approach for extracting concepts from the clinical narrative that does not rely on underlying grammar.

Methods

For this project, the data consisted of ten patient histories from the Baltimore VA Medical Center's VISTA EHR. A number of medical conditions were represented. Each history spanned several days and contained a sequence of 50-100 different notes pertaining to the clinical encounters during a hospital admission, and included notes from several different types of health care providers including physicians, nurses, radiologists, surgeons, and physical therapists.

The prototype system focused on increasing patient quality of care by tracking references to patient-reported pain, including the severity, location, onset, and duration in the hospital record. Each line of text was labeled as positive or negative for the existence of any of these elements, and was preprocessed to remove punctuation and digits. The text was tokenized by white space and then featurized using the log of the term frequencies of unigrams and bigrams. A SVM was trained on these features and tested using ten-fold cross validation.

Results

This new approach achieved precision, recall, and F-score greater than 70% in discovering references to patient-reported pain in clinical narratives.

Reference

1. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inform. 2009 Oct;42(5):760-72.
2. Heintzelman NH, Taylor RJ, Simonsen L, Lustig R, Anderko D, Haythornthwaite JA, Childs LC, Bova GS. Longitudinal analysis of pain in patients with metastatic prostate cancer using natural language processing of medical record text. J Am Med Inform Assoc. 2013 Sep-Oct;20(5):898-905.

Electronic Consenting Program

Timothy E. Grose, BS, Ronald L. Jenks, BS, Wes D. Kincaid, BS, Miguel Restrepo, PhD
Moffitt Cancer Center, Tampa, Florida

Abstract
Identifying subjects that meet inclusion criteria for a specific research protocol and consenting them are critical to protocol success. Integrating our enterprise scheduling, registration, and protocol management systems aided us in matching a patient to a trial more quickly and efficiently. To make this a paperless process, capturing 21 CFR, part 11 compliant electronic signatures was imperative. Interfacing with enterprise systems assures notifications of protocol accruals are distributed appropriately.

Introduction
An open source application has been developed that integrates our enterprise registration, scheduling, and protocol management systems to facilitate the accrual process to protocols. Patients meeting the inclusion criteria for a protocol are identified by filtering on attributes stored within these systems. Connection with our scheduling system allows us to identify where and when a patient can be located for a study. The application provides protocol specific educational materials and consent documentation which can be electronically signed; over the internet or within the facility, complying with 21 CFR part 11[1]. After consenting, the distribution of the accrual to other enterprise systems reduces the amount of manual processes needed in management of the patient enrolling on a protocol.

Subject Identification
Using existing HL7 interfaces from our patient registration, scheduling, and protocol management systems we then select attributes of inclusion to filter patients to a targeted cohort. These attributes could be demographic or other limited clinical attributes. The scheduling integration joins the patients in the cohort to their appointments so we know when and where the patient can be found. Integrating the protocol management system provides additional attributes to identify patients who require consenting to new versions or amendments to the protocols they are already accrued to.

Patient Encounter
During a patient encounter, protocol specific education materials are presented to the patient within the application. The consent document is then presented and electronically initialed and/or signed where required. At the completion of the consenting process a copy of the consent can be sent via email to the patient. In the event of a non-consenting patient the encounter result is recorded to know how to approach the patient at a future opportunity. Examples of these would be those who deferred, declined, or the consent process was interrupted.

Electronic Signature Capture
We have two CFR compliant ways to electronically capture the patient's signature: guided and un-guided. Guided signature capture is done capturing a graphical signature with the Consentor at the patient's side. Un-guided signature capture is accomplished when patients accrue to non-therapeutic trials over the internet. To assure CFR compliance[1] an account is created by the patient using unique identifiers to verify the patient is who they identify as.

Consent Distribution and Notification
Once the consent is signed it is transformed to a PDF and centrally stored on the enterprise network, making it available for reference by other systems. Metadata regarding the consent process is sent via HL7 message types to the protocol management system which electronically creates a consent record for patient management on the study. Additionally, an update is sent to the electronic medical record which is then capable of auto-ordering additional research specimens when a specimen collection event occurs for the patient.

Conclusion
The consolidation of our hospital enterprise systems facilitates the finding of patients and reduces the time required to document accruals. Interfacing the hospital systems and electronically creating the consent records has reduced the hands-on time required for patient management on our protocols.

References
1. Electronic Records; Electronic Signatures, 21 C.F.R., Part 11 (2013)

How to de-identify a large clinical corpus in 10 days

C. Grouin, PhD[1], L. Deléger, PhD[1], J.B. Escudié, MD[2], G. Groisy, MD[2], A.S. Jannot, MD, PhD[2,3], B. Rance, PhD[2,3], X. Tannier, PhD[1,4], A. Névéol, PhD[1]

[1]CNRS UPR 3251 LIMSI, Orsay, France; [2]AP-HP, University Hospital HEGP, Biomedical Informatics and Public Health Department, Paris, France; [3]INSERM U1138, Université Paris Descartes, Sorbonne Paris Cité, Faculté de médecine, Paris; [4]Université Paris Sud, Orsay, France

Abstract

To de-identify a large corpus of clinical documents in French supplied by two different health care institutions, we apply a protocol built from previous work. We show that the protocol can be installed and executed by outside collaborators with little guidance from the authors. The automatic de-identification method used reaches 0.94 F-measure and human validation requires about 1 minute per document.

Introduction

Clinical corpora are useful for scientists to develop natural language processing methods for the clinical narrative. In order to ensure the robustness of those methods, access to real data is a critical point. As clinical records contain personal health information, de-identification tools have been designed to help protecting privacy.[1]

Methods

Herein, we apply the protocol we designed to rapidly de-identify a new set of clinical records:[2] new documents to be de-identified are automatically pre-annotated with a statistical model built on similar resources. For new documents from the same hospital as our previous study[2] (corpus 1), we directly apply a model built with 100 training documents. For documents coming from a distinct hospital (corpus 2) we iteratively build a model as follows:

1. 20 documents were de-identified automatically, using a rule-based method[3] for the first iteration and a CRF (Conditional Random Fields) model for subsequent iterations, and manually validated by scientists and physicians

2. a statistical (CRF) model was built using all the manually validated de-identified documents available from all iterations. The latest model is then used for the next iteration of step 1. The features used to build the CRF model include surface features *(case, punctuation, digit, token length, etc.)*, deep features *(part-of-speech, lexical look-up, etc.)*, and external features *(position of the token in the record, cluster id).*

For the initial seed corpus of 20 documents, six different types of clinical notes were chosen to ensure robustness: 3 medical certificates, 3 consultation reports, 3 exam certificates, 3 hospitalization reports, 5 follow-up care letters, and 3 staff consultation reports.

Results and Conclusion

Corpus 1: The automatic de-identification method reached 0.94 F-measure and human validation required less than 1 minute per document for 800 documents. Corpus 2: About 1 day was needed to install the tools to use in the de-identification protocol. The automatic de-identification method reached 0.92 F-measure and human validation required as little as 1 minute per document for 100 documents. Validation time decreased with annotator experience and improvement in the performance of the de-identification method (i.e. after more training documents become available as we progress through more iterations). The protocol can be executed by outside collaborators with no prior experience of de-identification. It provides adequate support for the de-identification of a large corpus requiring little time and guidance.

References

[1]Meystre SM, Friedlin FJ, South BR, Shen S, Samore MH. Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Med Res Methodol. 2010;10.

[2]Grouin C, Névéol A. De-identification of clinical notes in French: towards a protocol for reference corpus development. J Biomed Inform. 2014. In press.

[3]Grouin C, Zweigenbaum P. Automatic de-identification of French clinical records: comparison of rule-based and machine-learning approaches. Stud Health Technol Inform. 2013;192(Part 1):476–80.

An Informatics Approach to Building a Surgical Residents Log

Michael J. Grove[1], PhD, Jeff Emch[1], MBI, Mark Engelstad[1], DDS, MD, MHI
[1]Oregon Health & Science University, Portland, OR

Abstract

Tracking of oral and maxillofacial surgical resident training experiences is reliant on non-standardized tools and methods resulting in compromised data for individual and program evaluation. To address this, we developed a domain ontology built on a SNOMED-CT foundation that drives a new resident surgical tracking log. Initial results are a new >7,000 concept domain ontology with properties linking surgical diagnostic and procedural experiences to a set of new educational concepts.

Introduction

Oral and maxillofacial surgery (OMS) is a surgical specialty that manages disorders of mouth, face, neck, and jaw. To maintain program accreditation and assess individual trainee progress, OMS trainees must log their surgical experiences, including disorders they have managed and procedures they have performed. Across all OMS training sites, current systems for logging and educational reporting are localized and non-standardized and rely on the use of payment coding schemes such as ICD-9 and Current Procedural Terminology (CPT). Reimbursement-based coding systems do not provide sufficient granularity of surgical experiences and do not always provide accurate clinical research data[(1)] . The consequence is a heterogeneous collection of surgical training data within and across programs that does not enable adequate assessment or comparison.

Methods

We are developing an ontology-driven OMS resident education log ("OMSLog") for use within the Oregon Health & Science University (OHSU) OMS training program and designed to scale for respective implementations at other OMS programs. The log has a blended architecture that consists of a browser interface for trainees to record their experiences, and faculty, and/or administrators to review and report on these events. The Java-based system is supported by a traditional relational database (RDB) back-end for data storage and reporting and a Resource Description Framework (RDF) triple-store for diagnosis and procedure data. Both the RDB and RDF triple-store are driven by a new domain ontology as terminological and semantic foundation. The initial ontology class hierarchy was developed by an OMS expert who incorporated domain-enhanced concepts into a curated SNOMED-CT extract.

Results

Current results are a 7,488 concept domain ontology being curated in the Protégé Ontology Editor environment. The ontology is modeled to represent surgical diagnosis and procedure activities and to characterize them according to a set of newly defined OMS educational categories. The use of the ontology provides increased granularity of surgical experiences and allows sophisticated classification of the educational characteristics of resident surgical activity for improved reporting capabilities. The ontology will also serve as a domain knowledge feedback mechanism and will support subsequent secondary use of de-identified data. Future steps include evaluating the ontology through graph metrics, demonstration of the ontology's logic structure to answer domain competency questions, and usability testing for the log user interface.

References

1. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses: ICD code accuracy. Health Services Research. 2005: 1620–39.

Automation in Healthcare: is Automation Bias a side effect?

Anupama E. Gururaj, PhD & Dean F. Sittig, PhD
School of Biomedical Informatics, University of Texas Health Sciences Center at Houston, Houston, TX, USA

Introduction: Use of automation in the healthcare industry has resulted in unintended errors that were not foreseen previously. Automation plays a prominent role in clinical decision support systems (CDSS) in the healthcare setting. CDSS acceptance and performance are determined by two end-user characteristics: 1) Trust in Automation that determines the extent to which CDSS is used and 2) Automation Bias (AB) that defines the tendency of the user to over-rely on automation. AB is one of the major underlying causes for errors during CDSS use (1). Therefore, a review of literature focused on studies exploring AB in healthcare is warranted.

Methods: A survey of PubMed literature using specific keywords relating to CDSS and medical errors such as "CDSS errors", "decision support system error", "error bias automation", "automation clinical decision support error", "automation error", "Sittig D", "Bates W", "Kaushal R error", "Kate Goddard" was conducted. The keyword selection was based on a quick survey of literature to determine the names of primary groups working on CDSS error identification as well as terms relating to error and AB. The titles and abstracts of the search results were manually scanned for appropriateness and inclusion. Articles were included if the studies commented on negative effects of CDSS in the medical domain only. Where possible, complete texts of the papers were retrieved and the texts were perused for relevant information. Resulting relevant literature was used to discuss and characterize the errors due to AB in the medical domain.

Results & Discussion: The PubMed search with the different keywords resulted in 2560 unique records that were retrieved. Of these, 22 were found to be relevant to the topic of study for which we were able to obtain full texts. To a large extent, studies defining AB in healthcare have focused on computer-aided detection (CAD) in radiology. Most of the studies showed heterogeneous results with reference to the effect of AB on errors. Modifiers that influence AB have been extensively investigated in other domains such as the aviation industry. A listing of these factors and their effects on AB is shown in Figure 1 which is a theoretical framework developed by Goddard et.al., based on the Theory of Technological Dominance (2). On the basis of our review of literature with regard to AB and its effectors in the medical domain, we have defined the primary role of four elements (encircled in red in Figure 1). Task inexperience lead to increased AB. AB increased when physicians were less confident of their own diagnosis and the converse was also found to be true. Therefore AB is a trade-off between self-confidence and trust in CDSS. Studies showed that increased task complexity increased reliance on CDSS and therefore increased AB. Detailed analysis of each investigation also underlined the parallels in etiology of AB between medical and other fields. Therefore, strategies that have been effective in mitigating AB in other domains (3) could be successful in healthcare too. Increased user accountability lead to decreased AB. Likewise, educating users of the reasoning behind DSS decisions and training users to be aware of appropriate reliance on the DSS reduce AB. The position of the DSS advice and relaying the decision as information rather than a command or recommendation decreases AB. Finally, decreasing screen details which translates to less clutter and more focused information decreased AB. Details of the studies determining mitigation strategies will be discussed in the poster. Given the potentially serious outcomes resulting from incorrect medical decisions, it would be beneficial to examine the negative impact of introducing automated clinical advice, as well as the overall positive effects of CDSS on medical decision-making.

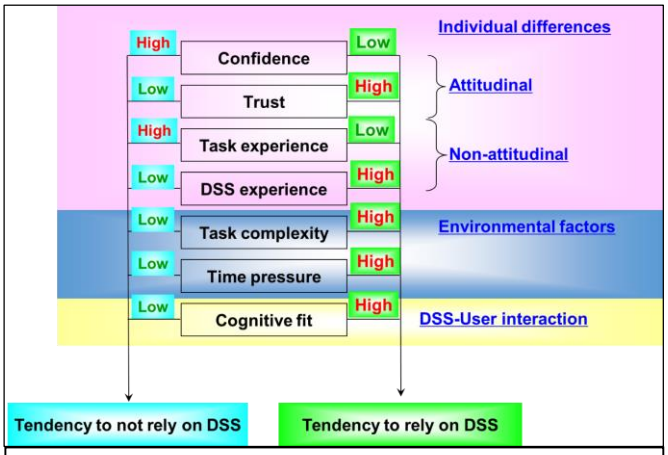


Figure 1. List of factors affecting AB as determined from all domains. The major players in the medical domain are encircled in red (Adapted from 2)

1) Jaspers MW, Smeulers M, Vermeulen H, Peute LW. Effects of clinical decision-support systems on practitioner performance and patient outcomes: a synthesis of high-quality systematic review findings. J Am Med Inform Assoc. 2011;18(3):327-34
2) Goddard, K., Roudsari, A., Wyatt, J. Decision Support and Automation Bias: Methodology and Preliminary Results of a Systematic Review. Int. Pers. Health Infor. 2011; 3-7
3) Lee, J.D., Review of a Pivotal Human Factors Article: "Humans and Automation: Use, Misuse, Disuse, Abuse". Human Factors: The Journal of the Human Factors and Ergonomics Society. 2008; 50(3): 404-410

The Business Value of IT in Healthcare: The Case of Cleveland Clinic's Online Second Opinion System

Peter Haddad [1, 3], Jonathan Schaffer[2] Nilmini Wickramasinghe [1, 3]

[1]RMIT University, Melbourne, Victoria, Australia, [2] Cleveland Clinic, Cleveland, Ohio, USA, [3]Epworth HealthCare, Melbourne, Victoria, Australia

Abstract

In an environment of rapid development of new clinical informatics solutions claiming to provide better healthcare delivery, there is a paucity of systematic frameworks to robustly measure the actual value of these systems. The following proffers such a model grounded in the information systems literature and assessment of business value to address this significant void.

Introduction

Today numerous clinical informatics solutions are being designed and developed claiming to provide superior healthcare delivery. However, too many of these solutions fail and many are dubious regarding their real value. To address this void an integrated model to assess the business value of these systems is developed.

Defining Value in Healthcare Delivery

'Business value of IT' is used to refer to the organizational performance impacts of IT, including cost reduction, profitability improvement, productivity enhancement and competitive advantage [1]. While specific perspectives are a key to viewing the critical impact of any IT system, there are advantages from the business, administrative, technology and most importantly, clinical reference points such as patient experience. For healthcare deliver, access and cost are cruical. Further, value is often defined in terms of the expenditure outcome benefits, divided by the cost expenditure while benefits, from a patient's perspective, include the quality of healthcare outcomes, the safety of the delivery process, and the services associated with the delivery process [2].

The Proposed Model

In order to develop an integrative model that will assess the value of a clincial informatics solution, perspectives of healthcare value from the respective points of view of all key stakeholders must be considered. To operationalize the IT resource, the IT portfolio is classified into infrastructure, transactional, informational, and strategic [3]. From an organizational perspective, The Enterprise of Healthcare Delivery Model provides useful insights. Finally, it is necessary to recognize the socio-technical perspective of these systems at four interrelated levels: (i) Clinical practices (people); (ii) Delivery operations (processes); (iii) System structure (organizations); and (iv) Healthcare ecosystem (society) [4] which all must work togehter to provide a better patient experience.

To test the potential of the proposed model, the specific case of the MyConsult® second opinion programat the Cleveland Clinic is applied. The subsequent analysis serves to demontrate that not only does this highlight the strengths and value of the MyConsult program but it also serves to demosntrate the robustness of the proposed model.

Conclusions

There is a clear need for a systematic framework to assess the business value of clincial informatics solutions. To address this need, a suitably robust model is developed from various bodies of IS and business literature. The proffered model is then tested using the second opinion program developed at the Cleveland Clinic.

References

[1] Melville, N., Kraemer, K. & Gurmaxani, V. 2004. "Review: Information Technology and Organizational Performance: An Integrative Model of IT Business Value". MIS Quarterly, Vol. 28 No. 2, pp. 283-322.

[2] Porter, M. & Teisberg, E. 2006. "Redefining Healthcare: Creating Value-Based Competition on Results". Harvard Business School Press, Boston

[3] Weill, P. & Broadbent, M. 1998. "Levering the New Infrastructure: How Market Leaders Capitalize on Information Technology". Harvard Business School Press, Cambridge, MA

[4] Rouse, W.B. & Cortese, D.A. 2010. "Engineering the System on Healthcare Delivery", IOS Press.

Automatic Extraction of Electronic Health Record System Data to Assist Evaluation of a Status Epilepticus Clinical Protocol

Baria Hafeez[1], Juliann Paolicchi[3,4,5], Steven Pon[2,4,5], Joy Howell[2,4,5], Zachary Grinspan[1,3,4,5]

[1]Department of Healthcare Policy and Research, Weill Cornell Graduate School of Medical Sciences, New York, NY
[2]Division of Pediatric Critical Care, Weill Cornell Medical College, New York, NY
[3]Division of Pediatric Neurology, Weill Cornell Medical College, New York, NY
[4]Komansky Center for Children's Health, Weill Cornell Medical College, New York, NY
[5]New York Presbyterian Hospital, New York, NY

Background: Status epilepticus is a neurological emergency characterized by unremitting seizures. Under these circumstances, immediate and precise actions are required to stop the seizure, and prevent brain injury. The Komansky Center for Children's Health has developed a "status protocol," which describes the suggested therapies to stop a patient from seizing. It is unclear the extent to which an evaluation of adherence to this protocol can be done electronically. Conducting such an adherence evaluation may be difficult because of the amount of data that needs to be reviewed. An automated tool to extract and display the data would have tremendous value both for quality assurance, to determine whether a protocol is properly being adhered to, and for research, to determine the effectiveness of the protocol. Electronic health record (EHR) systems may potentially solve the challenge of data collection and management to support evaluation of clinical protocols.

Methods: A retrospective chart analysis was conducted using information from the EHR system on a convenience sample of seven children with status epilepticus. We reviewed charts to qualitatively determine how well the data can be extracted and visualized. We used the R software environment, with the "timeline" package to visualize each patient's first 24 hours of care.

Results: Qualitatively, our observations are as follows: (1) most clinical data is well labeled in structured fields within EHR; (2) medication administration data is encoded in multiple places, with differing levels of detail and occasional inconsistencies; (3) the exact time of transfer from the emergency department to the intensive care unit (ICU) is not clearly indicated. However, this information can be inferred by searching for the first vital signs recorded in the ICU; (4) proxy measures need to be used to determine precise times of endotracheal intubation, arterial, and central line placement; (5) a "timeline" style visualization can rapidly clarify a patient's clinical course; (6) a clinical expert is needed to determine seizure onset and cessation; (7) free text EEG reports may not contain the exact time of seizure control.

Conclusion: Our observations indicate that several key elements of a protocol evaluation could be automated and visualized; however, manual chart review may still be necessary to obtain all the relevant data.

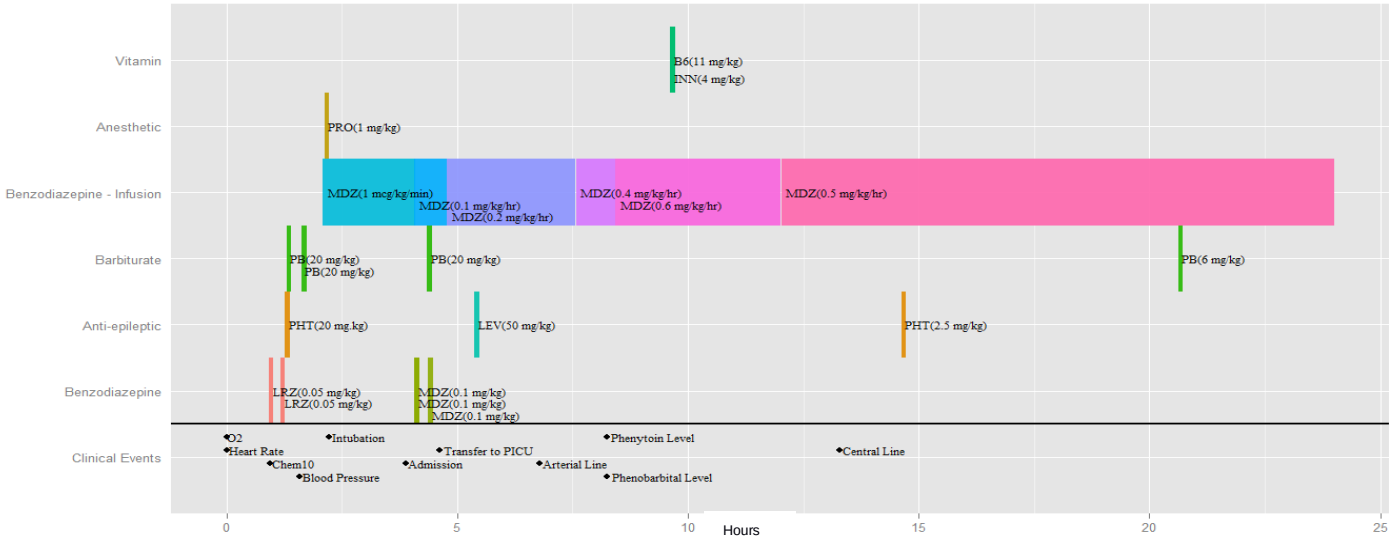


Figure 1: Timeline representation of a 1 year old boy who presented with Leigh syndrome.
Legend: LRZ – Lorazepam, PHT – Phenytoin, PB – Phenobarbital, PRO – Propofol, MDZ – Midazolam, LEV - Levetiracetam

Linking Adenomas Between Colonoscopy And Pathology Notes For PROSPR

Scott Russell Halgrim, MA[1], Leslie Sizemore, BFA[1], Edward Pham, MS[1], Gabrielle Gundersen, BA[1], David S. Carrell, PhD[1], Karen Wernli, PhD[1], Jessica Chubak, PhD[1], Carolyn M. Rutter, PhD[1]
[1]Group Health Research Institute, Seattle, WA

Abstract

In colorectal cancer screening, the number and size of adenomas found drive the recommended follow-up interval, but are often only available in unstructured text fields and split across reports. We augment an NLP tool with heuristics related to colonoscopy processes to increase the system's ability to count and size adenomas.

Introduction

Colonoscopy is a colorectal cancer screening test whose follow-up time is largely based on the number and size of adenomas found[1]. At Group Health, this information is electronic text in the colonoscopy procedure and pathology reports; the pathologist determines if a polyp is adenomatous, yet the procedure note often has more accurate counts and sizes. We use a natural language processing (NLP) system to find results above size and count thresholds.

Methods

We used the procedure and pathology notes of 248 patients selected randomly from all 2011 colonoscopies with pathology at Group Health. The gold standard was created by manual chart review. To detect adenomas, we extended the GATE[2] system of Harkema et al[3]. With initial modifications, the system detected cases of adenomas with 96% sensitivity, but performed worse at identifying cases of multiple (three or more) or large (greater than 10 mm) adenomas. We describe changes made to find more of these cases.

Gastroenterologists group samples into jars by colon area and pathologists write reports accordingly. To find cases of multiple adenomas, we use this method: if a jar's location is unique in a report, the sample is described as "fragments," and all jar contents are described as adenoma, then we set the number of polyps in that jar to match the number in that location in the procedure report. We classify a case as having a large adenoma if all samples in a jar with a unique location are adenomas and the procedure report lists a large polyp at that location.

Results and Discussion

Table 1 shows inter-annotator agreement and how our rules improved sensitivity. Although our rules' conditions are strict, we find many more cases of interest with little loss in specificity. We plan to assess the system's portability by testing at other sites and assess the significance of these findings by testing on larger sample sizes.

Table 1. Sensitivity and specificity of the NLP system on categories of interest before and after adding heuristics.

Measure	IAA (Cohen's Kappa)	Original		Modified	
		Sensitivity	Specificity	Sensitivity	Specificity
Has 3+ Adenomas (N=46)	.917	.717	.990	.870	.960
Has Adenoma >= 10 mm (N=12)	.880	0.000	1.000	1.000	.983

Conclusion

Identifying count and size of adenomas detected during a colonoscopy is important in colorectal cancer screening. We increased our ability to count and size adenomas found by augmenting an NLP system with heuristics to synthesize information between the procedure report and the pathology reports.

References

1. Lieberman DA, Weiss DG, Harford WV, et al. Five-Year Colon Surveillance After Screening Colonoscopy. Gastroenterology. 2007;133:1077-85.
2. Cunningham H, Tablan V, Roberts A, Bontcheva K. Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. PLoS Comput Biol 9(2): e1002854. doi:10.1371/journal.pcbi.1002854.
3. Harkema H, Chapman WW, Sual M, et al. Developing a natural language processing application for measuring the quality of colonoscopy procedures. J Am Med Inform Assoc. 2011:18;150-156.

Older Adults Use of Online and Offline Sources of Health Information and Constructs of Reliance and Self-Efficacy for Medical Decision Making.

Amanda K. Hall, PhD[1], Jay M. Bernhardt, PhD, MPH[2], Virginia Dodd, PhD, MPH[2]

[1]University of Washington, Seattle, WA; [2]University of Florida, Gainesville, FL

Abstract

We know little about older adults' use of online and offline health information sources for medical decision-making despite increasing numbers of older adults who report using the Internet for health information to aid in patient/provider communication and medical decision-making. Therefore we investigated older adult users and nonusers of online and offline sources of health information and factors related to medical decision-making. This study found significant differences between users and nonusers of online health information related to their preferred health information sources and examined factors related to online health information engagement for medical decision-making.

Introduction

Patients' interest in accessing and utilizing information to inform their healthcare decision-making and increase their healthcare knowledge and management is well known. However, prior to the World Wide Web, patients usually obtained information from healthcare providers, mass media sources, or local community members deemed knowledgeable about health issues. Today the Internet offers access to a wide range of disease prevention and health management information that can be a useful aid in decision-making. The literature contains few studies investigating health information sources, reliance for medical decisions, and the role of self-efficacy in medical decision-making between older American users and nonusers of online health information. To address this gap, this study (1) assessed the relationship between users and nonusers of online health information and self-efficacy for medical decision-making (2) examined the relationship between users and nonusers of online health information and their Reliance, either self-reliant or physician-reliant, for medical decisions, and (3) investigated the differences between users and nonusers of online and offline health information sources.

Methods

We conducted a survey using random-digit-dialing of Florida residents' landline telephones. Using the Decision Self-Efficacy Scale and the Reliance Scale, we measured relationships between users and nonusers of online and offline sources of health information.

Results

Study respondents were 225 older adults (age range 50–92, M = 68.9, SD = 10.4), which included users (n = 105) and nonusers (n = 119) of online health information. Approximately 75% of all respondents sought information regarding health from offline *healthcare professionals* followed by the *Internet* (46.9%). However, users and nonusers differed in frequency and types of sources sought. Users of online health information preferred a self-reliant approach and nonusers of online health information preferred a physician-reliant approach to involvement in medical decisions on the Reliance Scale, $t(218) = -.3.09$, $p = .001$. No significant differences were found between users and nonusers on the Decision Self-Efficacy Scale.

Conclusions

More empirical research is needed to extend the literature on the use of the Internet as a patient decision aid for medical decision-making. As the Internet continues to be a predominate source of available health information, further research is needed to test constructs that predict use of online health information to bridge information and communication gaps between healthcare providers and older adult patients for improved health outcomes and shared medical decision-making.

Acknowledgments

This work was supported in part by the National Institutes of Health, National Library of Medicine (NLM) Biomedical and Health Informatics Training Program at the University of Washington (Grant Nr. T15LM007442).

Coordination-Based Analysis of Inter-Unit Handoffs

Saira Haque, PhD, MHSA[1], Craig Kuziemsky, PhD[2]
[1]RTI International, Center for the Advancement of Health IT, Research Triangle Park, NC; [2] Telfer School of Management, Ottawa, ON, Canada

Abstract
Routines are repetitive patterns of action that guide organizational activities, including handoffs. Electronic health records (EHRs) are often designed based on ideal-type routines that do not match how activities occur in practice. The mismatch can lead to issues during transitions such as inter-unit transfers. We conducted observations of routines, participants and information sharing during inter-unit transfers at two institutions. Variations in patient severity, type of transfer, other activities occurring in the unit and the dependencies determine if deviations from ideal-type routines lead to issues in handoffs. Issues were mitigated by articulation work and clarity of roles and tasks in routines. Our findings can be used to support EHR design to support handoffs.

Introduction
Handoffs between units are points which are prone to errors and other issues [1]. Thus, it is important to understand more about the issues that occur in transfers to develop processes, systems and contingency plans for them. Conceptualizing the handoff as a coordination routine with multiple aspects can yield important insights into how issues with handoffs occur and how they can be prevented.

Methods: Site Selection
We studied different types of inter-unit transfers at a community hospital in the United States and an academic medical center in Canada. The community hospital handoffs involved transfers from the interventional cardiology unit to an inpatient nursing unit or to an outpatient unit. The academic medical center handoffs focused on transfers to the operating room (OR) and then from the OR to the recovery or patient care unit.

Results: Conceptual Framework
Table 1 shows the coding scheme used from a previous study [2]. Routines, or repetitive patterns of action, guide organizational processes such as handoffs and was used to establish the boundaries of the handoff for analysis. Coding is in process for different variations of routines and the coordination and work that exists within the routine in practice to validate the conceptual framework and coding scheme.

Table 1 – Coding Scheme to Analyze Inter-Unit Transfers

Concept	Description
Articulation Work	Work required to continue or go on
Coordination	Managing dependencies between tasks
Routine	Repetitive patterns of action that govern organizational work
Ostensive routine	The ideal-type or theoretical routine (policies and procedures etc)
Performative routine	The actual performances of the routine
Organizational routine	Patterns that govern work for a process throughout an organization
Coordination routine	Repetitive patterns of action used to manage dependencies

Conclusion
Errors and other issues in inter-unit handoffs occurred for a number of reasons. Varying ideas of the ideal-type handoff across staff led to inconsistency in processes made it difficult to integrate knowledge about patients and communicate that knowledge when the patient transitioned between units. Differences from ideal-types can occur because of communication and coordination issues, staff changes, other activities occurring in the unit or patient severity issues. Applying these related concepts to inter-unit transfers across institutions can help further understanding of errors in handoffs.

References
[1] Patterson, E.S., et al., *Handoff strategies in settings with high consequences for failure: lessons for health care operations.* International Journal for Quality in Health Care, 2004. **16**(2): p. 125-132.
[2] Haque, S. N., Oesterlund, C., & Fagan, L. (2013). *Do deviations from ideal routines cause coordination errors? An exploration of coordination in an ambulatory care setting*. In Proceedings of the American Medical Informatics Association Annual Meeting, pp. 1.324023–1.324023. Washington: American Medical Informatics Association.

Connecting CCC to SNOMED CT and LOINC in a Terminology Server

Tiffany Harman, RN; Rachael Howe, RN; Tosh Kartchner, RN; Susan Matney, RN MSN
3M Health Information Systems, Inc., Murray, UT

Introduction

Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) and Logical Observations Names and Codes (LOINC) have been mandated as the terminologies to use when messaging coded data between settings – SNOMED CT for coded problems and LOINC for outcome measures. Nursing data is key to providing a complete clinical patient picture. The Clinical Care Classification (CCC) is an American Nurses Association recognized nursing terminology. It contains nursing care components that provide the framework for classifying nursing diagnosis, interventions and outcomes. Linking to SNOMED CT and LOINC will allow the nursing data to be messaged.

The 3M Healthcare data dictionary (HDD) is a terminology server that houses the standard terminologies required to document clinical care, including SNOMED CT and LOINC. CCC can be added to the HDD and linked to SNOMED CT and LOINC to facilitate interoperability between systems.

Methods

A terminology model is designed to incorporate all of the CCC entities (care components, diagnoses, interventions and outcomes) as concepts, relationships and representations. Each CCC concept is given a numeric unique concept identifier (NCID) and appropriately related to each other. After all the CCC concepts are created in the HDD, relationships are built between the CCC concepts and SNOMED CT and LOINC concepts.

Results

The CCC terminology model encompasses 21care components, 60 major diagnoses, 77 major interventions, 116 sub-categories for diagnosis, and 124 sub-interventions. Each sub-diagnosis has three outcomes: improved, stabilized and deteriorated. Each sub-intervention has four action types: assess, perform, teach and manage. Figure 1 shows an example each for diagnosis and intervention.

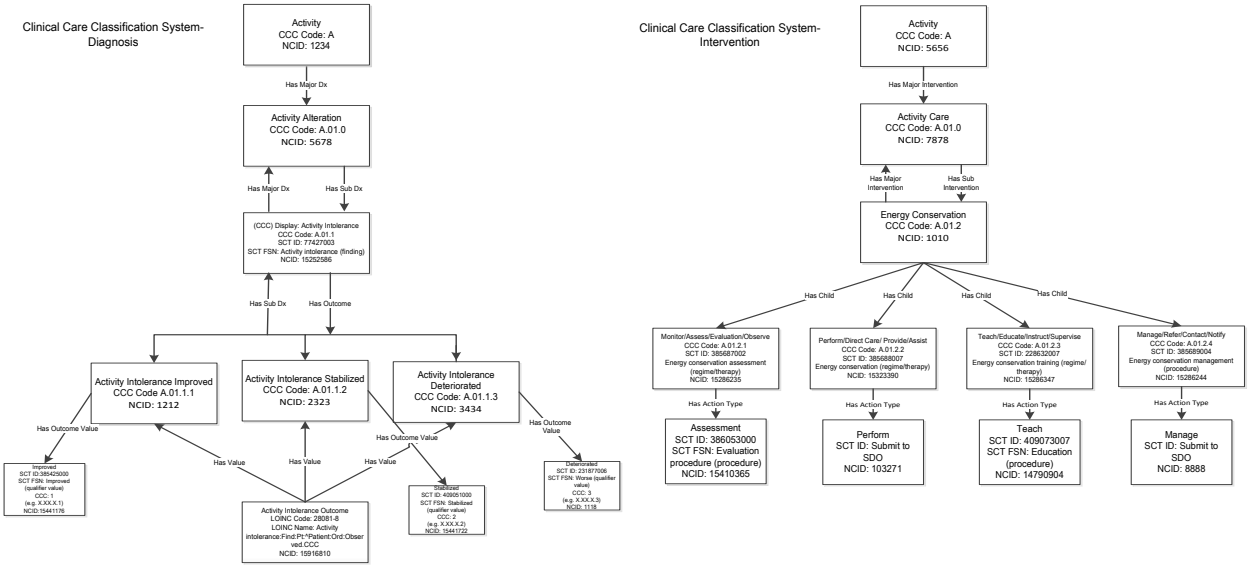


Figure 1

Conclusion

The HDD functions as a technology solution to implement standard terminologies. Incorporating the CCC within the HDD enables the continued use of CCC in systems and facilitates interoperability by providing terminology translations between CCC, LOINC and SNOMED CT. Continued use of standard nursing terminologies maintains nursing knowledge and promotes standardized documentation of the nursing process. The mapping to SNOMED and LOINC supports national messaging standards.

Development, testing, and refining the severe sepsis and septic shock sniffer

A.M. Harrison, BS; C. Thongprayoon, MD; R. Kashyap, MBBS; V. Smith, MD, A. Hanson, O. Gajic, MD, MSc; B.W. Pickering, MD, MSc; V. Herasevich, MD, PhD

Mayo Clinic Rochester and Mayo Clinic Arizona

Background: From 1979-2000, the incidence of sepsis in the United States increased from 164,000 to 660,000 cases[1]. Sepsis was also reported as the most expensive condition treated in US hospitals in 2011, with an aggregated cost of $20.3 billion.[2] However, current sepsis detection algorithms have not considered alert action in the context of failure to rescue.[3]

Objective: Develop and test an automated alert algorithm ("sniffer") for detection of severe sepsis failure to rescue.

Methods: Retrospective cross-sectional study using independent derivation and validation cohorts (Table 1). We examined all adult first-admissions to the medical ICU at Mayo Clinic in Rochester, MN, from January through March 2013 (N = 587).

Table 1: Cohort demographics

Variable	Derivation (N = 293)	Validation (N = 294)	P value
Age (± SD)	63.7 ± 18.5 years	63.3 ± 19.0 years	0.7936
Sex (% male)	52% (N = 151)	54% (N = 159)	0.5367
Hospital LOS, (± SD)	7.1 ± 6.7 days	7.6 ± 8.4 days	0.4426
ICU LOS, (± SD)	2.1 ± 2.6 days	2.2 ± 3.1 days	0.7081
SOFA score, Day 1 (± SD)	4.6 ± 3.5	4.5 ± 3.5	0.7201
APACHE III score, Hour 1 (± SD)	61.6 ± 25.6	59.1 ± 23.0	0.2275

Algorithm validation was performed against the "gold standard" of manual chart review by two trained reviewers, with one super-reviewer for cases of disagreement. Algorithm development and testing was performed using iterative recursive data partitioning and critical appraisal of false positive and negative alerts. The algorithm is based on the following variables: suspicion of infection, system inflammatory response syndrome, organ dysfunction (lactate and systolic blood pressure), and shock (vasopressors and fluid resistant hypotension).

Results: The ability of the first technical iteration of the severe sepsis sniffer on the derivation cohort to detect severe sepsis and/or septic shock was suboptimal: 59% sensitivity, 97% specificity, 92% Positive Predictive Value, and 83% Negative Predictive Values. Critical appraisal of false positive and negative alerts, along with iterative introduction of new clinical variables (mean arterial blood pressure, bilirubin, platelets, INR, mechanical ventilation, creatinine, PaO2/FiO2 ratio, urine output, GCS score, and fluid balance) into the algorithm, was then performed, which resulted in an increased sensitivity of 82%. Testing this algorithm on the validation cohort shows similar diagnostic performance (Table 2).

Table 2: Diagnostic performance of sniffer algorithm iterations

Algorithm	Sensitivity	Specificity	PPV	NPV
Algorithm 1 (Initial)	59	97	92	83
Algorithm 2 (Debugging)	82	97	93	92
Algorithm 3 (Validation)	80	96	92	91

Discussion: Current detection performance of our sepsis alert systems is similar to previous reports.[4,5] Recursive data partitioning and validation of the failure to rescue domain are anticipated to further refine the sniffer for eventual implementation in the clinical setting.

Conclusion: The validated sepsis sniffer showed sensitivity in good agreement with final derivation algorithm. However, this sniffer could be improved further by adding a failure to rescue component to the algorithm.

References:

1. Martin GS, Mannino DM, Eaton S, Moss M. The epidemiology of sepsis in the United States from 1979 through 2000. *N Engl J Med.* Apr 17 2003;348(16):1546-1554.
2. Torio CM AR. National Inpatient Hospital Costs: The Most Expensive Conditions by Payer, 2011. *HCUP Statistical Brief.* August 2013 2013;#160.
3. Silber JH, Rosenbaum PR, Schwartz JS, Ross RN, Williams SV. Evaluation of the complication rate as a measure of quality of care in coronary artery bypass graft surgery. *JAMA.* Jul 26 1995;274(4):317-323.
4. Hooper MH, Weavind L, Wheeler AP, et al. Randomized trial of automated, electronic monitoring to facilitate early detection of sepsis in the intensive care unit*. *Crit Care Med.* Jul 2012;40(7):2096-2101.
5. Larosa JA, Ahmad N, Feinberg M, Shah M, Dibrienza R, Studer S. The use of an early alert system to improve compliance with sepsis bundles and to assess impact on mortality. *Crit Care Res Pract.* 2012;2012:980369.

Quantifying the effect of weight reduction on progression to type II diabetes in high risk patients

Jesper Havsol (PhD)[1], Martin Karpefors(PhD)[1], Thérèse Rosklint (PhD)[2] and Cecilia Karlsson (MD, PhD)[3]

1) Advanced Analytics Center, Biometrics and Information Science, AstraZeneca R&D
2) Research and development Information (RDI), AstraZeneca R&D
3) Translational Medicine Unit CVMD, Early Clinical Development, AstraZeneca R&D

Pepparedsleden 1, SE-431 83 Mölndal, Sweden

It is well known that obesity is an important risk factor for type II diabetes. It has been shown for patients in a lifestyle modification program that weight loss is the most important predictor for reducing diabetes incidence [1]. Here we collect all relevant literature aiming to quantify the impact of weight loss on the risk of progressing to diabetes in patients with high risk for diabetes including different interventions and population types. We identified clinical trials on overweight or obese subjects with increased diabetes risk or pre-diabetes, treated with obesity drugs, lifestyle interventions or undergoing bariatric surgery. The material includes patients with an increased diabetes risk based on impaired glucose tolerance (IGT), impaired fasting glucose (IFG) as well as the FINDRISC scoring. Each study involved a weight loss phase of variable length followed by a follow up period of 1-4 years of changes in diabetic status. These criteria were met for 10 published clinical trials involving 9,811 overweight or obese subjects [2-11].

We extracted meta information on trial and population characteristics, as well as follow-up data on weight-loss and incidence of diabetes for both active treatment and internal controls. Exponential and linear models, to quantify the effect of weight reduction on diabetes incidence reductions compared with control, were fitted to the data.

Our results indicate that the association of diabetic risk was approximately linear for weight losses ranging from 0 to 10 kg, with an average risk reduction in the active arm of 8-13% compared with control for 1 kg average body weight reduction (corrected for weight loss in control group).

In conclusion, in this work we have gathered all available literature and quantified the risk reduction in progressing to diabetes after losing weight.

1. Hamman RF, Wing RR, Edelstein SL, Lachin JM, Bray GA, Delahanty L, Hoskin M, Kriska AM, Mayer-Davis EJ, Pi-Sunyer X. et al. Diabetes Care. 2006;11:2102–2107. doi: 10.2337/dc06-0560.
2. Knowler WC, Barrett-Connor E, Fowler SE, Hamman RF, Lachin JM, Walker EA, Nathan DM. N Engl J Med. 2002;11:393–403.
3. Katula JA, Vitolins MZ, Rosenberger EL, et al. Diabetes Care. 2011;34:1451–7.
4. Tuomilehto J, Lindström J, Eriksson JG, et al. N Eng J Med. 2001;344(18):1343–50.
5. Kosaka,K., Noda,M.;Kuzuya. 2005 Feb;67(2):152-62.
6. Penn L, White M, Oldroyd J, Walker M, Alberti KG, Mathers JC. BMC Public Health. 2009 Sep 16;9:342. doi: 10.1186/1471-2458-9-342.
7. Pontiroli AE, Folli F, Paganelli M, et al. Diabetes Care. 2005 Nov;28(11):2703-9.
8. Sjöström CD, Peltonen M, Wedel H. et al. Hypertension. 2000 Jul;36(1):20-5.
9. Torgerson JS, Hauptman J, Boldrin MN et al. Diabetes Care. 2004 Jan;27(1):155-61.
10. Vermunt PW, Milder IE, Wielaard F et al. Diabet Med. 2012 Aug;29(8):e223-31. doi: 10.1111/j.1464-5491.2012.03648.x.
11. Saaristo T, Moilanen L, Korpi-Hyövälti E et al. Diabetes Care. 2010 Oct;33(10):2146-51. doi: 10.2337/dc10-0410. Epub 2010 Jul 27.

Thermia: Simplifying Childhood Fevers with Mobile Decision Support

Jared B. Hawkins[1,2], PhD, Jane E. Huston[2], MPH, Florence T. Bourgeois[2], MD, MPH, John S. Brownstein[1,2], PhD

[1]Harvard Medical School, Boston, MA, [2]Boston Children's Hospital, Boston, MA

Summary

We developed a decision support tool and mobile application to provide information on fever and febrile illnesses, such as respiratory and gastrointestinal illnesses, to concerned parents. The primary goal is to provide information on supportive treatment that parents can provide at home and guidance on when to seek medical care via web-based and smartphone apps. A secondary goal is to integrate user data into public health biosurveillance tools (e.g., HeathMap.org and FluNearYou.org), to better track emerging disease and provide tailored treatment recommendations. Our tool will motivate parents to take a direct role in their child's health care by providing education and information necessary to support their decision-making around children with fevers and infectious febrile illnesses.

Introduction

Over 2 million children die each year due to infectious febrile illnesses [1]. In addition, medical costs are staggering; in the US, influenza treatment alone for children is ~$1.7 billion annually [2], with emergency visits for children under 5 years of age accounting for as much as $289 million [3]. Active engagement of patients and their families has the potential to lower medical costs through improved patient literacy and resource utilization, and may ultimately lead to more favorable health outcomes. We believe that by educating parents and providing at-home decision support, we can enable parents to provide supportive care when appropriate and optimize efficient use of healthcare resources, thereby reducing unnecessary medical costs.

Methods

We developed a mobile decision support tool, named Thermia (http://thermia.io), to provide at-home treatment recommendations for fever and infectious febrile illness symptoms based on established guidelines from Boston Children's Hospital. Thermia is a rule-based web framework that engages parents to answer questions regarding symptoms associated with a febrile illness. The user then receives information on possible causes for the symptoms, signs of serious illness, home treatment recommendations, and guidance on when to seek medical care.

Results and Discussion

Researchers have previously shown that at-home decision support for streptococcal pharyngitis is a viable approach to triaging patients and significantly reducing unnecessary office visits [4]. We are now utilizing this approach for febrile illnesses - specifically respiratory and gastrointestinal disease - which has not previously been done before. The Thermia website soft-launched in January 2014, and has had modest use to date. The majority of users are seeking treatment recommendations for children less than 5 years of age with a fever and most common symptoms of cough and/or sore throat. We are currently working to iterate and improve upon our decision support framework, with the goal of making the user experience as intuitive and informative as possible; to this end; we are working with popular parenting online communities. Our next step is to integrate user data into public health biosurveillance tools, such as HealthMap.org and FluNearYou.org, which have been extremely successful in using crowd-sourced data to track infectious disease outbreak and enable real-time monitoring. Integrating user symptoms from our application will strengthen these public health systems with the addition of a new, previously untapped data stream containing emerging symptoms at a local, regional and national level. This will allow us to inform users of emerging respiratory diseases via our application, and recommend preventative and therapeutic measures in the case of a local outbreak. Our poster presentation will discuss our progress to date on these endeavors.

References

1. Evaluation IfHMa. Global Burden of Disease Study 2010 (GBD 2010). Institute for Health Metrics and Evaluation (IHME), 2013.
2. Molinari NA, Ortega-Sanchez IR, Messonnier ML, Thompson WW, Wortley PM, Weintraub E, et al. The annual impact of seasonal influenza in the US: measuring disease burden and costs. Vaccine. 2007 Jun 28;25(27):5086-96. PubMed PMID: 17544181.
3. Fairbrother G, Cassedy A, Ortega-Sanchez IR, Szilagyi PG, Edwards KM, Molinari NA, et al. High costs of influenza: Direct medical costs of influenza disease in young children. Vaccine. 2010 Jul 12;28(31):4913-9. PubMed PMID: 20576536.
4. Fine AM, Nizet V, Mandl KD. Participatory medicine: A home score for streptococcal pharyngitis enabled by real-time biosurveillance: a cohort study. Annals of internal medicine. 2013 Nov 5;159(9):577-83. PubMed PMID: 24189592.

Streamlining Health Information Exchange Workflows through Automated Disclosure Control Services

Shan He, PhD[1], Darren K. Mann[1], Lance Schildknecht[1], Sidney N. Thornton, PhD[1, 2]
[1]Intermountain Healthcare, Salt Lake City, Utah; [2]University of Utah, Salt Lake City, Utah

Abstract

Patient consent remains an unintegrated and less automated component in the current health information exchange (HIE) infrastructure. Clinical workflows have been interrupted to procure administrative artifacts due to providers' fears and concerns associated with the potential misuse of protected health information (PHI). An automated solution that facilitates the reuse of existing data to enable compliant PHI disclosure is critical to streamline the HIE workflow while protecting providers from liability risks and preserving trust with patients.

Introduction

Although there are various specifications and services being developed to promote HIE, patient consent, which is central to the issue of privacy, remains unintegrated in the HIE infrastructure. For example, the Direct Project assumes that the sender is responsible for the collection of patient consent where appropriate, which is referred to as "out of band" verification[1]. Currently Intermountain Healthcare requires patient consent on file before disclosing PHI electronically to external providers. The collection of patient consent is time consuming and an interruption of the clinical workflow, which potentially limit the amount and type patient data available via HIE[2]. To alleviate the administrative burdens while addressing the privacy concern by preventing inappropriate disclosures, we are implementing a self-contained disclosure control service to automate the regulation-compliant PHI disclosures.

Methods

The automated disclosure control (ADC) service generates and enforces a set of rules for each PHI disclosing transaction in the format of when, what, to whom, for what purpose. These rules can be generated on the fly using existing data from various administrative and clinical sources during an encounter such as 1) the consent for treatment ; 2) current and previous care team participants; 3) Current episodic treatment/transition plans; 4) payer HIE stipulations. The key functionality in the ADC service is to automatically infer the treatment or payment relationship between the recipient of the disclosure and the patient based on the above information. When such relationships cannot be inferred, the clinician is required to provide supplementary attestation to justify the disclosure. Furthermore, the current proactive mechanisms for detecting potentially inappropriate EHR access can be reused to preempt suspicious disclosures. Finally, all disclosure details are audited by the ADC service. Even though the ADC service is designed to alleviate the need to collect patient consent, it is extendable to incorporate with structured patient consent to enforce the disclosure rules for situations when patient consent is collected (e.g. opt-out patients) or required (e.g. patient records maintained by federally assisted alcohol and drug abuse program).

Results

Regulation-compliant PHI disclosure via HIE is enabled by the ADC service when a valid treatment or payment relationship can be automatically inferred. The ADC service streamlines the clinical workflow by eliminating interruption of the care delivery process with unnecessary administrative tasks for the clinician. The ADC disclosure auditing provides transparency for patients and privacy advocacy for the healthcare providers.

Conclusions

A trade-off exists between the level of privacy assurance and the convenience of HIE participants. Leveraging automated services that minimize the administrative burden while maximizing privacy protection is critical for successful HIE adoption and sustainability.

References

1. The Direct Project Overview. 2010 October [cited 2014 Mar 2] Available from: http://wiki.directproject.org/file/view/DirectProjectOverview.pdf
2. Goldstein MM, Rein AL. Consumer consent options for electronic health information exchange: policy considerations and analysis. 2010 March [cited 2014 Mar 2]. Available from: http://www.healthit.gov/sites/default/files/privacy-security/choice-model-final032610.pdf

A Qualitative Preliminary Study of Older Adults' Personal Health Information Tracking Behaviors

Yuqi He, MLIS
University of Wisconsin-Madison, Madison, WI

Abstract
Personal Health Records (PHRs) create great opportunities for older adults to track and manage their health; individual interviews with people age over 60 were conducted to gather information about their personal health information tracking behaviors.

Introduction
The current growth in the number and the proportion of older adults in the United States is unprecedented. By 2050, the over-65 population in the U.S. will increase to 88.5 million, representing one in five Americans.[1] 86% of older adults have at least one chronic disease, and 56% have two or more.[2] The aging population with high proportions of chronic conditions will increase demands on medical services and poses great challenges to our already strained health care system.

PHRs can be potential technologies to help track and manage health. PHRs may be a particularly useful tool for older adults because of increased occurrence of chronic diseases and the need for long-term care. Several applications, such as Project HealthDesign's Colorado Care Tablet (CCT),[3] were designed to help older adults manage medications. We lack knowledge, however, about individual factors of why and how older adults keep track of their health information, how they use this information, their perceived benefits, and under what conditions they decide to continue or stop tracking. To help design information technologies that promote health tracking by older adults, I conducted individual interviews to explore their personal health information tracking behaviors.

Methods
Ten subjects (4 males and 6 females) aged 60 or older participated in this study. Their educational levels ranged from non-completed high school to PhD degrees. Seven of them were White. Face-to-face, semi-structured interviews were conducted. Each interview lasted an average of 40 minutes. The sessions were recorded using a digital recorder. All interviews were transcribed verbatim. The transcriptions were analyzed using Nvivo 10. A grounded theory approach was used to analyze the data and develop a model of older adults' health information tracking process.

Results
Five components consisted of this behavior model: motivation, tracking & managing process, outcomes (information sharing and perceived benefits), decision (continue or stop), and social contexts (doctors and families). Motivations prompt older adults' tracking needs, which in turn initiate the tracking process. Three themes were identified: goal-oriented (e.g. chronic disease management, medication management, and keeping fit), beliefs and values (e.g. informed health consumer, self-responsibility), and health-related experiences (e.g. health event, family health history, age-related memory loss). Older adults used pillboxes, paper files, and electronic gadgets such as PHRs as tracking tools. Six perceived benefits emerged: improved accuracy, keeping conditions in control, medication adherence, good communication with doctors, self-evaluation, and positive emotional reinforcement.

Conclusion
Tracking behavior is a complicated, iterative process. Older adults become a more diverse population in terms of their physical capacity, psychological performance, and life style. Every subject's tracking behavior is unique, in terms of their motivations, tools use, and perceived outcomes. This implies that designing information technologies that try to capture all behavioral actions is inherently complex and practically impossible. This complexity may suggest a layered strategy for IT designers to design a personalized, customizable product for older adults.

References

1. Vincent GK, Velkoff VA. The next four decades-the older population in the United States: 2010 to 2050. Current Population Reports, P25-1138; US Census Bureau: 2010.
2. Centers for Disease Control and Prevention. Percent of U.S. adults 55 and over with chronic conditions. 2009. Available from: http://www.cdc.gov/nchs/data/health_policy/adult_chronic_conditions.pdf.
3. Siek KA, Ross SE, Khan DU, Haverhals LM, Cali SR, Meyers J. Colorado Care Tablet: The design of an interoperable Personal Health Application to help older adults with multimorbidity manage their medications. J Biomed Inform. 2010 Oct; 43(5, Supplement):S22–S26.

Admission data predict risk as well as discharge data in patients with pneumonia: A readmission risk-model evaluation

Courtney L. Hebert, MD, MS[1], Peter J. Embi, MD, MS[1]
[1]The Ohio State University, Columbus, OH

Abstract

Real-time risk prediction using continuously updated electronic health record (EHR) data could help to target timely and personalized interventions. Few studies have examined how the change in patient-specific EHR variables on different days of a hospitalization might affect risk prediction. Using a previously validated readmission risk score on a cohort of hospitalized patients, this study describes how the score changes over time, and on which day the score is the most predictive of readmission.

Methods

The risk model used in this study was originally developed by our group to predict readmission in hospitalized patients with pneumonia on the day of discharge from the hospital. It was created using logistic regression with stepwise removal and performed well on a validation cohort (area under the receiver operating curve (AUC) 0.71). In addition to a history of prior admission in 30 days, variables in the final model included ones that could change over a hospitalization such as lab values, medications, number of medications and comorbidities. The model categorizes patients into high, medium or low risk based on their risk score.

This study applies the risk-model to a new cohort of patients admitted from 5/2012 to 1/2013 with a primary discharge diagnosis of pneumonia. Variable definitions were adapted in order to reflect risk per day of hospitalization. For example, the problem list was used to define comorbidities instead of discharge diagnosis billing codes (these only being available after discharge). The change in risk score for individual patients as well as the change in risk score over time was evaluated. An ROC test of equality was used to determine if change in AUC from different days of hospitalization was statistically significant.

Results

356 of 418 (85%) patients had a calculable risk-score (no missing values) on the day of admission. 274 (77.0%) patients had an increase in their score from the day of admission to the day of discharge and 72 (20.2%) patients had a decrease. Only 18 (5.0%) patients went from high or medium risk on the day of admission to low risk by discharge. 241 (67.7%) patients had no change in their risk category. The predictive model performed less well on this cohort with an AUC of 0.60 on the day of discharge. **Figure 1** shows the change in AUC over hospitalization. Overall, the AUCs of the model on different days were not statistically different with a p-value of 0.10.

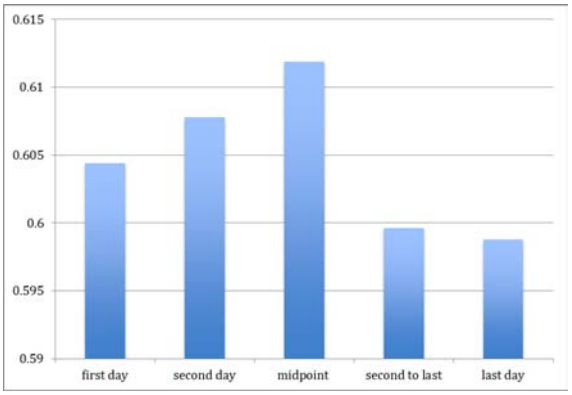


Figure 1: Change in AUC of model at different points during hospitalization

Conclusion

The predictive ability of the model varied minimally over the hospitalization, with the best performance at the midpoint of the stay. This suggests that, for this model, risk prediction earlier in the hospital stay would be reflective of the risk at discharge allowing for earlier interventions.

Design and Development of Team Builder –

Matching Funding Opportunities to Research Profiles

Andrew C. Helsley, BS[1,2], Robert A. Dennis, PhD[1,2], Marianne Zachariah, BA[1,2], and Douglas S. Bell, MD, PhD[1,2]

([1]Department of Medicine, David Geffen School of Medicine at UCLA, Los Angeles, CA, [2]UCLA, Clinical and Translational Sciences Institute, Los Angeles, CA)

The UCLA CTSI has developed a web-based open source system that mines NIH funding opportunity announcements (FOAs), processes salient passages of them with the National Library of Medicine's natural language processing tools to produce Medical Subject Heading (MeSH) terms that are then used to semi-automatically search against investigator web research profiles. We call this system the Team Builder, and we have engineered the basic components to address two similar use cases. The first begins with a funding opportunity announcement and the need to identify qualified investigators within our cross-institutional clinical translational science consortium. The second use case takes the perspective of an individual investigator who wishes to stay aware of possible funding opportunities that closely align with his/her area(s) of research. In both cases, the challenges are to match expertise as represented by a web-based research profile with a funding opportunity represented by its announcement.

Our poster will present the design and functionality of the Team Builder in the context of our initial work focused on the first use case. We have developed this system to leverage the rich data found in public research profiles with particular attention on MeSH keywords relating to publications. We have implemented a framework that processes NIH FOAs and combines them with the automatic classification capabilities of the Medical Text Indexer (MTI) to help improve the recall of staff research facilitators whose role is to invite qualified individuals to collaboration brainstorming session.

The central component of the Team Builder engine is the Unified Medical Language System (ULMS) and more specifically the Medical Subject Headings (MeSH. The MeSH terms are attached to publications and transitively to the authors of the publications. Team Builder associates MeSH terms with FOAs using the MTI. The result is a well-curated semantically meaningful common basis for matching the two together. Our initial user interface (UI) has been tailored with the recognition that pursuing a funding opportunity requires the participation of individuals in several different roles, and the raw output of MTI may not be completely satisfactory. Our UI allows staff to partition and augment the set of MeSH terms that are automatically associated with the FOA into smaller sets that are expected to improve the precision of the Team Builder beyond a simple fully automated approach.

The UCLA CTSI Team Builder has been developed using the open source framework called OpenACS. It is an advanced toolkit for building scalable, community-oriented web applications and services. OpenACS runs on OpenNSD or AOLserver, open source multi-threaded and high performance HTTP servers with an embedded TCL scripting language.

After reading our poster and ideally speaking directly with one of the co-authors, the learner will be better able to:

- Formulate a plan to implement a similar approach to making meaningful use of his/her institutions research profiles system and publically available NLM tools to match qualified investigators with high-value funding opportunities
- Decide whether to adopt our approach and software and request access to our source code to implement at his/her own institution

Use of an Iterative Search Strategy in Critical Care Informatics

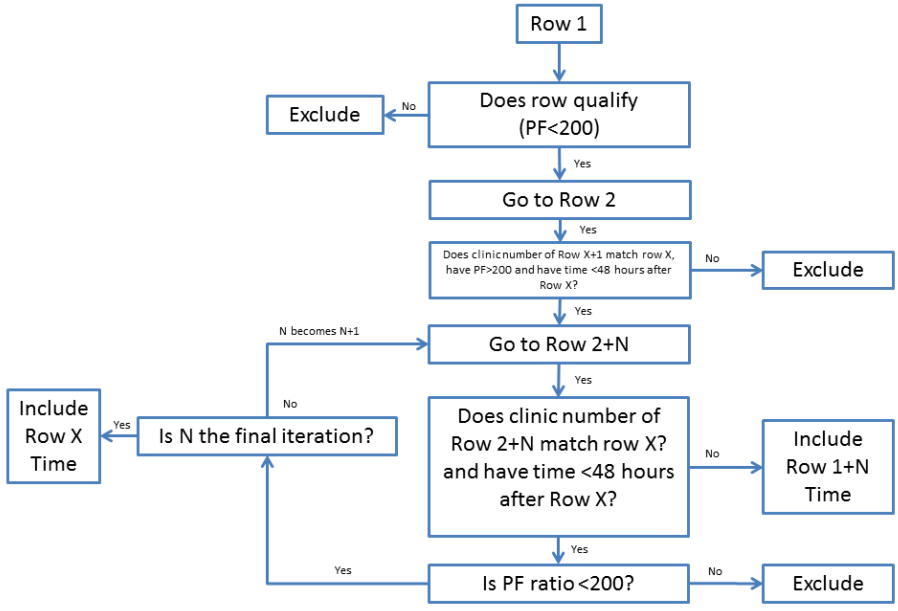
S. Kaur MD, LY. Garcia Arguello MD, JC. O'Horo MD, O. Gajic, MD, V. Herasevich MD, PhD, B. Pickering MD, R. Kashyap, MBBS
Multidisciplinary Epidemiology and Translational Research in Intensive Care (METRIC), Mayo Clinic, Rochester, MN

Rationale

Databases are excellent tools capturing a variety of patient data, but pulling trends is challenging. A change in hemoglobin over 24 hours or an increase in oxygen requirement may well be more significant than an isolated value. However, standard statistical software, like JMP, stores these values on separate rows, making aggregation or trending difficult. We attempted to use an iterative search strategy of a database of critical care patients to identify resolution time for patients with acute respiratory distress syndrome (ARDS).

Methods

All analysis was done with JMP statistical software. Cases of ARDS were identified using American European Consensus Conference (AECC) guidelines, corresponding to "moderate or severe" ARDS category of the New Berlin definition. ARDS resolution time was defined by the first P/F ratio >200 which is sustained >48 hours (i.e. no subsequent drops in that interval <200). All blood gas results (Pao2) and inspired oxygen (Fio2) values were obtained from the database, and combined to create P/F ratios for each lab draw interval. Subjects were sorted by medical record number and date of onset (earliest to later), such that all results were listed sequentially for each patient, where the number of distinct records for patient ranged from 2 -107 rows. A multi-iteration search was designed as shown in the figure below. Analyses were run for 1, 5, 10 and 20 iterations. This was then compared against a manual chart review of the same patients to determine the diagnostic performance of the tool.



Results

Each iteration had improved specificity, with no additional cases excluded between the 10th and 20th iteration. When compared against manual chart review, the 10 iteration model had excellent diagnostic performance at 96% sensitivity and 95% specificity.

Figure: Iterative search strategy logic. Iterations are run with N=1,2,3… until a final iteration.

Conclusions

An iterative search strategy is a viable approach to tracking variables change over time in critical care databases and has potential to speed up research by saving man power and minimizing errors related with fatigue.

Patient Factors Associated With Provider Response to a Weight Management Best Practice Alert

Annemarie Hirsch, PhD, MPH[1]; Lisa Bailey-Davis, DEd[1]; Craig Wood, MS[2]; Christopher Still, DO[3] Geisinger Center for Health Research[1], Weis Center for Research[2], Geisinger Obesity Institute[3], Danville, PA

Abstract

There has been little exploration regarding what patient factors are associated with provider acceptance of clinical decision support (CDS). We conducted a study using electronic health record (EHR) data to examine patient factors associated with referrals to weight management in response to a weight management best practice alert. BMI, age, race, gender, co-morbidities, and history of weight loss were associated with providers' response to the alert. When developing CDS tools, patient factors should be considered.

Introduction

To qualify for meaningful use (MU) incentives, hospitals and eligible providers are required to implement at least one clinical decision support (CDS) rule. Studies report that providers override up to 90 percent of CDS alerts.[1] Most of the research on providers' response to CDS tools has focused on characteristics of the alert itself, with little exploration of patient factors.[2] We evaluated what patient factors are associated with providers' response to a weight management best practice alert (BPA) in a primary care setting.

Methods

We conducted a retrospective study using EHR data to examine patient factors associated with referrals to weight management treatment in response to a weight management BPA. The alert fired when a patient with a body mass index (BMI) greater than 25kg/m^2 had an office encounter. The alert notified providers of the patient's weight status and allowed physicians to make a referral to weight management. We confined our analysis to alerts that fired during primary care encounters between March 2011 and January 2012. We used multiple logistics regression to identify patient characteristics that were associated with a referral that was ordered in response to the BPA.

Results

The BPA fired for 111,444 patients and 2004 (1.8%) had a referral to weight management. Alerts fired from 1 to 40 times for patients, for a total of 249, 508 fires. Patients had a mean age of 50.1 years; 52% were female; and 97% were white. Higher referral rates were associated with lower BMI, younger age, female gender, non-white race, and recent weight loss ($p<0.0001$ for all). In patients with BMI 25-35, those with a diagnosis of diabetes were more likely to get a referral ($p<0.0001$).

Conclusions

Patient factors were associated with how providers responded to the weight management BPA. When developing CDS tools, it is important to consider patient factors that will facilitate or inhibit provider acceptance of the tool. With a steady increase in the adoption of EHRs and the MU CDS requirement, we can expect a dramatic increase in the implementation of CDS tools, making it critical to increase our understanding of factors associated with the use of these tools.[3]

References

1. Weingard SN, Toth M, Sands DZ, Aronson MD, Davis RB, Phillips RS. Physicians' decisions to override computerized drug alerts in primary care. Arch Internal Med. 2003; 163: 2625-2631.
2. Seidling HM, Phasalkar S, Seger DL, Marilyn PD, Shaykevich S, Haefeli WE, Bates DW. Factors influencing alert acceptance: a novel approach for predicting the success of clinical decision support. JAMIA. 2011; 18: 479-484.
3. The ONC. Update on the adoption of health information technology. June 2013. [Cited 2014 March]. Available from: http://www.healthit.gov/sites/default/files/rtc_adoption_of_healthit_and_relatedefforts.pdf

Design and Evaluation of a Glomerular Disease Ontology

Jamie S. Hirsch, MD, MSB,[1,2] Chunhua Weng, PhD[1]
[1]Department of Biomedical Informatics, Columbia University, New York;
[2]Division of Nephrology, Department of Medicine, Columbia University, New York

Abstract

Glomerular diseases have been defined and classified by their pathologic findings, but many glomerular diseases share pathologic features. The existing taxonomies for glomerular diseases contain ambiguity and can lead to errors when using these disease definitions. The Glomerular Disease Ontology (GDO) was developed to formalize the knowledge of this domain, clarifying disease entities and their relationships to etiology, pathology, clinical manifestations, laboratory data, and treatments. The current version of GDO contains 247 classes, 43 object properties, and 664 class relationships. All concepts were encoded using the Unified Medical Language System (UMLS). The GDO has the potential to provide diagnosis and treatment decision support for glomerular diseases.

Introduction

Glomerular diseases suffer from taxonomic complexities due to the conflation of kidney biopsy pathology and disease definitions. This class of diseases is often referred to by their pathologic descriptors, which yet do not imply pathophysiology or treatment. This ambiguity has negative ramifications for clinical care and research coordination. No glomerular disease-specific ontology exists, while current ontologies rely on traditional pathologic descriptors as disease entities, preserving the conflation between pathology and disease. The goal of this research is to create a domain-specific Glomerular Disease Ontology (GDO), solidifying clear disease definitions with linkages to pathologic findings, clinical presentations, and treatments, in order to facilitate clinical decision support.

Methods

Core competency questions were created to establish the ontology's scope. Domain-specific knowledge and entities were gathered from the nephrology literature and international professional society guidelines and reports. Existing related ontologies were reviewed to identify reusable disease classes. Knowledge representation was performed using Protégé version 4.3.2 in the Web Ontology Language (OWL). The ontology's class and logical consistency was evaluated with FaCT++, an automated description logic reasoner within Protégé. Expert evaluations were undertaken to assess the domain accuracy, core competency, relevance, and usefulness.

Results

The ontology currently defines 247 classes, including 105 disease-based classes, with 75 of which currently mapped to UMLS. Most non-glomerular diseases and seven glomerular diseases were imported from the Disease Ontology. Eighty-one classes have an annotated definition and 11 glomerular classes are presently defined. The "glomerular disease" class has 7 direct children and total 55 unique classes (see Figure). It currently maps most glomerular diseases to basic clinical syndromes, systemic disease associations, histologic manifestations, and clinical guideline treatments.

Figure: Hierarchical structure of glomerular diseases within the GDO

Conclusions

The GDO represents the first step toward formalizing the taxonomy of glomerular disease. Missing in the current version are rarer glomerular diseases, ultrastructural pathologic findings, clinical manifestations, laboratory findings, classification and prognostication tools, and granularity in treatment protocols. The GDO will facilitate improvements in clinician education, clinical workflow, and research organization in the field of glomerular disease.

Pictogram-based tablet application enabling patient input of emotions and behaviors

Teruyoshi Hishiki, MD, PhD[1], Hiroki Yasui, MD, PhD[2], Yoshiko Matsunaga, RN, PhD[1], Megumi Itoshima, BPh[3], Keiko Abe, RN, PhD[2], Takahiko Norose, MBA[4], Takuro Tamura[5]
[1]Toho University, Tokyo, Japan; [2]Nagoya University, Nagoya, Japan; [3]Shimokawa Pharmacy Ltd, Kumamoto, Japan; [4]Hokkaido Pharmaceutical University, Otaru, Japan; [5]LINE Co., Ltd, Tokyo, Japan

Abstract

We developed and are testing a tablet-based Web application in which patients can record their actions and emotional states as they occur. The resulting data can be shared with healthcare professionals. On the touch-screen, patients identify what to record via pictograms, which are categorized into perceived well-being, daily activities, compliance with instructions, and self-measured physical data. The application could also be used to evaluate the patient's awareness of their disease and evaluate patient–provider communication.

Introduction

At patient visits, it is useful for medical providers to gain an understanding of the patient's emotional and behavioral wellbeing over the period since the last visit, but this is difficult to obtain because of time pressure. Self-recording by the patient offers an alternative approach, and its effectiveness in chronic disease management is supported by the findings of past studies; however, adherence to this method has previously been problematic. One solution is for patients to carry a device in which they can easily record their feelings and actions at the time that they occur.

Problem to Address

We aimed to develop a touch-screen-based user interface for a portable device that would enable patients to record their feelings and actions.

The Pictogram Approach

After discussion with medical providers about the type of between-visit information that would generally be valuable, four categories were identified: (a) perceived well-being, (b) daily activities, (c) compliance with instructions, and (d) self-measured body weight and blood pressure. These categories were then expanded to a total of 12 types of information.

We developed a Web-based application for seven-inch tablet devices, on which the home page displays a panel with a pictogram for each of the 12 types of information. The user has only to touch the relevant pictogram, and in most cases this is all that is required. Some pictograms, such as "how well I am feeling" and "blood pressure" have heat map scales or slide bar scales for input. Information entered by the patients is time-stamped and stored on a server. Patients can browse their information in time order or by pictogram type, and their healthcare providers can log on to browse the same information. A healthcare provider can append a text message to the patient's input history.

Current Status

We are presently analyzing the logs of seven male patients (age range, 36–82 years; median age, 66 years) and their healthcare providers who have completed a 4-month test of the application. Two of the patients were employed, and the others were at home, while all had stable health during the test. The patients spontaneously entered 1–14 inputs per day at the beginning of the test; those who inputted frequently from the start continued this pace to the end. Pictograms for "had a meal", "feeling good", "took my medication", "measured my blood pressure", and "took a bath" were the most frequently used. Some of the healthcare providers have reported that the recorded data enable a better understanding of the patients' lifestyle.

Conclusion

The pictogram method provides a simplified input method and is a tool for identifying patients' scope of attention, which may also reflect their awareness of their disease. Factors related to the patient and to the patient–provider relationship may provide insight into the patterns of usage.

Point of entry notification of shortages using a drug auditing program

S. Paul Hmiel MD, PhD[1], Charles H Andrus, MHA [2], Kevin O'Bryan MD[2], Phillip Asaro MD[1], Feliciano Yu, MD, MSPH [1]

[1]Washington University School of Medicine, St. Louis MO; [2]St. Louis Children's Hospital, St. Louis MO

Abstract

Communicating critical drug shortages to clinicians has become a major challenge, but opportunities exist within computerized clinician order entry systems to provide appropriate alerts at the time of order entry. We added drug shortage alerting to a dose range checking and auditing system, which was previously developed as an adjunct CDS system. We added rules representing medication shortages along with degree of availability, such as unrestricted use, in shortage, and unavailable. Analysis of alerts by medication and care unit facilitated education by clinical pharmacists. Regular monitoring and user feedback identified unnecessarily repetitive alerts. Shortage status rules were subsequently adjusted to alert only once per user, per drug, per week. Medication shortage alerts require careful design, as well as providing easily updated information to be effective.

Introduction

Communication of critical drug shortages to clinicians has become an ongoing challenge for hospital pharmacy services, due to rapid changes in availability, and difficulties in identifying the appropriate clinicians[1]. The most common approach utilizes combinations of paper-based, electronic communications, and/or direct conversations, to inform the affected clinicians, as not one approach is satisfactory by itself. Computerized clinician order entry systems within the electronic medical record allow alerting at the point of entry, which would be expected to be the most efficient, with the appropriate information delivered directly to the intended ordering clinician(s), without relying on clinician memory/awareness. To be maximally effective, however, the information needs to be accurate, up to date, easily updated, with minimal alert fatigue and annoyance.

Methods

A dose range checking and auditing system was previously developed as an adjunct clinical decision support system to provide more comprehensive information to clinicians, expanding on the vendor supplied dose range checking features of the electronic medical record (EMR). Additional rules were developed to represent medication shortage and degree of availability, such as unrestricted use, in shortage, and unavailable. In addition, a dashboard was developed to monitor number and types of alerts in near real time, over varying time intervals. When the rule was applied to an order, it could result in any of 3 alert types: warning (user can proceed without a comment), soft stop (user must comment to proceed), hard stop (user cannot proceed). A recent upgrade to the EMR provides the ability to link an alert directly to an alternative medication.

Results

The Pediatric and Cardiac Intensive care units accounted for just over half of the total shortage alerts, at 31.3 % and 21.5%, respectively, of the 6110 alerts analyzed. This is reflected by the most frequently alerted agents, with 792 alerts for calcium gluconate, 457 for bumetanide (bolus or infusion), and 308 for sodium bicarbonate, with these three agents used primarily in the intensive care units. Regular monitoring of the dashboard rapidly allowed analysts to identify nuisance drug shortage alerts, which was noted especially for frequently used agents, as the frequency of alert acknowledgement rapidly decreased. Consequentially, shortage alerts went from changing practice only 21.3% of the time to 40.7% of the time. The rules for shortage status were subsequently adjusted to only alert only once per user, per drug, per week, with marked increase in clinician satisfaction (decreased annoyance).

Conclusion

Point of entry notification of drug shortages provides an additional route of communication that is focused on providers, when it is most useful, without the disadvantages of more traditional communication tools. Attention to the design and frequency of alerts is essential to user acceptance however, while ease of maintenance is essential to capture rapidly evolving supply situations. The ability to redirect clinicians for a medication in shortage to an acceptable alternative will provide additional strategies to manage shortages.

Efficient Translation of EHR Free-Text Data to Coded Data PRN

Chad M. Hodge, MS[1, 2], Trilok Prithvi, MS[1, 2], Naveen Maram, MD[1], Kathryn G. Kuttler, PhD[1], Scott P. Narus, PhD[1, 2]
[1]Intermountain Healthcare, Salt Lake City, UT; [2] University of Utah, Salt Lake City, UT

Abstract
When developing a new clinical system, Intermountain Healthcare (IH) recognized an opportunity to try new means of improving and maintaining the underlying terminology. Implementing the new process and tool has increased clinician engagement, reduced uncoded patient data, and helped meet Meaningful Use goals.

Introduction
Intermountain Healthcare (IH) works with clinicians prior to implementation of vocabularies to define, pre-coordinate, and load only the concepts and representations that are clinically useful. Post-implementation, clinicians require new content, and request synonyms for existing content, allowing for efficient searching. When terminology is incomplete, clinicians revert to adding patient data as free-text rather than as coded concepts. Current processes for handling free-text data requests are ad-hoc, and entail many months of effort. This can leave clinicians disengaged from content governance, and patient data remains uncoded and thus unavailable to decision support. To address these issues, IH developed a new terminology feedback process to engage clinicians. Novel and efficient mechanisms were developed to identify gaps and allow clinicians to interactively review and approve recommended content in the context of patient care.

Methods
During a six-month period, hospitalists, providers at seven NICUs and one STICU, and TeleHealth services had access to the new feedback tool, made available initially through our EMR's problem management module. When clinicians added a free-text problem, a structured request was automatically routed to the terminology work queue. Clinical modeling engineers (CME) then inspected the free-text term, and searched for and evaluated matches for the term in the existing dictionary to determine if the term was misspelled, an acronym, a synonym, or missing. CMEs then added the new content or created mappings between the synonym/acronym and root concept, ensuring future searches return proper results. CMEs then created a list of candidate substitutions for the clinician's free-text problem, and sent that list back to the application for review by the clinician. The next time a clinician opened the same patient's problem list, the application presented the clinician with the substitution candidates for the free-text problem (Figure 1). The original free-text was viewed alongside the list of candidates. The clinician could select a single choice to serve as the replacement for the free-text problem, or could alternatively choose to reject all provided choices with an accompanying reason. If rejected, the process was performed one more time, resulting in a new set of candidates. If the clinician chose to accept one of the newly provided coded candidates, the free-text problem was automatically replaced by the coded concept. If rejected, the problem would remain free-text.

Results
The first three days the tool was available, over 1,200 free-text requests were created during normal clinician use of the problem list. The new process handled these requests daily. The old process added new concepts to the dictionary, *if* the missing content was discovered by the CMEs. This allowed future use of the term, but left previously entered patient problems as free-text. This new approach has permitted coded content to replace the original free-text problems, with clinician interaction. In fact, free-text problems remaining on problem lists dropped from 12% using the legacy system, down to 1% using the new process.

Conclusion
Allowing clinicians to choose coded substitutes of free-text problems in the workflow has reduced free-text problems remaining on lists, engaged clinicians, and reduced turnaround time to add coded content.

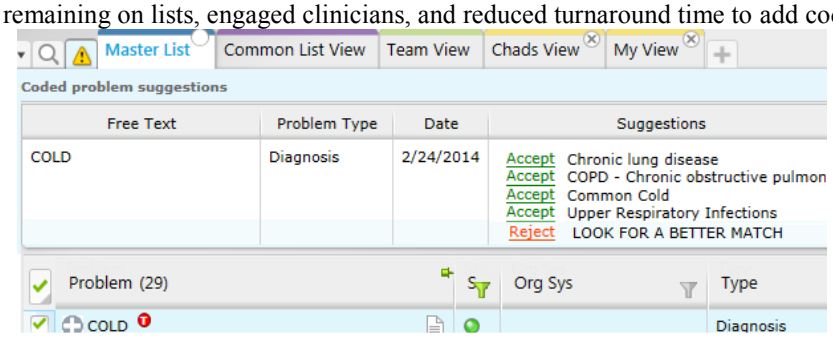


Figure 1 - Functionality in the Problem Management Module that allows a clinician to substitute a free-text problem with a coded problem, based on CME & clinical governance suggestions, or to reject candidates and request a new set of choices.

Reporting the Results of Safety-Enhanced Design Evaluations for Meaningful Use Stage 2 Certification: Are they Comparable?

Jan Horsky, PhD[1,2], Michael Swerdloff, BS[3]

[1]Brigham & Women's Hospital; [2]Harvard Medical School, [3]Partners HealthCare, Boston

Abstract

Effective, safe and routine use of HIT by clinicians is predicated on the availability of well-designed systems that have excellent usability characteristics and can be integrated into common clinical workflows The ONC has required, for the first time, that institutions and vendors developing EHRs submit results of summative usability evaluations as part of their application for Meaningful Use Stage 2 certification. The intent was to let developers show evidence of usability of their product so consumers could make informed purchase decisions. The expected content of each usability report was specified by the NIST in the Common Industry Format Template. However, a review of published test protocols shows wide variations in methods, content and study size. It may prove to be difficult for consumers to make effective comparisons across systems and make informed purchase decisions.

Description

Effective use of health information technology (HIT) and continuing progress in its adoption by clinicians are the objectives of national policy initiatives in many countries.[1] The rationale driving the implementation of electronic health record systems (EHR) and decision support is their potential for monitoring and improvement of care quality and the safety of patients. There is an emerging consensus among leaders in the industry, academia and government that a sustained positive effect of HIT can only be achieved by using systems that are specifically designed for the complex healthcare environment and that provide cognitive support, have high standards of usability, advanced interface design and are well integrated into clinical workflows.

The Office of the National Coordinator for HIT (ONC) has underscored the significance of high-quality design as a pre-requisite for high performance and safety of information systems in Stage 2 of its Meaningful Use Criteria for 2014. It has added a requirement to report the results of a summative usability evaluation and an attestation that development was done in accordance with a user-centered design process in order to receive certification and financial incentives. The intent was to help vendors and developers at academic and public institutions demonstrate evidence of EHR usability in a format that allows both independent evaluation and comparison across products.[2] The test results are made public on the ONC website to allow consumers and procurers at large institutions to review and assess basic usability characteristics of systems they may consider purchasing. Developers may also use the insights and lessons learned from the test process to further refine the design of their systems and to focus on problem areas that may not have been previously identified.

The format and required content of result reports was specified by a common industry format template[3] but methods and the extent of evaluations (sample size, scenario complexity) varied by system and institution. For example, the number of clinicians participating in observational studies varied from 3 to 30 (as of June, 2014). Test settings content of scenarios and the interpretation of results such as path deviations, type and severity of errors and task success or failure criteria almost certainly varied from one report to another and some criteria were not described in reports at all. The evaluators were sometimes usability professionals hired by vendors to run the study or experts already working for the institution but often just information system developers whose appropriate training or experience in usability evaluation was not reported. A comparison of reports published on the ONC Meaningful Use site shows wide variations in the number of clinicians tested, methods used to derive the results and completeness. The intended comparability of results across systems may be difficult to achieve.

References

1. Simborg DW, Detmer DE, Berner ES. The wave has finally broken: Now what? Journal of the American Medical Informatics Association. 2013 June 1, 2013;20(e1):e21-e5.
2. Lowry SZ, Quinn MT, Ramaiah M, Schumacher RM, Patterson ES, North R, et al. Technical Evaluation, Testing and Validation of the Usability of Electronic Health Records. 2012 NISTIR 7804.
3. Schumacher RM, Lowry SZ. Customized Common Industry Format Template for Electronic Health Record Usability Testing. Washington, D.C.: National Institute of Standards and Technology, 2010 NISTIR 7742.

A modular approach for Consolidating CCDs from multiple data sources

Masoud Hosseini, MSc[1,2], Jonathan Meade[3], Jamie Schnitzius[3], Brian E. Dixon, MPA, PhD[1,2,4]
[1] Indiana University School of Informatics and Computing, Department of BioHealth Informatics
[2] Regenstrief Institute, Center for Biomedical Informatics
[3] CreateIT Healthcare Solutions, Inc.
[4] Center for Health Information and Communication, Department of Veterans Affairs, Veterans Health Administration, Health Services Research and Development Service
Indianapolis, Indiana, USA

Abstract

Clinical information is fragmented across many different organizations and computer systems. To view complete, longitudinal medical records for a patient, clinical systems are starting to routinely, electronically request information from various sources in the form of a Continuity of Care Document (CCD). Given the heterogeneity and fragmentation of data across sources, systems are likely to receive multiple CCDs for a single patient. Current tools for viewing CCDs make it cumbersome for providers to explore different CCDs to find a specific data which can be duplicated or even conflicted. We have developed a modular approach for integrating and de-duplicating multiple CCDs into one consolidated document. Our system is designed to support nationwide and regional health information exchange efforts to support integration of standard clinical documents into clinical information systems for presentation to clinicians.

Introduction

The Continuity of Care Document (CCD) is an electronic document exchange standard for sharing patients' health information summary between providers and organizations which is required under Meaningful Use criteria. Given the fragmented and heterogeneous nature of clinical data, requests for complete medical records results in multiple CCDs from a variety of sources, however, care providers prefer to review consolidated information that represents a single, comprehensive picture of a patient's medical history and current condition, rather than multiple CCD documents that may include duplicate or conflicting information.

Methods

We designed a modular, open source system that aims to consolidate and de-duplicate received CCDs. The key component is the CCD Consolidation Engine, which executes a set of rules against a list of CCDs to produce a single, clinically useful CCD. Four types of rules are executed: 1) Pre-Format Rules examine incoming CCDs and inspect them for quality; 2) Primary Rules ensure the program returns valid information to the user; 3) De-duplication Rules merge, de-duplicate and handle conflicting information in the list of CCDs; and 4) Post-Format rules clean up the returned, consolidated and de-duplicated CCD. The system further audits all actions performed to comply with federal regulations. We further developed an application programming interface (API) to enable remote interaction through web services. The system is developed using the C# language and we used Microsoft .NET libraries to manipulate XML documents. Audit data is retained in a NOSQL database, such a Hadoop.

Results

A prototype was developed for a local health software competition, for which it took top honors. The prototype currently executes consolidation and deduplication rules against a de-identified reference dataset at a rate of approximately .009 to .03 seconds per rule depending on complexity. Evaluation of precision and recall using a larger dataset from an operational health information exchange is currently underway.

Conclusion

Given that meaningful use regulations and many HIE organizations are less than three years old, there do not yet exist a plethora of proven solutions for consolidating CCDs. Consolidation of CCDs, however, is challenging and requires processes that can interpret, merge, de-duplicate, and resolve conflicts across complex documents involving complex data types. Accuracy of the de-duplication and consolidation greatly depends on the logic used in the Consolidation Engine. The benefit of the audit model chosen for this project is that it can be used to analyze the logic and clinical usefulness of the output. Health informatics researchers and scientists will be able to improve the algorithms and methodologies used in the rule logic over time increasing the clinical usefulness of the CCD.

Evaluation of SNOMED CT Content Coverage for a Decision Support System

Rachael Howe BSN, RN, Tosh Kartchner RN, Kristen Humpherys BSN, RN, Tiffany Harman RN, Susan Matney MSN, RN, PhD (c), FAAN, Senthil K. Nachimuthu MD PhD
3M Health Information Systems, Inc., Murray, Utah

Introduction

The 3M Healthcare Data Dictionary (HDD) was used to standardize the clinical data used by a decision support system (DSS) which was used for clinical quality assurance and reimbursement. The HDD is a vocabulary server that stores standard and local terminologies. The goal of this project was to map the DSS concepts to Systemized Nomenclature of Medicine Clinical Terms (SNOMED CT) in order to analyze the EHR data. The DSS data included nine disease processes: chronic obstructive lung disease, pneumonia, acute coronary syndrome, heart failure, etc.

Methods

There were 3 phases of this project: analysis of content to identify patterns, terminology mapping, and evaluation. Analysis consisted of reviewing the DSS data, identifying the concept models or patterns within the data, and creating the rules necessary to map to SNOMED CT. 'Rule concepts' or 'questions' are necessary for clinical decision support, and linkage to the standards mapping and were made in conjunction with the DSS' needs and knowledge. For example the rule might state, 'postictal state > 10 min', and is used to determine that the patient has been postictal for a certain duration of time. The mapping phase consisted of linking all the DSS concepts in the HDD to represent its hierarchy, and then linking all the clinical concepts in the DSS terminology to SNOMED CT concepts within the HDD. The evaluation phase consisted of looking at the SNOMED CT concepts used in mapping to determine difficulties that could arise when the DSS implements the mapping in their system.

Results

Analysis resulted in the creation of base concepts within the concept model. Base concepts are considered 'domains' that were used for grouping the answer choices. The base concepts were created based on the DSS' clinical knowledge. The mapping phase consisted of adding all local representations and relationships into the HDD, and then linking these and the base concepts to the appropriate SNOMED CT concepts. Once all concepts were mapped both the SNOMED CT concepts and our mappings were evaluated for consistency and accuracy. Inconsistencies were found in SNOMED CT concept coverage causing difficulties in implementing the DSS' rule concepts and some examples are illustrated in **Table 1**.

Table 1 Inconsistencies in SNOMED CT concept coverage

DSS Concept	Patient History SNOMED CT Concept	Family History SNOMED CT Concept	Issue
History of back pain	1. History of (contextual qualifier) -392521001 2. Backache (finding) -161891005	Family history of backache (situation)- 429976008	There is a 'family history of backpain' but not a 'patient history of back pain'
History of GI bleed	History of - gastrointestinal bleed (situation) - 275551007	1. Family history with explicit context (situation) -57177007 2. Gastrointestinal hemorrhage (disorder)- 74474003	There is 'patient history' but not 'family history'
History of stroke	History of - cerebrovascular accident (situation) - 275526006		Pre coordinated found for this but not for 'history of dysphagia' as shown above

Conclusion

Results demonstrated inconsistencies in SNOMED CT concept coverage. Types of inconsistencies included pre/post coordination of past history, family history, contraindications, negation, and normal/abnormal results. This caused the rules utilized in the DSS to be quite complex. There needs to be consistency in SNOMED CT concept coverage for systems to implement in a consistent way. Otherwise, users will not know when to look for pre- versus post-coordinated terms. SNOMED CT is a clinical terminology standard that could be utilized in EHRs and clinical systems, however for this to be achieved further work in the consistency of concept coverage is needed.

Automated notification of primary care providers upon patient admission: pilot results from a randomized controlled trial

Gregory W. Hruby MA[1], Hojjat Salmasian MD[1], Rimma Pivovarov[1] MA, Daniel G. Fort MPH [1], Nancy M. Chang MD[2], David K. Vawdrey PhD[1]
[1]Department of Biomedical Informatics; [2]Department of Medicine
Columbia University Medical Center, New York, NY

Background

Healthcare reform in the U.S. demands greater coordination and continuity across care settings. Building on our previous work to automate the process of identifying patients' primary care providers (PCPs) in an electronic health record, we implemented an electronic notification system to alert PCPs when their patients were admitted to our hospital. In preparation for a randomized controlled trial to evaluate the impact of the system, we conducted a pilot study involving attending PCPs caring for patients in an ambulatory internal medicine clinic at our academic medical center. This work reports the results of the pilot evaluation.

Methods

The automatic PCP notification system was developed in collaboration with clinical, informatics, and information technology groups at New York-Presbyterian Hospital. Using the Arden Syntax, a medical logic module (MLM) was created that sent a secure health message (SHM) to a PCP when one of his/her patients visited the emergency department (ED) or was admitted to our hospital. From November 2, 2013 to January 1, 2014, we measured baseline 30-day post-discharge follow-up visit to the clinic as well as 30-day hospital readmission rates. From January 15 to March 1, 2014, we measured the impact of SHM notifications using 30-day post-discharge follow-up visit to the clinic as our primary outcome. We obtained consent from 16 attending physicians and randomized them to two study groups of n=8: the intervention group received SHM notifications, while the control group did not. In addition to our primary outcome, we calculated 30-day hospital readmission rates and conducted a survey that was issued to PCPs within 24 hours after each patient admission or ED visit. The survey was designed to verify the accuracy of the PCP identification mechanism, assess the PCP's awareness of his/her patient's ED or hospital encounter, and evaluate the PCP's perception of the usefulness of the SHM notification.

Results

Survey. The overall survey response rate was 66.3% (77/116). The survey indicated that our identification mechanism was still performing as intended, with an accuracy rate of 92.2% (71/77). PCP awareness of the admission or ED event for the study group and control group was 62.8% (27/43) and 21.4% (6/28), respectively ($p=0.043$). In the intervention group, physicians who were initially notified through the SHM, 88.9% reported it useful.

Clinical Outcomes. The baseline rate for 30-day readmission and 30-day follow-up visit in clinic was 20% (8/40) and 94.6% (88/93) respectively. The 30-day readmission rate for the intervention and control group was 30.4% (7/23) and 26.6% (4/15) respectively, ($p>0.999$). The 30-day follow-up visit in clinic rate for the intervention group and control group was 80.0% (56/70) and 73.9% (34/46), respectively ($p=0.498$).

Conclusion

We observed a significant increase in primary care providers' awareness of their patients' ED visits or hospital admissions in the group that received SHM notifications. We observed a 6% absolute increase in the rate of 30-day follow-up visit to the clinic from the control group compared to the intervention group, although it was not statistically significant. Furthermore, we confirmed that the automated PCP identification mechanism we previously studied has remained accurate over time.

Acknowledgements: This work was partially funded by National Library of Medicine Fellowship #5T15LM007079-19 and National Science Foundation IGERT #1144854 (RP).

Development and Evaluation of a Mobile Medication Patient Symptom Support (M^2-PASS) System to Strengthen Cancer Patients' Transition from Hospital to Home

Angela Hu, BA[1], , Kenneth Patrick, MD[1], J. Robert Beck, MD[1], DeLinda Pendleton, RN, MSN, CPHQ[1], Mike Korostelev, MSECE[2], Ning Gong, MSECE[2], Li Bai, PhD[2] Kuang-Yi Wen, PhD[1]

[1]Fox Chase Cancer Center, Philadelphia, PA; [2]Temple University, Philadelphia, PA

Abstract

During the transition from hospital to home, cancer patients are particularly vulnerable to adverse drug events, a known driver for hospital readmission rates. Text-based mobile health interventions represent a highly customizable and low-cost platform that can be integrated with telephonic nurse follow-ups to enhance standard care delivery. This study evaluates the feasibility of an integrated 2-week patient medication and symptom management intervention (M2-PASS) delivered via short message services (SMS) and telephonic nurse follow-ups.

Introduction

Transitional care refers to a set of actions which ensure the coordination and continuity of health care as patients transfer between different locations or different levels of care within the same location. Cancer patients are particularly vulnerable to adverse drug events (ADEs), a known driver for hospital readmission rates, during their transition from hospital to home. Improving the quality of transitional care and discharge planning is critical for reducing high hospital readmissions rates and will require multidisciplinary cooperation to develop novel methods for addressing medication management, symptom monitoring, and patient education. Text-based mobile health (mHealth) interventions represent a highly customizable and low-cost platform that can be integrated with telephonic nurse follow-ups to enhance standard care delivery and engage patients with calls to action. This study evaluates the feasibility of an integrated 2-week patient medication and symptom management intervention (M^2-PASS) delivered via short message services (SMS) and telephonic nurse follow-ups.

Methods

Content development was overseen by an expert group who advised on common drug side-effects, patient empowerment strategies, and health literacy. Information technology systems were developed iteratively on the front and back-end to support secure, automatic delivery and receipt of SMS messages on a user-friendly interface. This study will recruit N = 100 newly discharged adult cancer patients to the 2-week intervention. Participants will receive medication reminder texts the morning after their discharge, be prompted to respond confirming receipt of medications reminder texts, complete a nurse follow-up call two days after discharge, and receive twice daily symptom management texts. Feasibility will be determined by participants' confirmed receipt of medication reminder texts and satisfaction with the intervention.

Results

This study will be completed in summer 2014 and data will be available for conference presentation. We will evaluate user participation based on confirmed receipt of medication texts and completion of nurse follow-up calls. Changes in medication adherence, symptom severity and distress, self-efficacy for symptom management, and health-related quality of life will be assessed between baseline and post-intervention assessment. We will also document the frequencies of readmission rate and evaluate their potential causes. Post-intervention patient interviews will be conducted to determine user satisfaction and quantitatively analyzed to modify or improve the intervention in terms of message frequency, content, and intervention duration.

Discussion

Medication adherence is a critical component of transitional care. This study is innovative in its integration of text-based mobile health solutions and telephonic nurse follow-up calls to enhance the delivery of transitional care in adult cancer patients. Text-based symptom management also serves as a low-burden method of improving patient self-efficacy and confidence in their own care. This integrated health delivery system offers a novel approach to reducing health care costs related to hospital readmissions.

Usability of a Novel Wearable Camera System to Inform Tailored Intervention with Dementia Family Caregivers

Lu Hu, MSN[1], Jennifer Lingler, CRNP, PhD[1], Julie Klinger, MA[1], Laurel Mecca, MA[1], Grace Campbell, PhD, MSW[1], Amanda Hunsaker, LSW, MPH[1], Sally Hostein, BA[1], Bernardo Pires, PhD[2], Martial Hebert, PhD[2], Richard Schulz, PhD[1], Judith Matthews, PhD, MPH[1]
[1]Univ. of Pittsburgh, Pittsburgh, PA, USA; [2]Carnegie Mellon Univ., Pittsburgh, PA, USA

Abstract: *Clinicians usually rely on family caregivers' (CGs) reports, which may be selective or biased by imperfect recall, to understand challenges faced at home. Our multidisciplinary team designed a wearable camera system to capture daily interaction between persons with dementia (PWDs) and their CGs. Clips derived from the video data provided a springboard for discussing strategies to deal with dementia. Preliminary findings suggest that our system facilitates tailored interventions which may increase CG confidence and reduce burden.*

Introduction: Family CGs of PWDs often suffer distress from dementia-related difficult behaviors exhibited at home. However, primary care providers have limited opportunity to observe dyadic interaction during brief clinical encounters, and they have no objective method for assessing these caregiving concerns *in situ* to inform their plan of care. We have harnessed advances in mobile computing and camera technology to develop such an objective method, and in this presentation we report preliminary findings regarding the usability of our system with the first five dyads.

Methods: We visited community-residing PWDs and their CGs to obtain their consent and collect baseline demographic, health, and functional status information; to learn CGs' attitudes toward technology; and to assess CGs' perceptions of caregiving burden (Zarit Burden Interview) and self-efficacy. CGs were shown how to use our system, which includes a tiny camera attached to overglasses and a vest that holds the battery and electronics. They were asked to wear the system for as many waking hours as desired at home over two 3- to 7-day periods. CGs received a daily call to remind them to change the battery and SD card and to insert the SD card into a USB port on a laptop left in the home, to enable remote uploading of their camera data. Video clips derived from the first wear period plus CG reports of difficult behaviors exhibited by PWDs formed the basis for an intervention that included suggestions of individualized strategies that the CG could implement. These suggestions were made during a home visit by our interventionist, which was followed by three booster calls to reinforce the intervention. Usability was assessed immediately post-intervention, after the second wear period. Measures of CG burden and self-efficacy were repeated 3 months later.

Results: Five dyads (4 Caucasian, 1 African-American) have completed this protocol to date. Two caregivers were husbands and three were daughters, with a mean age of 65.2±19.4 years. All PWDs were female (M=79.8±5.2 years), and their mean Mini-Mental State Examination score was 18.4±11.7. All study participants had at least a high school education (CGs: 15.6±0.9; PWDs: 14±2.0 years). A total of 711 video clips was generated over an average of 13 days per dyad, with wear times of up to 8.5 hours per day beginning as early as 7:00 AM and ending as late as 1:00 AM the next day. Tailored interventions included suggestions to create distraction, simplify instructions, anticipate threats to safety, accept or increase respite from other family members or community agencies, and pre-medicate with short-acting analgesics to prevent or reduce activity-induced pain during bathing or exercise. After the second wear period CGs rated statements about our system from 1 (not at all accurate) to 10 (extremely accurate) and reported it to be fairly easy to learn to use (M=9.2±1.3), with little concern about invasion of privacy (M=2.4±1.9) or system flimsiness (M=2.8±4.0). They reported experiencing minimal nervousness (M=1.6±0.9), anxiety (M=2.2±2.7), embarrassment (M=2.8±4.0), or confusion (M=3.4±3.3) related to using it. In contrast, CGs did not regard the system as particularly attractive (M=3.2±3.2), capable of making life easier (M=2.8±2.0), or helpful in achieving important goals (M=2.8±1.5). Their views were neutral regarding the system's portability (M=5.0±4.2) and potential benefit vs. cost (M=4.2±3.3). Compared to baseline, CGs' self-efficacy at 3 months post-intervention increased from 58.3 to 66.1 and their burden decreased from 27.4 to 24.3. Positive comments included "Had my doctor had a copy of that tape [showing] her constricting [from severe contractures] like that for the past one or two years…[her pain] would have been addressed much sooner" and "Maybe a one or two-day video…should be supplied…to the doctor, so he can see for himself what's going on."

Conclusion: CGs are willing to wear a camera system to capture daily interaction with their PWDs, providing evidence to inform tailored intervention which may, in turn, increase CGs' confidence and reduce their burden.

Integrating a Diagnostic Decision Support Tool into an Electronic Health Record and Relevant Clinical Workflows through Standards-Based Exchange

Nathan C. Hulse, PhD[1,2], Grant M. Wood[1], Siew Lam, MD,MS[1], Michael Segal MD,PhD[3]
[1]Intermountain Healthcare, Salt Lake City, UT; [2]Department of Biomedical Informatics, University of Utah, Salt Lake City, UT; [3] SimulConsult Inc., Brookline, MA

Abstract

Diagnostic decision support systems (DDSS) have a rich history that dates back to very early experiments in medical informatics. Yet the overall uptake and usage of these types of tools in a clinical setting has been characterized as lower than expected. In an effort to make better use of a DDSS that specializes in rare diseases and genetic diseases, we are integrating the software more tightly into clinical workflows within our electronic medical record, making it available as an integrated part of our clinical notes module. The integration uses relevant HL7 standards, including the Context Aware Knowledge Retrieval Application (Infobutton) standard and the Clinical Document Architecture (CDA), to facilitate communication of the data in and out of the DDSS. In addition, the system integration assists the physicians in creating clinical notes, capturing the information entered into the software and returning to the medical record recommendations and inputs ranked by pertinence for the physician to review. We anticipate that the integration of the resource will make it easier for physicians to use the software and benefit not only from the diagnostic aid, but also in the creation of necessary documentation required for care and even the justification for the ordering of specific genetic tests.

Introduction/ Background

Diagnostic error has been highlighted as an important issue in medicine, and diagnostic decision support systems (DDSS) hold great potential for addressing this issue. Yet their overall uptake in regular practice has been described as far lower than would be expected[1]. Researchers have explored the reasons behind this and have characterized several themes that have led to lower uptake of DDSS in practice. Some of the primary reasons include 1) that the average physician does not feel the need to seek out DDSS in routine care 2) time pressures, in part due to double data entry and 3) perception that the likelihood of the DDSS suggesting something that would significantly alter their treatment for the patient is low.

Methods

In 2013, Intermountain Healthcare, Geisinger Health System and SimulConsult were awarded an SBIR grant in which we jointly proposed a tighter integration between the medical record and the DDSS offered by SimulConsult. In this effort we have attempted to address some of the underlying reasons for lower DDSS usage mentioned above by making the tool available and connected within clinical workflows. At Intermountain, we have met with a group of geneticists and neurologists (likely candidates for using the tool) and have identified some key functions for addressing workflow. These include 1) Integration with the clinical notes module, a typical 'kick-off point' for the use of this type of DDSS 2) standards-based data input into the tool, preventing the need for having users re-enter data that is already present in the medical record and 3) the ability for the SimulConsult tool to pass back template clinical notes and a session summary report, using standards-based exchange. Users will also be able to review other physicians' interactions with the tool, as well as re-launch a previous session with all previous findings intact.

Results/Discussion

In this poster, we will present results from our integration efforts, including details about our approach and design, overall usage data, user feedback, and how relevant HL7 standards for data exchange fit into the effort.

Conclusion

We feel that our approach provides value for our users by lowering barriers to usage, including improved workflow consideration, less need for duplicate data entry, and greater practical value in assisting users with necessary documentation. Aspects of our approach would likely benefit other DDSS implementations in achieving greater use.

References

1. Berner ES. Diagnostic decision support systems: why aren't they used more and what can we do about it? AMIA Symposium Proceedings. 2006:1167–8.

Evaluating the size of deceased patient EHR research data sets: A multi-year trend analysis

Vojtech Huser MD, PhD[1], Aaron Miller PhD[2], David K. Vawdrey PhD[3]
[1]NIH Clinical Center, Bethesda, MD; [2]Marshfield Clinic Research Foundation, Marshfield, WI; [3]Columbia University Department of Biomedical Informatics, New York, NY

Abstract

In recent years there has been an increased interest in streamlining research access to Electronic Health Record (EHR) data. In addition to deidentification and rending EHR data free of Protected Health Information (PHI), records of deceased patients can be also used in research. Decedent research clause within Health Insurance Portability Act (HIPAA) regulation enable such research use without the requirement of full IRB review or informed consent.[1] To investigate the value of deceased subject Integrated Data Repository (dsIDR), we analyzed the size of dsIDR at various timepoints. We demonstrate on an IDR of a large integrated delivery network (Marshfield Clinic), that an IDR will contain more deceased patients than living patients in 2056 by projecting current data trends into the future. A confirmatory assessment at IDR of Columbia University Medical Center showed a similar trend.

Background and Methods: The Marshfield clinic (MC) health system consists of 57 centers and employs over 800 physicians and more than 3,000 nurses. Since 1985, MC has been a pioneer in the use of electronic health record with a home-grown EHR system called CattailsMD. It represents a large integrated delivery network (IDN) and captures a large portion of the medical care in central and northern Wisconsin. MC's integrated data repository is hence a suitable model data repository to study the size of deceased patient population in time. We tested the hypothesis that at some point in the future, the number of decedents stored in the RDW will be larger than the number of extant patients in the warehouse. Due to gradual adoption of EHR, the current coverage of the local population by any given warehouse (of an integrated delivery network) should grow in time and the warehouse will eventually contain data on all long-term residents and deceased long-term residents in a geographical area of an IDN. We analyzed the size of dsIDR at various times (December 31st of each year from 2002–2013). To avoid bias by patients that move to an area temporarily and do not receive their long-term care within the IDN, we only included patients with at least a 15–year time span between their first and last lab test (L15, "long-term" patients) as a proxy for long term residents.

Preliminary Results: The Table shows the computed *dsR ratio* for years 2002-2013 as a fraction of deceased patients over the total patients (only L15 patients are included in the table).

Discussion and Conclusion: In the study period, we see a growing number of living L15 patients. This reflects a growing coverage of IDR of the local population. We expect that this number will approximate to the number of care-seeking long-term residents in the covered geographical area (reduced by the IDN market share proportion). The dsR column indicates growing ratio and increase by 0.89 percentage points per year. Extrapolating this trend into the future we expect MC's IDR to contain more deceased patients in 2056. Although the results assume a dominant role a given IDN in a geographical area, the results apply clinical settings with multiple provider networks, if the patients maintain their provider preference over time. To confirm our findings, we computed similar measures for IDR at Columbia University Medical Center (CUMC) in New York City for years 2005–2013. CUMC data showed a similar rising trend in dsR, from 3.67% in 2005 to 8.97% in 2013. Differences in the dsR rate between the two institutions may be related to differences in populations and variation in the accuracy of death records in the institutional clinical data warehouses.

Table: dsR ratio data for Marshfield Clinic.

Year	Total	Deceased	Living	dsR
2002	25,263	563	24,700	2.23
2003	37,921	1,356	36,565	3.58
2004	48,480	2,210	46,270	4.56
2005	57,764	3,333	54,431	5.77
2006	67,159	4,516	62,643	6.72
2007	76,375	5,843	70,532	7.65
2008	85,469	7,253	78,216	8.49
2009	95,780	8,846	86,934	9.23
2010	107,032	10,714	96,318	10.01
2011	119,696	12,727	106,969	10.63
2012	131,608	14,884	116,724	11.31

Reference

1. Huser V, Cimino JJ. Don't take your EHR to heaven, donate it to science: legal and research policies for EHR post mortem. J Am Med Inform Assoc, 2014, 21(1), p8-12.

Encoding Performance Measures For Automated Quality Assessment

Tammy S. Hwang, [1]Susana B. Martins, MD MSc,[1] Samson W. Tu, MS [1,2], Dan Y. Wang, PhD[1], Paul Heidenreich, MD[1, 2] Mary K. Goldstein, MD, MS[1, 2]
[1]VA Palo Alto Health Care System, Palo Alto, CA; [2]Stanford University, Stanford, CA

Abstract

Performance measures evaluate the proportion of relevant patients for whom appropriate medical care was provided. Performance measures for heart failure (HF) are complex to evaluate, requiring multiple sources of clinical data, such as diagnoses, procedures, labs, and medications. Currently, most health care systems use human reviewers to do manual review of patient data to report quality measures for HF. We created an automated HF performance measurement system utilizing a Protégé knowledge base and EON software as an execution engine.

Introduction

National Quality Forum (NQF) reviews and endorses standardized healthcare performance measures (PMs), also called quality measures, used by many hospitals and healthcare systems to monitor and improve care. Manual chart abstraction for data to compute PMs is labor-intensive and hence costly; automated systems could decrease the time of professional staff in chart review to reserve professional expertise for cases requiring human judgment. In this project, we aimed to develop a pipeline for automated performance measurement of heart failure (HF) management.

Methods

Subject matter experts identified high priority HF PMs endorsed by NQF. We focused on measures pertaining to Angiotensin-Converting Enzyme inhibitors/Angiotensin Receptor Blocker therapy and beta-blocker therapy in HF patients (NQF 0081 and 0083). The automated pipeline is designed to identify (1) whether a patient is eligible for a recommendation during a PM's measurement period (denominator) and (2) whether the patient's medical regimen during the measurement period adhered to the recommendation (numerator). The process becomes complex when introducing denominator exclusions, which may be medical-, system- and/or patient related-reasons. Where the target PM listed a condition without fully specifying it, we expanded our knowledge sources to related PMs (e.g., NQF 0615) for specifics. We encoded the numerator, denominator, and exclusion criteria in a Protégé knowledge base (KB). For each PM, we could specify multiple methods for defining inclusion or exclusion criterion, for example: for the exclusion criterion "hyperkalemia," patients who had an ICD9 code for hyperkalemia as an inpatient admission diagnosis (from inpatient encounter domain) or who had a potassium greater than 5.0 (from the LabChem data domain using LOINC) would be excluded. For each inclusion or exclusion criteria, we defined the time period to search for the data (e.g. performance measurement period), decision rules for missing data, appropriate data domains to search (e.g. ICD9 codes, CPT codes, allergies, or labs), and appropriate data ranges when applicable (e.g. creatinine > 2.5). We adapted the EON execution engine, previously developed to provide guideline-based CDS, to process and evaluate the criteria encoded in the Protégé KB, and we then stored the results in a SQL server database. The output is an automated PM report of proportion of eligible cases, computed as the numerator/denominator as well as detailed information about patients meeting exclusion criteria.

Results

We encoded 6 inclusion criteria and 34 exclusion criteria. Preliminary evaluation, with a convenience sample of 20 outpatient and 17 inpatient cases, compared human reference standard with system output; the system made the correct conclusions in all cases regarding eligibility for the PM, lists of exclusion criteria, and percent attained on the PM for those eligible.

Conclusion

Our results suggest that our HF PM system automates performance measurement of HF management taking account of complex criteria and multiple domains of structured patient data. This could allow for real time performance measure and support interventions to improve health care outcomes by quickly identifying patients who would benefit from specific therapeutic regimens. Further validation and refinement with a larger dataset is needed.

Acknowledgments. VA HSR&D QUERI Heart Failure RRP 11-428 (PI: Goldstein). Views expressed are those of the authors and not necessarily those of the Department of Veterans Affairs or other affiliated institutions.

Smart infusion pump limit violations and high alert medications: the role of the Single Step Titration Error Prevention system

Authors: Cary Ikemoto RPh; Tim Hoh RPh; Idal Beer, MD, MBA, MPH. Fluid Systems Therapeutic Area, Baxter Healthcare Corporation, Deerfield, IL. cary_ikemoto@baxter.com

Problem: Smart pump technology is intended to reduce programming errors with Dose Error Reduction Software (DERS), which triggers alerts when programmed values exceed facility-defined upper or lower dosing limits. Titration programming, where a significant incremental dose or rate change does not exceed soft or hard dosing limits is not intercepted by traditional DERS. High alert medications, which bear a heightened risk of patient harm when used in error, are frequently titrated. Titration of high alert medications may cause harm, even within conventional soft dosing limits. **Purpose:** An additional drug limit alert related to percent titration is available from at least one device manufacturer. It captures titration programming entries that exceed a predefined rate change percentage (both high and low). The purpose of this investigation is to review smart infusion pump data to understand the frequency of titrations and related alerts which may demonstrate the significance of having an extra safety check within the soft limits for high-alert medications. **Methods:** Blinded infusion pump data from two (2) U.S. IDNs were analyzed: IDN 1: average of 2198 pump channels, 9 months of data (2013); IDN 2: average of 692 pump channels, 7 months of data (2012). Initial programming and titrations were analyzed for the following high alert medications: Amiodarone, Dobutamine, Dopamine, Epinephrine, Esmolol, Fentanyl, Heparin, and Norepinephrine. The SIGMA Spectrum Infusion system, used for this analysis, includes a Single Step Titration Error Prevention system in addition to traditional DERS. The Single Step Titration Error Prevention system captures programming events when the incremental dose change exceeds the facility defined percentage (%) increase or decrease. **Results:** There were 19,711 programming events, 43% related to titration programming (Figure 1). Of the titration programming, soft dosing limit violations accounted for 44% of alerts, hard limit violations accounted for 10% of alerts, 43% of alerts were titration limit violations and 3% were a combination of soft limit and titration limit alerts (Figure 2). **Conclusion**: Titrating high alert medications, even with soft and/or hard dosing limits can potentially cause patient harm. The Single Step Tritation Error Prevention System prompted clinician review, confirmation, and intercepted 3646 potential patient safety events in this analysis.

Figure 1. Total Programming Events

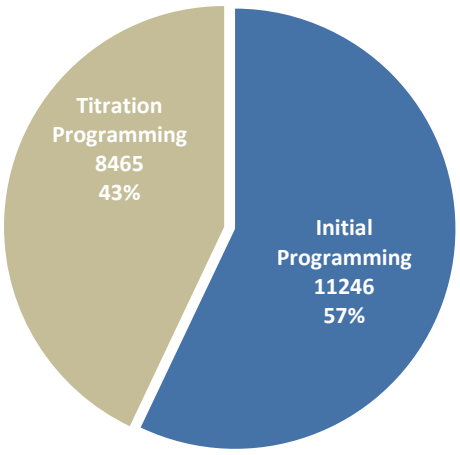
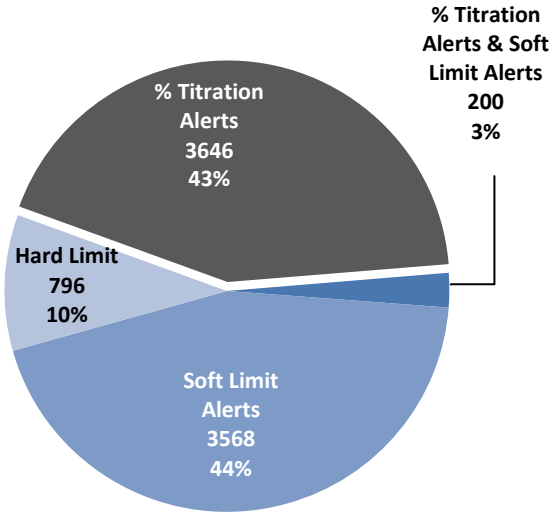


Figure 2. Titration Programming Alert



Deploying Informatics Tools to Improve an Interactive Medical Education Experience

Timothy D. Imler, M.D., M.S.[1-3], and Jason Cadwallader, M.D., M.S.[2,4]

1. Division of Gastroenterology and Hepatology, Indiana University School of Medicine, Indianapolis, IN.
2. Department of Medicine, Indiana University School of Medicine, Indianapolis, IN.
3. Division of Biomedical Informatics, Regenstrief Institute, Inc., Indianapolis, IN.
4. Richard L. Roudebush VA Medical Center, Indianapolis, IN.

Problem Addressed

The morning report session can be anxiety provoking in many learners as they are often put on the spot (pimped) during the session to extract their understanding of the case. This often leads to learners either avoiding the interaction, or not submitting their excellent contributions for fear of public humiliation.

Informatics System Solution

Morning report has been a staple of medical education dating back to Sir William Osler at John Hopkins in the early 20th century. The style of morning report differs quite significantly around the country with some academic centers using cases admitted from the previous night, while others utilize interesting cases discovered during the course of a month's clinical rotation. Almost universally the case is run by a "facilitator" (faculty member or chief resident) while the "learners" consist of medical students, residents, and sometimes fellows.

We hypothesized that "learners" would prefer a model that allowed a crowdsourcing of learning with the ability to share their knowledge without risk of public humiliation. After discussion with key stakeholders within the medical education system a framework was devised in order to support a dynamic, mobile enabled system that allowed direct interaction with the case in an anonymous form.

The system (From the Case Files) was developed to allow a "facilitator" to have a mobile device that enables them to directly place information both on a projected screen, and on the "learners" mobile device (Figure 1). This information could then be utilized to dynamically generate a differential diagnosis using SNOMED-CT concepts (Figure 2) and orders using LOINC concepts (Figure 3). Differential diagnosis generation and order entry are crowdsourced to allow information generated by other learners to be anonymously added for ranking by an individual learner.

Additional features such as Medicare based reimbursement costs, laboratory based result questioning, and group ranked differential (probable, possible, and unlikely) were all involved in the system creation. A demo video of the participants' interaction during a case presentation is available at https://www.youtube.com/watch?v=pqIvO5nGP5U.

The system is undergoing further development to provide individualized feedback to learners and facilitators while adding additional features of gamification. It is being beta tested at a single large academic institution for morning report with plans to release for general use.

Figure 1: Learner screen showing elements from history enabled by facilitator.

Figure 2: Learner screen showing dynamic and crowd sourced differential diagnosis ranking.

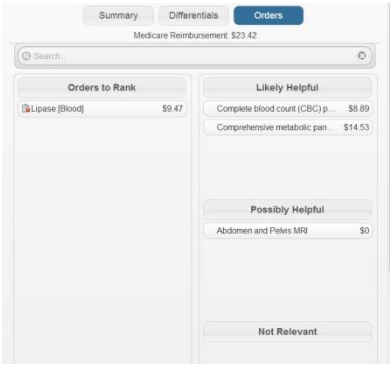


Figure 3: Learner screen showing dynamic and crowd sourced ordering and ranking.

A Systematic Review of eHealth Interventions to Improve Health Literacy

Robin J. Jacobs, PhD, MSW, MS; Jennie Q. Lou, MD, MSc; Raymond L. Ownby, M.D., Ph.D.; Joshua Caballero, Pharm.D., B.C.P.P.
Nova Southeastern University, Fort Lauderdale, FL

Understanding patients' health literacy (HL) in relation to behavioral risk factors is an important goal in the prevention and management of chronic diseases. One possible approach to addressing low HL is to create eHealth interventions that are acceptable, easily deployed, and cost-effective. Currently, few studies exist that systematically reviewed technology-based HL interventions. The purpose this study was to identify and summarize eHealth technologies employed to improve HL; discuss effectiveness of eHealth applications to improve HL based on reports of attributes; and identify the gaps in knowledge in eHealth applications to improve HL to guide future research.

A review of the current state of the science regarding types of eHealth technology for health literacy interventions was conducted. The study selection criteria flowed directly from the review question (i.e., What are the current eHealth interventions to improve health literacy?) and were specified a priori. Interventions had to include at least one eHealth delivery component (e.g., touchscreen computer, handheld electronic device, Internet delivered and one measure of [or components related to] health literacy to promote positive change in lifestyle behaviors for improved health outcomes). Abstracts were searched in a systematic fashion using the keyword search "health literacy" AND "health information technology" AND "eHealth" OR "e-Health" from scientific databases to assess the presence of eHealth applications targeting HL. A search of 16 databases (e.g., Biomedical Reference Collection, Health Technology Assessments, MEDLINE) yielded 466 abstracts, of which 12 were included in this review.

Compared to control interventions, the interventions using technology reported significant outcomes or showed promise for future positive outcomes regarding HL in a variety of settings, for different diseases, and with diverse populations. Understanding and measuring patients' health literacy in relation to behavioral risk factors is an important goal in the prevention, detection, and management of chronic diseases. A concern is the fact that overall health literacy rates are poor and even poorer for individuals from lower socioeconomic and/or ethnic minority backgrounds. Implementation of eHealth and health information technologies is being considered as an effective alternative in addressing current concerns about the health status and quality and safety of the U.S. health care consumer population. Thus it is imperative that we ascertain best practices for delivering health literacy interventions using information technology that is accessible and cost-effective. This review has indicated that it is possible to deliver eHealth interventions specifically designed to improve health literacy skills for people with different health conditions and risk factors. There is also evidence to suggest eHealth interventions may be more effective particularly for individuals with very low literacy. What remains less clear is the extent to which patients with will feel comfortable using a computer or handheld electronic device or will have access interactive eHealth programs using these modalities. It is also likely that understanding how the healthcare system works in addition to eHealth interventions is an important aspect of health literacy. Before eHealth interventions can be hailed as a behavior change intervention of the future, the effective components and mechanisms need to be identified, rigorously tested, and its cost-effectiveness established in different contexts.

An Analysis of Mayo Clinic Search Query Logs for Cardiovascular Diseases

Ashutosh Jadhav, MS[1], Amit Sheth, PhD[1], Jyotishman Pathak, PhD[2]
[1]Knoesis Center, Wright State University, Dayton, OH; [2]Mayo Clinic, Rochester, MN

Abstract

Increasingly, individuals are taking an active role in learning and managing their health by leveraging online resources. Understanding online health information searching behavior can help us to study what health topics users search for and how search queries are formulated. In this work, we analyzed 10 million cardiovascular disease (CVD) related search queries from MayoClinic.com. We performed a semantic analysis on the queries using UMLS MetaMap and analyzed structural and textual properties as well as linguistic characteristics of the queries.

Introduction

Since the early 2000's, Internet usage for health information searching has increased significantly. According to the latest 2013 Pew Survey, one in three American adults have gone online to find out information about a medical condition. According to the Center for Disease Control and Prevention (CDC), in the United States, CVD is one of the most common chronic diseases and the leading cause of death (1 in every 4 deaths) for both men and women. Prior studies have shown that online resources are 'significant information supplement' for the patients with chronic conditions. One of the most common ways to seek online health information is via Web search engines, such as Google. Therefore, studying search queries can help us to understand Online Health Information Seekers' (OHIS) "information needs" and how they formulate search queries, which in turn, will empower us with knowledge to improve the health search experience, as well as to develop more advanced next-generation knowledge and content delivery systems. Although chronic diseases affect large population, very few studies have investigated online health information searching for chronic diseases and especially for CVD. We address this knowledge gap in the community by analyzing CVD related search queries. Some of the potential beneficiaries of this work are Web search engines and health websites.

Methods

In this study, we collected 10 million CVD-related search queries that direct users from Web search engines to the Mayo Clinic's consumer health information portal (MayoClinic.com). We performed the following analysis on the CVD related search queries: 1) Top search queries associated with CVD, 2) categorization of the queries into health categories using UMLS Metamap based on UMLS concepts and semantic types, 3) Structural analysis: length of the search queries, usage of search query operators and special characters in the search queries, and 4) types of search queries (keyword based, Wh and Yes/No questions), misspellings in the queries, and linguistic structure of the search queries.

Results and Discussion

Most of the top CVD queries are related to major CVD diseases and blood pressure (high/low). Top searched health categories for CVD are 'Diseases and Conditions' and 'Vital Signs'. Even though CVD prevention is possible, very few OHIS search for prevention while most of them search for symptoms and post diseases information (Living with, Diet, Treatment, Drugs). The average length of a CVD search query (3.88 words and 22.22 characters) is longer than that of a general search query, which implies that OHIS describe health information needs in more detail. Usage of a search query operator (4%) is limited and variation of 'AND' (AND, &, +) is used more often (95%) followed by 'OR'. Only 3.2% of the search queries contain at least one spelling mistake. OHIS predominantly formulate search queries using keywords followed by Wh-Questions and Yes/No Questions. Almost all CVD search queries have at least one noun.

Conclusion

We found that using Metamap and UMLS concepts/semantic type is a very good approach for categorization of health related search queries into health categories. This study extends our knowledge about online health information searching behavior, and provides useful insights for Web search engines, health-centric websites, healthcare providers, and healthcare-centric application developers.

Evaluation of the Health Level Seven Fast Health Interoperable Resources (FHIR) Standard as a Query Data Model for the Arden Syntax

Robert A. Jenders, MD, MS, FACP, FACMI
Charles Drew University & University of California, Los Angeles, CA

Abstract

Context: Arden Syntax is a standard that encodes knowledge as Medical Logic Modules (MLMs) but that lacks a standard query data model. Objective: Assess to what extent the Health Level Seven (HL7) Fast Health Interoperable Resources (FHIR) standard can represent MLM query data elements. Method: 340 MLMs containing 3268 queries were examined. Result: FHIR can be used to represent all these query data elements. Conclusion: FHIR adequately represents data queried using the Arden Syntax.

Introduction

Arden Syntax is an American National Standards Institute (ANSI) formalism supervised by Health Level Seven (HL7) for representation of procedural medical knowledge with the goal of facilitating sharing units of knowledge known as MLMs. Some site-specific changes must occur in order for a knowledge base to be transferred from one site to another. Key to minimizing site-specific changes is the standardization of database linkages, which in turn requires identification of a standard data model, vocabularies and query syntax. This is sometimes known as the "curly braces problem" of Arden because of the syntactic construct used to enclose these site-specific references[1]. The HL7 FHIR draft standard for trial use (DSTU) Release 1 is a new framework that allows references to data to be defined, represented in XML and bound to terminologies in modular components known as "resources." For example, concepts such as "patient," "medication" and "observation" are key FHIR resources, each with structured attributes that may be other resources. Increasingly, implementations involving HL7 standards and the standards themselves are being developed using FHIR. The present work was undertaken to assess FHIR's utility as a standard data model for queries in the Arden Syntax.

Methods

A previously assembled, robust convenience sample of MLMs was examined. The data query statements were extracted from these MLMs, and the data elements therein were identified. Each then was assessed to ascertain whether it could be specifically represented using FHIR DSTU Release 1.

Results

A total of 340 MLMs were pooled from 6 source CDS systems, including 19 from 3 vendor knowledge bases and 312 from 3 academic medical centers. MLMs concerned mainly with lab tests were the most common (138/331 = 42%), followed by clinical assessment or classification (75/331 = 23%) and medication (45/331 = 14%). The remainder addressed administrative and miscellaneous topics. Each MLM contained at least one READ statement with a data query. A total of 3268 queries were identified, and the data elements therein were compared to current FHIR resources to assess whether they could be represented. All the data elements in these queries could be represented by FHIR resources. Some data elements were complex and required linked resources to represent fully (e.g., bacteriologic reports) but were still representable in FHIR. Elements that were not directly representable in other data models previously proposed for the Arden Syntax such as the HL7 Virtual Medical Record (e.g., references ranges and pregnancy status) were representable in FHIR Of note, the primary time of query data elements in Arden is an implicit attribute of every query variable and is not explicitly represented, but this, too, is representable in FHIR.

Conclusions

FHIR is adequate to represent the data elements found in a large set of query statements in a corpus of Arden Syntax MLMs. Consideration should be given for use of FHIR to represent data elements in a standard way in queries in the Arden Syntax in order to facilitate knowledge sharing.

References

1. Jenders RA, Corman R, Dasgupta B. Making the standard more standard: a data and query model for knowledge representation in the Arden Syntax. Proc AMIA Symp 2003;:323-330.

Exercise Intensity Adherence Improved by Remotely Controlled Cycle Ergometer

In cheol Jeong, PhD[1], Joseph Finkelstein, MD, PhD[1]
[1]Chronic Disease Informatics Program, Johns Hopkins University, Baltimore, MD

Introduction

Cycling exercise is an essential component of rehabilitation in older adults and individuals with various chronic conditions. Access to exercise facilities may in these individuals be limited by mobility impairment, lack of transportation or insurance coverage. Recently we introduced the interactive Biking Exercise (iBikE) system that was demonstrated to support patient-centered model of individualized self-care applications for telerehabilitation, disease management, and health promotion. However patients with compromised upper limb mobility may have difficulties in following their cycling exercise plan which affects safety and efficacy of home-based programs. To address this barrier, a remotely-controlled mode of cycling exercise intensity has been introduced. The purpose of this project was to compare adherence to cycling intensity during active (self-controlled) and passive (remotely-controlled) exercise.

System

The upper limb telerehabilitation system that supported the internet-controlled home cycling exercise consisted of a clinician unit, internet server and a home unit. The home unit consisted of a touch screen tablet, ergometer control unit, ergometer and wireless biosensors. LabVIEW 2011 SP1 was used to operate this system. A low-cost data acquisition device (NI USB-6008) which had plug-and play full-speed USB connectivity was designed to provide prescribed speeds by controlling the voltage output level in ergometer. The system design is shown in Figure 1.

Evaluation

The degree of adherence to prescribed exercise intensities was compared by investigating how well subjects followed varying target speeds at the self-effort mode that did not use the remote control mechanism, and the guidance mode that was controlled remotely. Two-minute periods of consecutive target cycling levels of 0.5, 1.0, 1.5, 1.25, 0.75 miles/hour were provided to subjects by displaying corresponding speed on exercise dashboard. The actual cycling speed was obtained by an electromagnetic sensor in the cycle machine with 50 Hz sampling rate. Each volunteer was assigned to both passive and active mode in a random order.

Results

Overall, 8 volunteers participated in the study. Absolute mean differences between prescribed and actual levels (AMD) and coefficients of variations (CV) were calculated to assess adherence. Average AMD for self-effort mode was: 0.054±0.006 (range 0.046-0.060 mi/hr), and for guidance mode: 0.022±0.010 (range 0.010-0.037 mi/hr); mean CV for the self-effort mode was: 9.622±6.300 (range 5.22-20.43%), and for guidance mode: 3.338±0.660 (range 2.79-4.46%). The differences between AMD and CV were statistically significant ($p<0.05$).

Discussion

The cross-over evaluation of exercise intensity adherence demonstrated higher adherence with upper limb exercise speed prescription during remotely-controlled mode compared to self-effort mode. We concluded that this approach may be a valuable option in frail individuals or patients with movement disabilities.

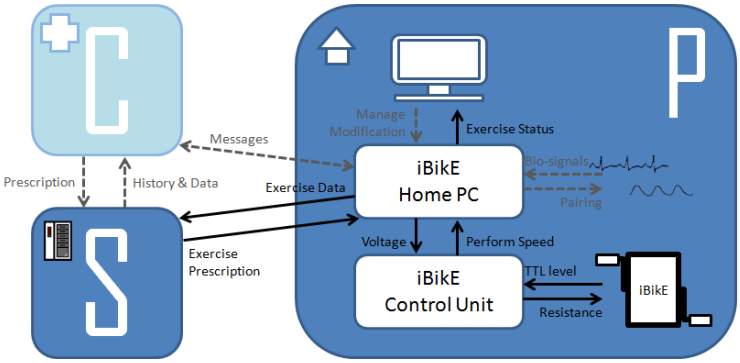


Figure 1. Remotely-controlled iBikE system

Social InfoButtons for Patient-oriented Healthcare Knowledge Support

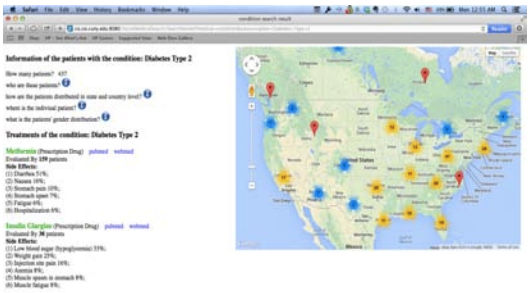
Xiang Ji, BE[1], Soon Ae Chun, PhD[2], James Geller, PhD[1]
[1] New Jersey Institute of Technology, Newark, NJ;
[2]City University of New York, College of Staten Island, Staten Island, NY

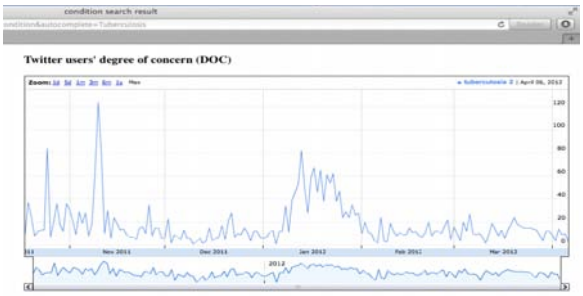
Abstract: Delivering contextually-relevant knowledge resources into EHR systems at the point of care was proposed as *Infobutton HL7 standard*. However, the *InfoButton* standard does not consider social media data or the patients' healthcare behaviors or practices. We present a *Social InfoButtons* system that collects patient-generated social media data and other open health data to provide insights on healthcare trends and patients' practices and issues, using the Semantic Integration model, that supports Social Healthcare Knowledge for clinicians, patients and policy makers.

Introduction: A study [1] showed that 34% of the Internet users utilized social media to read other patients' commentaries and experiences about health or medical issues. The InfoButtons system of Cimino et al. [3, 4] provides knowledge resources needed by clinicians but does not incorporate patient-generated social health data. With the Health 3.0 trend, it is increasingly becoming important to understand the patients' actual health practices, behaviors, trends and concerns. Our *Social InfoButtons* system generates contextually summarized information about social health practices by geographic or temporal dimensions, providing end-users (e.g. patients, clinicians, or government officials) with healthcare information, such as treatments, practices, conditions, experiences, sentiments, and behaviors reported by other patients through social media.

Methods: Providing integrated social health knowledge is challenging, since the online health data sources vary in formats and platforms. We present a semantic model for representing the integrated social health knowledge, and use Linked Open Data [2] to integrating heterogeneous social media health data sources. We developed the Social InfoButtons architecture to provide a social health analytics platform that can summarize, visualize and compare the contextually relevant health information for patients, clinicians, and healthcare government officials. The architecture consists of the data collector, RDF triple store, and analytics module. (1) The data collector extracts publicly available health data from various data sources, such as the social network site PatientsLikeMe, the government-maintained CDC website, the PubMed website, and the patient resource portal WebMD; (2) The extracted entities from different sources were integrated with RDF triples; (3) The coverage of the analytics module of Social InfoButtons ranges from simple SPARQL queries to interactive geographic and temporal visualizations to gather insights about social health trends and anomalies

Results: The integrated knowledge base currently has 612,017 triples representing 1228 conditions from different sources and their related information such as treatments, symptoms, and side effects of treatments. The figures below show the geographic knowledge retrieved by Social InfoButtons for Diabetes Type 2 (left), and a timeline trend of negative sentiments towards tuberculosis from Twitter users (right).





Conclusions: A Social InfoButtons prototype was developed to provide social health knowledge that is relevant to an end users' context. A semantic integration model and social analytics module are presented for social health knowledge, extracted from data sources ranging from social network sites, research communities, and government sites to patient Web resources.

References

1. The Social Life of Health Information. http://www.pewinternet.org/files/old-media//Files/Reports/2011/PIP_Social_Life_of_Health_Info.pdf
2. Bizer C, Heath T, Berners-Lee T. Linked Data - The Story So Far. Int. Journal on Semantic Web and Information Systems. 2009.
3. Cimino JJ, Li J, Bakken S, Patel VL. Theoretical, Empirical and Practical Approaches to Resolving the Unmet Information Needs of Clinical Information System Users. Proceedings of AMIA Annual Symposium. 2002.
4. Del Fiol G, Huser V, H. Strasberg HR, Maviglia SM, Curtis C, Cimino JJ. Implementations of the HL7 Context-Aware Knowledge Retrieval (''Infobutton'') Standard: Challenges, strengths, limitations, and uptake, J of BioMedcal Informatics 45, 2012.

Personalizing Statistical Models for Asthma Prognosis and Therapeutics

Hongyang Jia, Patricia Flatley Brennan, Shiyu Zhou, Junbo Son, Yu-Ting Hung
Department of Industrial and Systems Engineering, University of Wisconsin-Madison, Madison, WI

Abstract

Researchers often use statistical models when developing decision support methodologies for asthma prognosis and therapeutics. We conducted a selective literature review and contrast existing methodologies with our proposed statistical approach, which leverages new patient-generated data and addresses individual features.

Background

The Smart Asthma Management (SAM) project (NSF: IIS1343969) aims to apply a statistical modeling methodology employed in reliability engineering, to the challenge of asthma management. In literature, statistical models, e.g. Markov Models, Bayesian Networks, are most often used in asthma prognosis while Expert System forms the dominant methodology in recommending therapeutics. As patient-generated, frequently sampled data becomes ubiquitous, promising yet challenging opportunities arise to build alternative analytical method to capitalize it, and accelerate patient-centered asthma care.

Method

We reviewed selected literature within the scope of decision support methodology in asthma therapeutics and prognosis. We compared the model approaches on two aspects: purpose and data source, and then elaborated the comparison between SAM approach and traditional statistical methods.

Result

The results of selected literature review are presented in Table 1. All existing approaches leverage population-level data or expert experience to formulate models for either diagnosis or therapeutics. In SAM we propose to augment population-based prediction in three ways: First, SAM augments population prediction with continuous or frequently sampled data, for example real-time patient sensing or easy to capture self-reports rather than conventional clinical assessment data. Other statistical models are designed based on clinical data and cannot capitalize new data sources[1, 2]. Autocorrelation, missing data and imprecision are not addressed enough as well. We will construct a multistate model from population level estimates then refined with extension based on an individual patient's data. Thus, SAM will build a personalized diagnosis model for each patient while others use a generic model for every one[1, 2]. Third, because it incorporates both treatment and state indicators, SAM will provide recommendations for both diagnosis and therapeutics based on new data source. Overall, SAM provides new perspective on asthma diagnosis and therapeutics, and facilitates patient-centered asthma management.

Table 1. Contrast of different model approaches

Model	Purpose	Data Source for Model Building
Statistical Model	Diagnosis	Population
Neural Network	Diagnosis	Population
Expert System	Treatment Planning	Heuristics
Case Based Reasoning (AI)	Treatment Planning	Population

Discussion

Decision support technologies can capitalize on frequently sampled patient data, but the models to process that data will perform best if the model can manage the autocorrelation, missing and imprecision of patient-generated data.

References

1. Jackson CH, Sharples LD, Thompson SG, Duffy SW, Couto E. *Multistate Markov models for disease progression with classification error.* The Statistician, 2003. **52**(Pt 2): p.193-209.
2. Sanders DL, Aronsky D. *Detecting Asthma Exacerbations in a Pediatric Emergency Department Using a Bayesian Network*. AMIA Annu Symp Proc 2006:684-688.

Readmission Leakage Risk Stratification with Associative Classification

Yugang Jia, Lin Li, Usha Nandini Raghavan, and Nathan Cohen

Philips Research North America

Abstract

A high percentage of patient's readmissions[1] occur at other hospitals instead of the one they first admitted to, which is defined as readmission leakage. Associative classification method is applied to identify low and high readmission leakage patient cohorts for generating actionable insights for management purpose. Results show that admitted patient with non-acute cardiovascular disease or certain behavior pattern has a high chance of readmission leakage.

Introduction

Understanding and managing readmission leakage risk is very important for hospital to prevent revenue loss and improve their patient management cost-effectively in the whole continuum of care1. The identified leakage patterns can provide actionable insights to stratify the patients and support management improvement, including redesign of inpatient management practices and proactive outreach to patients, referring physicians and community hospitals to manage patient transitions and support care coordination.

Method and Results

Unplanned 30 days readmission data (around 77K) from eight large-scale hospitals in NY is extracted from Healthcare Cost and Utilization Project (HCUP) database (2009-2011) Agency for Healthcare Research and Quality, where the average readmission leakage rate is 28%. Associative classification method[2], is used to extract meaningful associations between patient features and their hospital choice. All patient cohorts with more than 5% of whole readmission population are identified. Then cohorts with more than 50% of readmission leakage or less than 20% of patient leakage are selected as high and low readmission leakage cohorts for rule generation respectively.

The 5-fold cross validation results as shown in left side of Figure 1 indicates the observed leakage ratio is statistically significant different cross three leakage risk levels (p-value < 0.01). If a hospital is capable of managing 10.8% readmissions (leakage risk classes), it can cover 19% leaked readmissions for leakage prevention.

Examples of identified high risk patient cohorts are shown in right side of Figure 1. Non-acute cardiovascular patients that discharged to home have significantly higher leakage risk compared the overall average (p-value < 0.01). For this type of patients, their leakage risk is tied to the patient satisfaction, which can be improved by inpatient management. Two other cohorts with significant leakage ratio are young male patients who live in metro area and patients with drug abuse (p-value < 0.01). To prevent their leakage, hospitals need to better maintain contact with patients by proactively outreach to patients, referring physicians and support coordination of care.

References

1. Gerhardt G, Yemane A, Hickman P, Oelschlaeger A, Rollins R, and Brennan N. Data Shows Reduction in Medicare Hospital Readmission Rates During 2012. Medicare & Medicaid Research Review 2013: Vol 3, No 2.
2. Welch S. R., S. Huff M.. Cohort Amplification: An Associative Classification Framework for Identification of Disease Cohorts in the Electronic Health Record, AMIA Annu Symp Proc. 2010; 2010: 862–866.

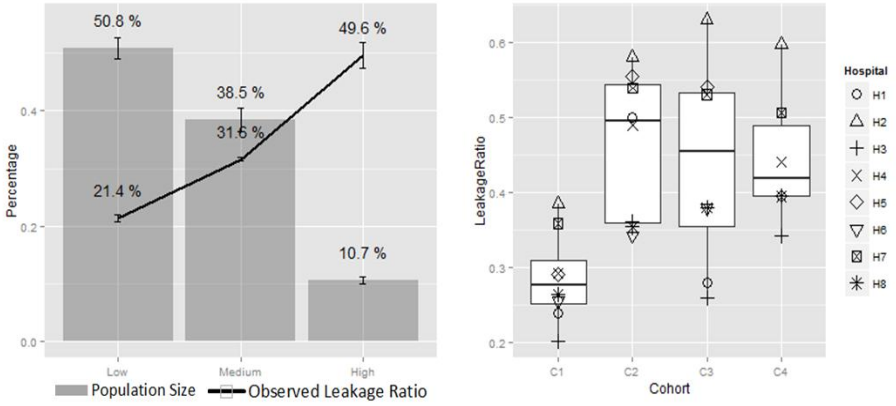


Figure 1. Left: 5-fold cross validation of readmission leakage risk model. Right: High risk patient cohort examples. C1: All population. C2: Comorbid with Drug Abuse.C3: Male from Metro Area with Medicaid or Selfpay.C4: Non-acute cardiovascular and discharged to home

Lexical Term Standardization of ICD-11 Using Semantic Web Technologies

Guoqian Jiang[1], Harold R. Solbrig[1], Bedirhan T. Ustun[2], Christopher G. Chute[1]
[1] Department of Health Sciences Research, Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN; [2] World Health Organization (WHO)

Abstract

As part of work plan of the World Health Organization (WHO), the lexical terms used in ICD-11 labels or properties of entities should follow a standard and homogeneous approach. The objective of the study is to develop and evaluate a solution to identify term heterogeneity in ICD titles using Semantic Web technologies.

1 Introduction

The beta phase of the 11th revision of International Classification of Diseases (ICD-11) started in May 2012. As part of work plan of the World Health Organization (WHO), the lexical terms used in ICD-11 labels or properties of entities should follow a standard and homogeneous approach. For example, the terms "cardiac," "renal" should always be used in ICD titles, definitions and values instead of the terms "heart," "kidney". The objective of the study is to develop and evaluate a hybrid solution to identify term heterogeneity in ICD titles.

2 Materials and Methods

We utilized a lexical toolset known as the Sub-Term Mapping Tools (STMT) developed at National Library of Medicine (NLM) [1]. Two main features of the tool were used: 1) to find all sub-terms for an entity; 2) to find all permutations of synonymous sub-term substitutions. We developed a Semantic Web-based wrapper service that links WHO ICD-11 content services with local STMT lexical sub-term services. Specifically, the wrapper service takes an ICD entity URI as the input (which retrieves the title of the ICD entity), and then renders the sub-terms of an entity and the synonyms of each sub-term in a Semantic Web Resource Description Framework (RDF) format [2] using the W3C standard Simple Knowledge Organization System (SKOS) signatures. Using the wrapper service, we harvested the sub-terms and their synonyms in RDF triples for all foundation entities (n= 29,445) and loaded them into an open source RDF triple store known as 4store. We enabled a SPARQL endpoint that provides standard SPARQL query services against the sub-term dataset and analyzed the dataset.

3 Results

We retrieved the sub-term pairs, in which the preferred label of a sub-term appears to be the synonym of the other sub-term, and identified 4,927 distinct sub-term pairs. We manually reviewed a small subset of the sub-term pairs and concluded that they reflect reasonably well the term heterogeneity in ICD titles. For example, for the sub-term "bleeding," we identified its synonymous sub-terms "haemorrhages," "haemorrhage," "haemorrhagic," "hemorrhage," "hemorrhagic," "blood loss," "ruptured," "rupture," "spot," and "spotted". The frequency distribution of such sub-terms would help the decision-making of subject-matter experts in choosing a preferred sub-term out of its synonymous sub-terms so that the preferred sub-term could be used in ICD titles in a consistent way. We plan to work with WHO to conduct a comprehensive review for the set of sub-term pairs and their frequency distribution in ICD titles.

4 Conclusion

In summary, we developed a hybrid approach that combines an NLP-based lexical tool with a Semantic Web-based approach, which would provide an effective and scalable solution for lexical term standardization of ICD-11.

Acknowledgement

This work is partly supported by a Mayo-WHO Contract 200822195-1.

Reference

1. Lu CJ, Browne AC. Development of Sub-Term Mapping Tools (STMT). AMIA 2012 Annual Symposium, Chicago, IL, November 3-7, 2012, p. 1845.
2. WHO ICD-11 Content Services: http://id.who.int/icd/entity/; Last visited at March 10, 2014.

Impact of barcode design on the medication administration process

Junghee Jo, MS[1], Jenna L. Marquard, PhD[2], Lori A. Clarke, PhD[2], Philip L. Henneman, MD[3,4]

[1]Electronics and Telecommunications Research Institute, Daejeon, Korea; [2]University of Massachusetts, Amherst, MA; [3]Baystate Medical Center, Springfield, MA; [4]Tufts University School of Medicine, Boston, MA

Abstract

Barcode medication administration (BCMA) systems have been recommended to help reduce errors during the verification of patient identifiers (VPI) process. BCMA manufacturers claim their systems ensure the "Five Rights" of medication administration. Such claims are valid, however, only if healthcare workers follow a medication administration process consistent with the specific barcode design being used. This poster shows how the VPI process depends on the barcode design, which often varies by state, healthcare organization, and manufacturer.

Introduction

Failure to perform the VPI process correctly is one of the major causes of medication errors.[1] BCMA has been shown to be effective in reducing VPI errors[2], but healthcare workers need to be aware of what information is encoded on the barcodes and perform a VPI process that is appropriate for that design.

Observations

We considered several barcode designs (2 designs are shown in Figure 1), where the ID band and medication label barcodes are encoded with zero, one, or two patient identifiers, and show how the VPI process during medication administration depends on this design.

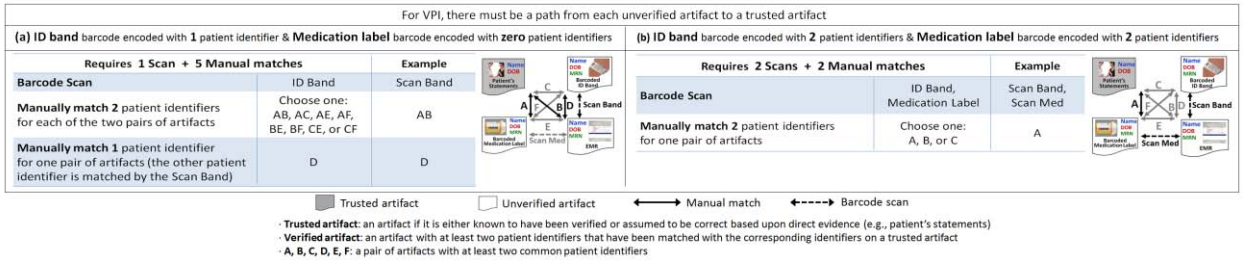


Figure 1. The VPI process varies by choice of barcode design; 2 designs are shown.

For each design, we determined the percentage of nurses complying with the Joint Commission guidelines for VPI by analyzing data collected from a previous clinical simulation study in which nurses administered medications to a mock patient[3]. Regardless of which design the nurses believed to be applicable, few nurses complied with the guidelines; the results ranged from 12% for Figure 1(a) to 32% for Figure 1(b).

Recommendations

Our findings suggest several ways to improve compliance, which may help reduce errors. First, healthcare workers should be aware of the information encoded on barcodes, as the appropriate VPI process depends on the barcode design. Second, as will be illustrated in the poster, encoding two patient identifiers on the ID band and medication label barcodes simplifies the process and reduces the number of required manual (i.e., visual and verbal) matches. Third, teaching health care workers a specific sequence of actions for VPI based on the chosen barcode design will prevent healthcare workers from having to choose among different alternatives. Finally, since unusual situations are bound to arise, healthcare workers should learn the underlying principles for correctly performing VPI.

References

1. Lisby M, Nielsen LP, Mainz J. Errors in the medication process: frequency, type, and potential clinical consequences. IJQHC. 2005;19(1):15–21.
2. Paoletti RD, Suess TM, Lesko MG, Feroli AA, Kennel JA, Mahler JM, et al. Using bar-code technology and medication observation methodology for safer medication administration. AJHP. 2007;64(5):536–543.
3. Henneman PL, Marquard JL, Fisher DL, Bleil J, Walsh B, Henneman JP, et al. Barcode verification: reducing but not eliminating medication errors. JONA. 2012;42(12):562–566.

Acknowledgements: This material is based upon work partially supported by NSF awards 1239334 and 1234070.

Online Quiz Taking: Does Allowing Repeated Tries Without Penalty for Incorrect Responses Increase Guessing?

Craig W. Johnson, PhD, Craig Harrington, MS, MSSW,
Rodney Howell, BSCET, Taiwo Akinwande, RN, BSN
University of Texas-Houston Health Science Center

Abstract

The study examined online quiz-taking behaviors of 121 students in a statistical methods in health informatics course. Four milestone lesson quizzes allowed repeated attempts per each multiple-choice question, with grades determined by performance on the final attempt. Results revealed incorrect responses increased and accuracy of responses decreased on later quizzes, consistent with a hypothesis that: Students learn to guess more frequently. Additional research needs to assess whether effects are related to lesson difficulty levels.

Background

According to cognitive learning theories, mental activities like practicing and rehearsing, are vital to the acquisition of knowledge[1(p6-7)]. Online techniques may provide fewer controls over the learning environment, allowing students to interact in unexpected ways. This study analyzes three operationalizations of measures of guessing where online students may retry each quiz question without penalties.

Research Question

Do changes that are consistent with increased frequency of guessing – specifically number of tries per quiz (nbr_tries), first-try correct responses (score), and percent of total tries that are correct on the first try (pct_correct) – occur across four quizzes selected within a course or across course iterations?

Method

Data were deidentified lesson responses of 121 students from 2008 to 2013. Four representative online hypertutorial lessons were selected from the 22 covered during a statistical methods in health informatics course. Selected lessons were separated by two-week intervals. A multivariate split-plot design included year as a between-subjects factor and lesson as a within-subjects factor and three dependent variables: nbr_tries, score, and pct_correct.

As variables were extremely positively skewed in the presence of unequal sample sizes and violated sphericity and Box M Test assumptions, the research employed SPSS v.22 to conduct parametric and nonparametric doubly multivariate analyses using the GLM/MANOVA Repeated Measures procedure for tests of linear, quadratic, and cubic trends among within-subjects effects. Puri-Sen *L* statistics from the nonparametric trend test analyses were calculated using ranked data.[2] Decisions regarding statistical significance of nonparametric tests did not differ substantively from those of parametric tests; ergo results of the more familiar parametric tests are reported.

Results

Over the four lessons, significant ($p < .005$) linear and cubic trends and profile plots revealed small mean differences in earlier lessons followed by significant and substantial mean differences that leveled off for later lessons. Score and pct_correct significantly decreased and nbr_tries increased with later lessons.

Discussion

Results are consistent with the hypothesis that guessing increases across lessons as shown by an increase in the number of tries per quiz, decreases in the number of correct responses on the first try of a quiz, and the percent of total tries that are correct on the first try. Analyses did not control for lesson difficulty levels and students may have been willing to select less certain answers as they experienced no penalty for incorrect responses. While the increased numbers of tries per quiz, decreased numbers of correct responses on the first try and percent of total tries correct on the first try were consistent with the guessing hypothesis, future research should address such potential threats to validity.

References

1. Mergel, B. (1998). *Instructional Design & Learning Theory*. University of Saskatchewan. http://goo.gl/v0Qaqd.
2. Thomas, J.R., Nelson, J.K., Thomas, K. T. (1999). A generalized rank-order for nonparametric analysis of data from exercise science: A tutorial. *Research Quarterly for Exercise and Sport*, 70(1), 11-23.

An Institutional Strategy to Support Clinical Research with Centrally Managed Custom Data Repositories

Stephen B. Johnson, PhD[1], Thomas R. Campion, Jr., PhD[1], Nonie E. Pegoraro, MA[1], Leon Rozenblit, JD, PhD[2], Charles Tirrell[2], Curtis L. Cole, MD[1]

[1]Weill Cornell Medical College, NY, NY; [2]Prometheus Research LLC, New Haven, CT

Introduction: Healthcare organizations face significant technical and organizational challenges to support clinical research. These barriers have historically led to fragmentation of the research enterprise, which in the extreme force individual researchers to manage their data separately. In some cases, groups of researchers who share common interests have developed custom data repositories that meet their collective needs.[1] These approaches do not scale as an institutional strategy. When there are multiple local repositories, many data management activities are redundant, regulatory processes are convoluted, and opportunities for data sharing and collaboration in new areas are limited. One potential institutional strategy is to develop a centralized data repository, typically by making operational clinical data available to researchers for secondary use.[2] This approach often fails to meet the specialized needs of heterogeneous research groups for primary data collection, study management, local data integration and curation. Exclusive focus on central processing can reduce opportunities to improve research productivity.[1] We describe here a new strategy that focuses on the needs of local research groups, while attempting to achieve economies of scale at the institutional level with a sustainable funding model.

Methods: Architecture for Research Computing in Health (ARCH) partitions information technology for clinical research into external data sources, shared infrastructure and custom research data repositories. External sources include electronic health records, research administration systems, electronic data capture and biobanks. Shared infrastructure includes a *loading zone* where external data can be organized and integrated, and a *working zone* where data are transformed into structures needed locally.[2] Research data repositories provide highly customized, self-service facilities for extracting data for analysis and identifying cohorts as well as data quality assessment and exploration. Shared components are largely funded centrally while custom components are charged back. Both shared and custom components are implemented using an open source platform (RexDB), which offers a model-driven architecture, meta-data management, configuration services and web-based query language (RexQL).

Results: The institution adopted the ARCH strategy with significant financial support from the Joint Clinical Trials Office and the Clinical and Translational Science Center to support hardware and staff for shared infrastructure, as well as a scientific advisory board to vet requests for creating new custom repositories. In addition, two research groups (anesthesia and digestive care) have adopted the strategy by investing in services to provide custom data transformations, user interfaces and reports. The demand for additional repositories from other research groups is high, with three planned for the coming year (pediatrics, urology and hematologic oncology) and new requests added every month.

Discussion: The ARCH strategy enables an institution to centralize research infrastructure for secondary data use, regulatory compliance, data transformation, quality assessment, security, needs assessment and training. Local research groups are fully empowered to collect and integrate specialized data, with customized workflow to maximize scientific coloration and research productivity. A transparent model for cost sharing ensures financial sustainabililty.

References

1 Hruby GW, McKiernan J, Bakken S, Weng C. A centralized research data repository enhances retrospective outcomes research capacity: a case report. J Am Med Inform Assoc. 2013 May 1;20(3):563-7.

2 Abend A, Housman D, Johnson B. Integrating Clinical Data into the i2b2 Repository. Summit on Translat Bioinforma. 2009 Mar 1;2009:1-5.

Semi-Automated Method to Extract Semantic Information from EHR Flowsheet Data for Pressure Ulcer Research

Steven G. Johnson, MS[1]; Jung In Park, BS, RN[2]; Matthew D. Byrne, PhD, RN[3]; Beverly Christie, DNP, RN[4]; Lisiane Pruinelli, MS, RN[2]; Suzan Sherman, PhD, RN[4]; Bonnie L. Westra, PhD, RN, FAAN, FACMI[2]

[1]University of Minnesota, Institute for Health Informatics; [2]University of Minnesota, School of Nursing; [3]St. Catherine's University; [4]Fairview Health System

Problem

Significant amounts of useful data are contained in non-standard, semi-structured electronic health record (EHR) flowsheet data. The goal of this study was to develop a method for semi-automatically mapping this data into semantic concepts in order to define a data model for pressure ulcer research.

Introduction and Background

Pressure ulcers occur for more than 2.5 million people each year and are estimated to result in $9.1B to $11.6B in unnecessary health care costs and causes pain, infection risk, increased hospital stays, and death [1]. Evidence-based practice guidelines exist for prevention of pressure ulcers and documentation captured in electronic health records (EHRs) can be used to relate patients' conditions and care provided to evaluate the effectiveness of these guidelines on outcomes. However, EHR documentation must be mapped to standard clinical models and concepts in order to analyze the benefit of these practices. A standard model for pressure ulcer data is being developed from HL7, but EHRs capture most pressure ulcer related data as customized fields.

Methods and Data Source

Clinical data, including flowsheets, were extracted from one Midwest health system's EHR and loaded into a clinical data repository at the University of Minnesota. A subset of 200,000 records was extracted and analyzed to develop a clinical data model and map concepts to national data standards to link patient characteristics and care to pressure ulcer outcomes. The flowsheets contained 15,000 different types of data, which would have been difficult and error prone to normalize and standardize for research. Therefore, a semi-automated method was developed to map these types of data into standard SNOMED CT concepts. The method uses natural language processing (NLP) tools on the flowsheet field definitions and attribute information to match semantically equivalent concepts.

Results / Discussion

The semi-automated process produced a set of SNOMED CT concepts that were used to create a model of pressure ulcer data useful for research. The method uncovered a number of data fields that were really the same concept and also helped group the concepts that were more closely related, which simplified the resulting model. This same process can be used with any health system's data to efficiently map custom flowsheet data fields to standard concepts and extract the relevant pressure ulcer related information.

Acknowledgment

"This was supported by Grant Number 1UL1RR033183 from the National Center for Research Resources (NCRR) of the National Institutes of Health (NIH) to the University of Minnesota Clinical and Translational Science Institute (CTSI). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the CTSI or the NIH. The University of Minnesota CTSI is part of a national Clinical and Translational Science Award (CTSA) consortium created to accelerate laboratory discoveries into treatments for patients."

References

1. Berlowitz, Dan, et al. Preventing Pressure Ulcers in Hospitals: A Toolkit for Improving Quality of Care. April 2011. Agency for Healthcare Research and Quality, Rockville, MD. Available from: http://www.ahrq.gov/professionals/systems/long-term-care/resources/pressure-ulcers/pressureulcertoolkit/index.html

The Extent to which U.S. Hospitals Promote Their Patient Engagement Activities and Outcomes: Preliminary Results of Quantitative Content Analysis Research

Josette F. Jones, Maryam Zolnoori, Samar Binkheder, Katherine Schilling, Michelle linox, Lakshmi Ravali Pondugala
School of Informatics and Computing, Indiana University- Indianapolis, IN 46204

Introduction

Patient engagement is "a concept that combines patient activation with interventions designed to increase activation and promote positive patient behavior, such as obtaining preventive care or exercising regularly [1]." Within the scope of healthcare information technology (HIT), patient engagement is managed by technologies that can be used for myriad purposes ranging from enabling patients to view their medical histories online, to communicating with practitioners online, to using electronic platforms for capturing and disseminating large quantities of patient data. The goals of technology trends around patient engagement include improving accuracy of diagnosis and treatment of patients.

Purpose

The purpose of this study is to describe and make inferences regarding the extent to which hospitals in the U.S. meet patient engagement criteria delineated by the National eHealth Collaborative (NeHC) [3].

Methodology

A quantitative content analysis method has been used to examine and code content from the website of 100 hospitals nationwide. The hospitals were randomly selected from a comprehensive list on health.usnews.com, Hospitals were categorized as "small," "medium," or "large" based on the number of outpatient visits recorded annually. The contents of 100 hospital websites are being analyzed using a patient engagement framework which describes an inclusive platform for achieving patient engagement [3]. This framework includes five stages for patient engagement: 1) inform me, 2) engage me, 3) empower me, 4) partner with me, and 5) build my e-community. In order to quantify the results of hospitals website analysis, data are labeled with the following binary criteria: applicable criteria found (+1), criteria not found (-1), not sure (0), and not applicable (9). Statistical techniques including t-test and chi-squared test are used to analyze relationship between patient engagement criteria and sizes of hospitals, hospital rankings, and number of outpatient visit. In addition, a descriptive analysis will be provided for every section of the framework to better understand how hospitals promote patient engagement through their public websites.

Preliminary Results

The researchers expect that the results of data analysis will illustrate the extent to which small, medium, or large hospitals (based on counts of outpatient visits) inform, engage, empower, and collaborate with patients. Results will reveal the strengths and weaknesses of hospital websites in demonstrating an organization's relative levels of patient engagement, as well as the extent to which the achievement of Stage 2 requirements of the Meaningful Use legislation have been achieved and communicated.

Conclusion

As we move from a physician-centric model of healthcare to a patient-centric model, the role of hospitals in publically promoting their patient engagement activities and outcomes becomes increasingly important. The findings of this study will provide a snapshot of current affairs and a discussion of relevant trends and themes.

References

1. "Health Policy Brief: Patient Engagement," Health Affairs, February 14, 2013.

2. Gruman, J., Rovner, M. H., French, M. E., Jeffress, D., Sofaer, S., Shaller, D., & Prager, D. J. (2010). From patient education to patient engagement: implications for the field of patient education. *Patient education and counseling*,*78*(3), 350-356.

3) National E-health Collaborative (2014, February) , http://www.nationalehealth.org/patient-engagement-framework,

Automated Early Warning System for Monitoring Workflow, Evaluating Patient Care and Predicting Risk in Secondary Care in the UK.

Julie S Jones-Diette, PhD[1],
Michael Brown[2], Gemma Housley[1], Jim Hatton[3], Dominick Shaw[1, 4]

[1]East Midlands Academic Health Science Network, Nottingham University Hospitals NHS Trust, Nottingham, UK. [2]Horizon Digital Economy Research Institute, University of Nottingham, UK. [3]Deputy-Director of Information and Performance, Nottingham University Hospitals NHS Trust, Nottingham, UK. [4]Division of Respiratory Medicine, School of Medicine, University of Nottingham, UK.

Introduction

The presentation will discuss the UK's first single-source big-data technology platform for NHS secondary care. Key to its technical and commercial innovative potential is the use of key data sources which are already collected daily by all NHS Trusts in the UK. To date there is no known integration of these sources and their combination will permit accurate assessment, information dissemination, near real-time monitoring, and clinical adverse incident measures at unprecedented levels in the NHS. The UK National Health Service (NHS) deals with 1 million patients every 36 hours. Hospital Trusts in the UK utilise many automated and bespoke systems to coordinate this level of care across all departments producing unique opportunities for data linkage, the prediction of risk and a system which can pull all disparate data streams together for resource planning and operational efficiency which is essential for the future of patient care.

Aim

The Nottingham University Hospitals has been chosen as the pilot site to launch a tool which will build on both existing infrastructure at the Nottingham University Hospitals and already established methods for the management of operational efficiency within the aviation sector. The resulting software system will improve patient experience, support resource planning and optimise hospital efficiency.

Method

Our award winning workflow handheld monitoring system[1], using the wireless system Nervecentre[2], will be combined with other data sources including the NUH Data Warehouse, DATIX Safety Database, ORMIS Electronic record of Theatre events, ED patient flow data, EMAS Ambulance statistics, ESR Electronic Staff Record and E-rostering across the two large teaching hospitals within the University of Nottingham NHS Trust (QMC and City Hospital Campus). A suite of applications for airport optimisation of passenger flow and operational efficiency will be adapted for use in the NUH secondary care setting using information assimilated from these data streams.

Results

The initial output will be a rich database of hospital and patient activity across the two hospitals including current resource allocation within the hospital. Using this information, data analysis will determine the most informative parameters for the prediction of future resource requirement and patient kinetics including areas for potential overburden. Once linked with suitable visualisations designed through a human-factors process with one of our partners, The Horizon Digital Economy Research Institute, we will use multi-method research techniques (walkthroughs, technology-enhanced diary studies, interviews, and surveys) to evaluate the degree to which the data infrastructure improves the sharing and usage of the data within operational and strategic affairs.

Conclusion

The final product will be a real-time suite of applications for patient care and hospital efficiency and for estimates of risk at the ward or hospital level. We will design the suite to support data collected by other NHS Trusts to adapt the system for a UK roll out to ultimately provide shared clinical benefits and evidence-based improvements to workflow nationally.

---

[1] Blakey JD, Guy D, Simpson C, Fearn A, Cannaby S, Wilson P, et al. Multimodal observational assessment of quality and productivity benefits from the implementation of wireless technology for out of hours working. BMJ Open. 2012 : 2(2).

[2] http://nervecentresoftware.com

Tissue-Experiment Inventory: A System to Enable Cataloguing of Experimental Results in Association with Tissue and Participant Information

Norman P. B. Joseph[1], Leonid Kvecher[1], Brenda Deyarmin[1], Lori Sturtz[1], Frank J. Cammarata[1], Caroline Larson[1], Craig D. Shriver[2], Richard J. Mural[1], Hai Hu[1]
[1]Windber Research Institute, Windber, PA; [2]Walter Reed National Military Medical Center, Bethesda, MD

Abstract
Historically our experimental results were managed neither systematically nor centrally. Experiments were performed on a variety of platforms and for diverse projects, and results were not directly linked to other tissue information. To solve this problem, we designed, developed and implemented a process to systematically collect and catalog experimental results in association with the tissue information, permitting us to link these results to existing clinicopathologic data in a data warehouse.

Introduction
Windber Research Institute (WRI) developed the Data Warehouse for Translational Research (DW4TR) with an advanced analytical workflow platform in collaboration with InforSense/IDBS[1]. WRI also maintains a large tissue bank of biological samples for cancer research. DW4TR contains participants' demographics, medical history, life style, risk factors, pathologic annotations of the biospecimens as well as biomarker information assayed by immunohistochemistry and FISH. However, the experimental results derived from these samples were historically managed and stored separately by individual Primary Investigators (PIs). WRI undertook the current work realizing it would be beneficial for researchers to be able to link experimental results directly to elements of DW4TR.

Methods and Results
We reviewed internal records to identify projects, PIs, technicians and platforms under which experimental results from tissue bank data had been generated. PIs were asked to provide lists of requested samples; technicians were asked to provide the locations for experimental results identified by samples and participants. All information was confirmed against the data in the DW4TR. Electronic versions of all result files were then copied to a central archive location, and tables were created in the DW4TR to track project, sample and experimental results location information. The analytical workflow builder (see Figure 1A) was used to develop workflows for integrating experimental results information with these tables, and for linking this information to sample and participant data already existing in the DW4TR. In addition, a workflow was developed to perform ad-hoc queries on the resulting tables, and published to the portal (Figure 1B). We are in the process of expanding this application to enable lab scientists to directly deposit their experimental results into this tissue-experiment inventory.

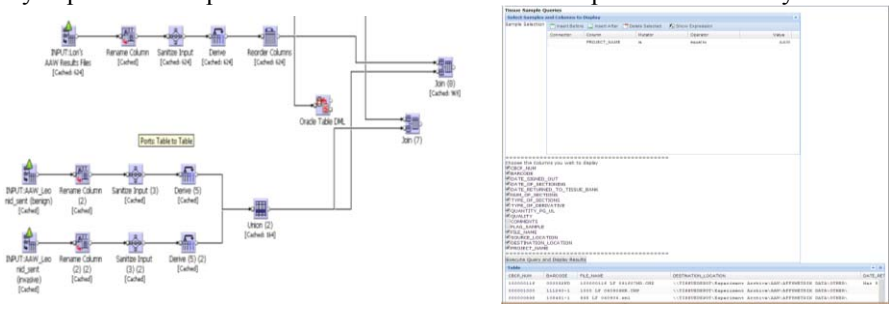


Figure 1. (A) Part of the analytical workflow for data integration. **(B)** Sample portal ad-hoc query.

Conclusions
The Tissue-Experiment Inventory allows us to link clinicopathologic and sample information directly to experimental results, facilitating research and enabling new discoveries.

Disclaimer
The views expressed in this abstract are those of the authors and do not reflect the official policy of the Department of Defense, or U. S. Government.

References

1. Hu H, Correll M, Kvecher L, et al. DW4TR: A data warehouse for translational research. J Biomed Inform. 2011 Dec; 44(6):1004-19. doi: 10.1016/j.jbi.2011.08.003. Epub 2011 Aug 22.

A Statistical Study of Words Used in Chinese Clinical Documents

Meizhi Ju,BS,Haomin Li,PhD*,Huilong Duan PhD
Zhejiang University, Hangzhou, China

Abstract

Corpus is important for developing NLP (Natural Language Processing) technology. For many reasons, corpus of clinical document is very valuable for MLP (medical language processing), especially in China. In this study, we first established a corpus of Chinese clinical documents, then calculated word normalized frequencies based on counting words in different categories, and finally represented them by data visualization.

Introduction

80% of the clinical documentation that exists in healthcare today is unstructured. Therefore, MLP has been widely studied in health care IT industry. While, MLP research in Chinese language is scarce. Lack of corpus of Chinese clinical documents is one of the obstacles. In this primary study, we collected clinical documents and did a basic statistical study of that corpus. Moreover, presenting word and frequency in a visual and intuitive form can be a basis for visualizing analysis research at a deeper level. Such researches based on medical lexicon corpus will be of great value. The corpus built in this research and the related statistical study will better serve further MLP research.

Method

Total 63,172 Chinese clinical documents were collected from a Chinese hospital to establish this corpus. Each document was firstly segmented in sentences and clauses based on punctuations. Then each clause was segmented in words or phrases based on a Chinese word segmentation library named Pangu. After that, we annotated each word with its part-of-speech and another eight customized attributes which could afford an opportunity to classify words.

To give a basic idea about the words used in Chinese clinical documents, we counted the frequency of each word in the corpus. In this step, Arabic numbers, punctuations, names and English letters were excluded from the study.

The clinical documents from different departments or of different note types have unique word usage. To compare this, the normalized word frequency was calculated. For each category, we counted the total number of words as TN and then counted the number of each word as N. We calculated the word normalized frequency through N divides TN for each word. Meanwhile，we sorted these words in a frequency descending order.

To provide a whole picture of the word used in Chinese clinical documents, we have designed a visualization tool to generate charts on word, word normalized frequency and category. In the chart, the horizontal axis represents each specific category. In the vertical direction of each integer point, we place bubbles in descending order related to word normalized frequency and the number of bubbles depends on how many we have selected. The size and color of each bubble is used to represent word normalized frequency and Chinese word, respectively. With purpose of getting distribution rules of words, we count sub-language classes attached to words as one customized attribute in different note types or departments and do large amount of analysis combining with a number of bubble charts.

Result

Through segmenting 63,172 Chinese clinical documents which involves 31 departments and processing the resulting words, we have established a Chinese corpus which involves 324,000 sentences, 12 million clauses, and 50 million words. Symptoms, diagnosis and body parts relevant as well as medication words count for a large percentage while chemical, time and organism words count for a very small percentage in different departments and note types excluding all irrelevant words. In different note types, contrary to operative notes and discharge summaries, pathology notes and progress notes have a large percentage on all sub-language classes especially on diagnosis and symptom sub-language classes. In different departments, emergency department has the lowest rate in diagnosis and symptom sub-language classes, and generic medical department has the highest rate. In addition, the trends of ups and downs about sub-language classes in different departments as well as note types don't have apparent differences.

NLP enhances Quality Care Measures in Heart Failure

Ravikumar K.E., Wagholikar K.B., and Liu, H
Department of Health Sciences Research, Mayo Clinic, Rochester, MN, 55901

Introduction: Congestive Heart Failure (CHF) is one of the leading causes for death across the globe. Ensuring quality care to CHF patients has been a great challenge in clinical practice. HEDIS[1] and NHQM[2] have prescribed guidelines to ensure quality care in the management of CHF patients. Identifying the patient with CHF very early is the first step in ensuring quality care. Clinical notes have been the major source in accurately identifying whether the patient has CHF.

Objective: The chief objective of the study is to investigate the role of NLP in the identification of CHF patients to ensure quality care.

Methods: Our approach to detect evidence for CHF in clinical notes of patient consists of the following 4 steps: 1) *Compile a comprehensive dictionary to identify terms relevant to CHF in the clinical notes* - Based on the criteria used by the nurse abstractors we compiled a dictionary from UMLS[3], which consists of 330 terms and organized into main categories a) Disease/Phenotype terms that describe CHF 2) Drugs used to treat CHF (Digoxin, Coreg), 3) Co-morbid disorders such as renal failure, Hypertension etc. 4) Drugs to manage secondary factors (Lasix etc.). The final dictionary consists of 330 terms. 2) *Use natural language processing (NLP) term-spotting techniques to detect those terms in clinical notes* - We used MedTagger[4], a Aho-Corasik lookup approach to detect the various terms compiled in the dictionary. 3) *Classify patients whether they have CHF or not based on the terms identified in the text* – We had two rules to classify if the patient has CHF. If the clinical note of a patient contains terms that describe both CHF condition and drug used to treat CHF then the patients are classified as CHF patients. If the term from only one of the category is mentioned then the secondary information such as presence of drugs to treat renal failure is relied upon to confirm the CHF condition of a patient.

Data set – Table 1 summarizes the statistics of the data sets that we used in this study. Our study was on 490 patients who visited Mayo Clinic over a period of 2 months (60 days starting May 15th 2012) with complaints related to heart disorders. The records were manually analyzed by four nurse abstractors and classified as CHF patient based on the evidences provided by the physician, lab tests, drugs prescribed, etc.

Table 1 – Data statistics

Data Set	Total Patients	Total clinical notes
Training	376	6239
Test	114	2078
Total	490	8317

Experiments: We retrieved 8317 clinical notes related to 490 patients. We used 6239 notes (from 376 patients) for training and the remaining (2078 from 114 patients) for testing. The training data was used to refine the dictionary and fine-tune the rules to improve the performance of the system on the training data.

Results and discussion: The system achieved a precision, recall and F-measure of 72.38%, 96.23% and 82.61% respectively on the blind test data. The first rule that considers the combined presence of terms from the first two categories is a strong indicator for CHF resulting in very high precision but lower recall (Row1 of Table 2). However it the second rule that classifies a patient to have CHF based on the presence of terms either the first or the second category significantly boosted the recall (96%). However, we observed significant drop in the precision (from 93% to 72.38%) due to the second rule (Row2 of Table 2). This rule helped the system to achieve the best overall F-measure.

Table 2 – Evaluation on test data

System	Precision	Recall	F-Measure
Rule1	91.17%	63.42%	74.80%
Rule 2	72.38%	96.23%	82.61%

Conclusions: We conclude that NLP based techniques significantly helped in the process of early identification of CHF patients. This is very important to ensure quality care and early intervention to the patients while treating for heart failure.

References

1) http://www.ncqa.org/Portals/0/HEDISQM/HEDIS2014/List_of_HEDIS_2014_Measures.pdf
2) http://www.jointcommission.org/assets/1/6/ICUManualPDF.zip
3) Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. (2004) Nucleic Acids Res. 32 (database issue), D267–D270.
4) http://ohnlp.org/index.php/MedTagger

Methods for Early Stakeholder Engagement for Implementation of Health Information Technology

Megha Kalsy, MS[1], Natalie Kelly, MBA[3], Jennifer H. Garvin, PhD MBA RHIA[1, 2, 3], Mary K. Goldstein MD MS[4,5]
[1]Department of Biomedical Informatics, [2]Division of Epidemiology, University of Utah, Salt Lake City, UT;
[3]IDEAS Center, Salt Lake City Health Care System, Salt Lake City, UT; [4]VA Palo Alto Health Care System, Palo Alto, CA, [5]Center for Primary Care & Outcomes Research, Stanford University, Stanford, CA

Abstract

As part of a Department of Veteran Affairs (VA) study, we undertook early stakeholder engagement in order to increase adoption and uptake of an automated system. We used semi-structured interviews and thematic analysis to understand the context of implementation using a combination of the Promoting Action on Research Implementation in Health Sciences (PARIHS) framework of Implementation Science and the Socio-Technical Model of Health Information Technology to inform our work. We identified themes associated with four dimensions of the socio-technical model: hardware and software, clinical content, workflow and communication, as well as internal organizational features that related to barriers and facilitators of our anticipated implementation.

Introduction

Our early stakeholder engagement process sought to understand the context of an anticipated implementation of an automated quality measurement system for inpatients with congestive heart failure. We used the Promoting Action on Research Implementation in Health Sciences (PARIHS) framework[1], which includes the elements of evidence, context, and facilitation, as well as the Socio-Technical Model of Health Information Technology (HIT), which includes the following 8 dimensions; 1) hardware and software, 2) clinical content, 3) human-computer interface, 4) people, 5) workflow and communication, 6) internal organizational features, 7) external rules and 8) measuring and monitoring, of which we used 4, to guide our approach. Evidence provides the clinical basis for the system, and we examine the context or environment in which the implementation will occur. Prior studies have shown that HIT can be used as a facilitator of evidence-based practice[2] and by using Implementation Science we hope to increase adoption and uptake of our system.

Methods

We used a snowball sampling technique to identify interviewees beginning with VA key informants. Key informants provided a broad understanding of QM related to inpatient chronic heart failure within the VA. We also interviewed other subject matter experts based on the recommendations of the key informants, and/or their job category. We developed an interview guide and used it to undertake interviews. Two independent note takers populated the interview guide, summarized the interview and then combined them into a single summary through a consensus process. The summarized interview was then sent to each interviewee for review and editing (validation). We used the collective set of validated summaries to generate preliminary themes (codes) to answer our research questions.

Preliminary Results

We interviewed 14 stakeholders; key informants consisting of VA Quality Management (QM) and measure automation experts with a minimum experience of 5 years working in the VA and a minimum experience of 5 years working in QM and subject matter experts consisting of VA employees with job categories such as: clinical quality specialists; quality management professionals, clinicians and pharmacists; and clinical program analysts. Key informants and subject matter experts' experience in the VA ranged from 2-35 years, and from 2-33 years in QM. Themes associated with internal organizational features related to: use of evidence based care, a culture of continuous quality improvement, and routine use of quality control reporting as a feedback loop. The main theme related to hardware and software was that the VA has multiple informatics tools for clinical care and extensive supporting informatics infrastructure. Themes associated with clinical content related to the importance of medication reconciliation, the presence of clinical data accessible for determining guideline-concordant care. The workflow and communication themes related to workflow associated with patient monitoring, the need for quality measurement data for timely decision making, and the impact of transitioning from retrospective measurement to the potential use of automatically extracted data that could be used in clinical decision support systems (CDSS).

Conclusion

The early stakeholder engagement process is essential for providing information that is useful in system design for successful implementation of an automated system.

VA Disclaimer

This publication is based upon work supported by the Department of Veterans Affairs, Veterans Health Administration, Office of Research and Development, HSR&D, Grant # IBE 09-069. The views expressed in this article are those of the authors and do not necessarily represent the views of the Department of Veterans Affairs or the University Academic Affiliates.

References

1. Rycroft-Malone, J. (2004). The PARIHS framework--a framework for guiding the implementation of evidence-based practice. *Journal of Nursing Care Quality*, *19*(4), 297–304.
2. Goldstein, M. K. (2008). Using health information technology to improve hypertension management. *Current Hypertension Reports*, *10*(3), 201–207.
3. Sittig, D. F., & Singh, H. (2010). A new sociotechnical model for studying health information technology in complex adaptive healthcare systems. *Quality and Safety in Health Care*, *19*(Suppl 3), i68–i74. doi:10.1136/qshc.2010.042085

Benefits and Challenges of a Height Sensing Approach for In-Home Gait Speed Detection

Avinash Kalyanaraman[1], Blaine Reeder, PhD[2], Kamin Whitehouse, PhD, MS[1]
[1]University of Virginia, Charlottesville, VA, [2]University of Colorado, Aurora, CO

Abstract

Changes in gait speed are a predictor of functional change in older adults. As the developed countries of the world experience demographic shift due to population aging, there is a need for informatics-based solutions to support independent living for older adults. This poster discusses potential benefits and challenges of applying a field-tested height sensor to the problem of in-home gait speed detection as an early indicator of functional decline.

Introduction

Early detection of functional decline is important to provide timely interventions that prevent adverse events and threaten the independence of older adults in the community. Detection of changes in gait speed is one strategy to monitor functional change in older adults at home[1]. *Doorjamb* is a door-mounted multi-sensor that integrates motion and height sensors and has been used to reliably track room occupancy for multiple people in home energy efficiency studies[2]. The purpose of this poster is to identify issues in the application of the Doorjamb sensor to the problem of in-home gait speed detection and present initial walk test results for a member of our research team. Walk test data from additional trials will be included for presentation at the AMIA symposium in November 2014.

Benefits and Challenges

One potential benefit of using Doorjamb to measure in-home gait speed is that it is a field-tested technology that will go to production in the near future. Another benefit of Doorjamb is that it was developed to disambiguate multiple residents to track room occupancy. However, one challenge is determining the reliability of Doorjamb to disambiguate multiple residents when measuring gait speed. Another challenge is to understand the potentially high variability of in-situ gait speeds to provide a meaningful measure of gait speeds for different individuals at home.

Walk Test

As a proof-of-concept, a member of the research team performed eight individual walk trials on an "ideal" 10'+ straight-line course, moving toward and perpendicular to the Doorjamb sensor (Figure 1). The blue bar indicates ground truth walking speed as measured by a DSLR video camera. The red and green bars are different ways of estimating walking speed using the doorway sensor. The third and fourth trials captured poor readings of the person walking, making these sensor measurements unreliable. Of the remaining trials, error rates were approximately 0.05m/s or less.

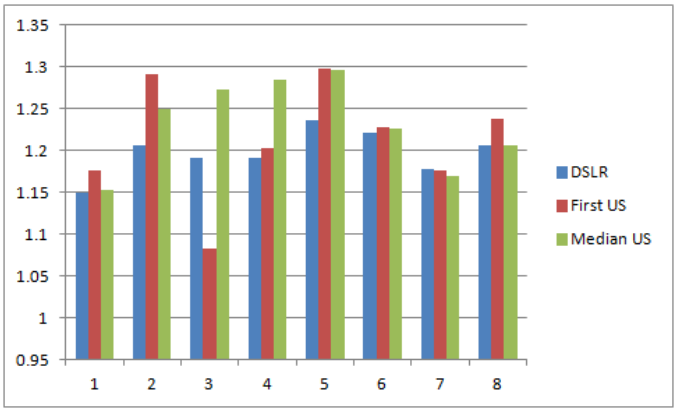


Figure 1. Initial proof-of-concept walk test results

Discussion

Based on our prior experiences in developing custom integrated sensors, we believe we can improve accuracy through additional sensors and algorithm refinement as well as capability to accommodate other than ideal straight-line walks. Future research will include formal walk trials that enroll a mix of young and older adults in a laboratory setting to validate Doorjamb performance before moving to field studies.

References

1. Montero-Odasso M, Schapira M, Soriano ER, et al. Gait velocity as a single predictor of adverse events in healthy seniors aged 75 years and older. *J Gerontol A Biol Sci Med Sci.* 2005;60(10):1304-1309.
2. Hnat TW, Griffiths E, Dawson R, Whitehouse K. Doorjamb: unobtrusive room-level tracking of people in homes using doorway sensors. Paper presented at: Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems 2012.

ABSTRACT

Youjeong Kang MPH, RN[1]
[1]University of Pennsylvania, School of Nursing, Philadelphia, PA

Patient characteristics associated with reshopitalization in older adults with heart failure receiving telehomecare.

Problem: Heart failure (HF) is the leading cause of rehospitalization in the United States. One potential way to reduce HF rehospitalizations is through the use of telehomecare technology. However, studies on telehomecare use in the United States have demonstrated inconsistent results in reducing HF rehospitalizations and lengthening time to rehospitalization. Additionally, little is known about risk factors for rehospitalization during the course of a telehomecare episode.

Study purpose: the aim of this study is to identify patient characteristics associated with HF rehospitalization, and with time-to–first rehospitalization, during receipt of telehomecare services.

Methods: This is a non-experimental retrospective analysis of the Outcome Assessment Information Set (OASIS) dataset from Medicare beneficiaries with heart failure, who were provided telehomecare services by a private home healthcare company. This study used the most current version of the OASIS dataset (OASIS-C) to examine patient characteristics associated with rehospitalization by using multiple logistic regression and survival analysis techniques.

Results: In terms patients with a documented formal pain assessment, patients who indicated severe pain using a standardized assessment tool, were 1.85 times more likely to be rehospitalized than patients without severe pain (p=0.01) within 60 days. Patients who had skin lesions or open wounds were twice more likely to be rehospitalized than those without skin lesions or open wounds (p=0.02) within 60 days. Patients who were able to safely dress their lower body had were nearly three times more likely to be rehospitalized than patients who were able to dress without assistance if clothing and shoes are laid out or handed to them (p=0.01) within 60 days. Patients with urinary incontinence were at 1.35 times greater risk of rehospitalization than those without urinary incontinence (p=0.03) at any given point in time. Patients who had hospitalizations more than twice in the past 12 months were at 1.40 times greater risk of rehospitalization than those with fewer than two hospitalizations in the past 12 months (p=0.03) at any given point in time .

Conclusion: This study's findings may provide home care clinicians with a set of risk factors for use in targeting the patients most likely to benefit from telehomecare or other additional interventions. For example, patients who report severe pain or patients who have been hospitalized more than twice in the past 12 months may be considered high-risk patients who need an additional intervention.

Key words: heart failure, telehomecare, rehospitalization, survival analysis

Automated Detection of Transient Lower Esophageal Sphincter Relaxations

Martin Karpefors, PhD[1] & Magnus Ruth, MD, Prof (affiliated)[2]

[1] Advanced analytics Center, B&I Sciences, GMD, AstraZeneca R&D
[2] Dept of Otolaryngology, Sahlgrenska University Hospital, Gothenburg, Sweden

Abstract

A new method to automatically detect Transient Lower Esophageal Sphincter Relaxations (TLESRs) was developed using a random forest supervised machine learning algorithm. Potentially, this type of method can make fast, objective and reproducible evaluation of manometric recordings, which is a prerequisite for using TLESRs as an endpoint in large clinical trials.

Introduction

Transient Lower Esophageal Sphincter Relaxations (TLESRs) are short relaxations in the tonus of the sphincter which are triggered by gastric distension. They allow venting of air from the stomach but constitute also the main mechanism for reflux of gastric contents into the esophagus. Therefore, the number of TLESRs can be used as a marker for reflux disease.

In clinical trials, experts manually inspect recordings of esophageal pressure to identify the TLESRs according to specific TLESR criteria. Such inspections take approximately 45 minutes even for an expert. In addition, both the intra- and inter-expert variability is relatively high[1]. Consequently it is difficult to use this endpoint in larger studies.An unsuccessful earlier attempt to automate detection was made by van Herwaarden et al.[2] However, they encourage the development of new algorithms.

Method and Results

Using trial data from an internal study, we have developed a supervised machine learning classification algorithm that automatically detects TLESRs. The algorithm was based on random forest and trained on recordings from 10 subjects. All recordings were evaluated by an expert to find the TLESRs. In the training material, the expert identified 113 TLESRs and these were used as the true TLESRs. Training examples of non-TLESRs were chosen at random in the recordings, and in excess by a factor of five. During the training we had an out-of-bag error of 5.75%. However, the main challenge was to use the model on full length recordings. By applying the algorithm stepwise, in a moving window fashion along the entire recording, the probability of having a TLESR at each time point was calculated. For a probability threshold of 0.6, the result showed that the sensitivity was as high as 90% and PPV 50%.

Conclusion

Manual evaluation of TLESR is time consuming and associated with a high reader variability which has implications for the use of the technique in large clinical trials. The present study describe an algorithm for automatic detection of TLESR with the potential to speed up the evaluation process and reduce variability, thus providing a cost effective alternative to manual evaluations.

References

[1] Holloway RH, Boeckxstaens GEE, Penagini R, Sifrm DA, Smout AJPM, Objective definition and detection of transient lower esophageal sphincter relaxation revisited: is there room for improvement? Neurogastroenterol Motil 2012 24, 54-60.

[2] van Herwaarden MA, Samsom M, Wolf C, Leong IS, Smout AJPM, Computer analysis of prolonged lower oesophageal sphincter pressure recordings, Neurogastroenterol Motil 2001 13, 37-44.

A Knowledge-Based Collaborative Clinical Case Mining Framework

Ramakanth Kavuluru, Ph.D[1], Anthony Rios, B.S[1],
Brandon Kulengowski[2], and Patrick McNamara, Ph.D[2]

[1]Division of Biomedical Informatics, Department of Biostatistics, University of Kentucky
[2]Department of Pharmaceutical Sciences, University of Kentucky

Introduction: An important challenge in developing and maintaining an effective Pharm.D curriculum is to design assessment methods that incorporate newer clinical case scenarios that model realistic situations students might encounter in their professional careers. Such scenarios often involve assessment of students' comprehension of a set of core concepts (in physiology, biochemistry, pharmacology) that interact in myriad ways to result in complex cases. Coming up with new assessment scenarios is also practically relevant in obviating the problem of students passing their test questions to future students. However, building an evolving assessment case repository based on these realistic scenarios is a highly time consuming task given there are over 2.4 million case reports and clinical trial publications indexed by PubMed. Our effort aims to expedite this process through knowledge-based collaborative methods.

Contribution: We built a collaborative clinical case mining tool that lets users search for cases based on different aspects of drugs and diseases that are the focus of a curricular component. Users can also assign tags that correspond to different core concepts of the curriculum for subsequent retrieval based on tags and named entities. We deployed an initial prototype that is being used by five researchers in the domain of bacterial infections at http://ccmining.uky.edu.

Methods: The input for the bacterial infection domain was a set of names of 50 infection causing microorganisms and 72 drugs with therapeutic relevance in the treatment of those infections. We passed each of these names through MetaMap with the *term processing* option to identify concept unique identifiers (CUIs). Next, using a disjunctive query of all synonymous English names of these CUIs from the UMLS Metathesaurus, we retrieved 215,000 citations with publication type 'case reports' or 'clinical trial' from PubMed. The disjunctive query is also run on clinical trials made available through http://clinicaltrials.gov/ which resulted in 8000 trials. We extracted all named entities using MetaMap from these two sets of documents and indexed them using Apache Solr for entity based search. Specific semantic types were chosen to restrict search based on drugs or microorganisms. Although PubMed facilitates MeSH based query expansion, here we also provide search based on named entities from title and abstract. An important set of core concepts in Pharm.D programs involves drug properties and their interactions with disease mechanisms. Hence finding cases based on properties of a drug such as drug targets, mechanisms of action, volume of distribution, solubility, and bioavailability is an important requirement for finding interesting clinical cases. However, drug properties are often not mentioned in the case reports or clinical trial descriptions although the names of the drugs are used. To overcome this, we established an indirect connection between drug properties and clinical cases based on the particular drug names mentioned in the case narratives. We noticed that even structured versions of drug databases such as DrugBank capture drug targets, mechanisms of action, and other descriptive properties using free text. Even numerical properties such as volume of distribution, solubility, and protein binding follow non-uniform notation and are essentially captured as strings. Using a set of expert selected semantic types, we extracted named entities from the target and mechanism of action text snippets in DrugBank for each of the drugs. The numerical properties were manually curated by domain experts. The drug property database thus curated was used to link properties to particular cases that employ the corresponding drugs.

Discussion and Future Enhancements: Currently the system allows users to quickly search for cases based on named entities or properties of drugs used in the cases. Users can also annotate the results with custom concept tags. Search results are displayed in a faceted browsing interface with MeSH terms and user tags as the facets. Five faculty members from the UKY College of Pharmacy are assigning tags that correspond to core concepts in their teaching activities. Future goals are to build an ontology of core concepts based on the tags, incorporate 'related cases', quality metrics (e.g., impact factors), and also expand the corpus and drugs beyond bacterial infections domain.

The Role of Big Data in Community-wide Population Healthcare Delivery and Research

Hadi Kharrazi, MD PhD[1]
[1]Johns Hopkins School of Public Health, MD

Introduction

The emergence of big data in healthcare is inevitable. The growing demand for big data has been empowered by major improvements in data sharing, steady rise of Electronic Health Record adoption, and supporting federal policies and incentives. The appetite for big data has already reached community-wide population health initiatives partly due to the recent roll out of Accountable Care Organizations (ACO). These policies and technological advancements have created a unique 'landscape' to draw upon big data for a 'greener' population health delivery system.

This viewpoint presentation attempts to expand on the challenges and opportunities that Johns Hopkins Medical Institute (JHMI), as an academic medical center, is facing to utilize existing and potential sources of big data to enhance its community-wide population health. These challenges and opportunities can be applicable to other academic medical centers in which big data will play a key role in the future of evidence-based community-wide healthcare delivery.

Methods

This presentation reviews a list of infrastructures, centers, projects, and initiatives that JHMI has established to propel the use of big data in improving the population health of their patient community. This list is then mapped against potential sources of Big Data in population health management to identify gaps and opportunities that are applicable to other academic medical centers as well. Furthermore, a conceptual framework is proposed and then contextualized for JHMI to identify stages involved in translating Big Data into effective population healthcare policies.

Direct implications of Big Data initiatives at JHMI includes the neighboring communities (i.e., surrounding zip codes) and translated through various population healthcare delivery mechanisms (e.g., Johns Hopkins Community Health Partnership and the Johns Hopkins Coordinated Antenatal Service Enhancement). Indirect implications of this review may surpass JHMI's patient communities as other academic medical centers may adapt similar Big Data population health frameworks and amend existing or potential gaps.

Results / Principal Findings

After applying the proposed conceptual big data population health framework at JHMI, the following items were identified as critical to overcome the barriers in translating big data into community-wide outcomes: (a) increasing the coordination of activities among all centers and institutes through a 'Population Health Big Data Committee' operating under the ACO management; (b) aligning incentives for participating units, centers and institutions to form a big data eco-system; and, (c) promote the culture of sharing in which population-level clinical data can be accessed and studied by researchers for the purpose of quality improvement, while research findings can be swiftly translated into population solutions and deployed back into operations.

Discussion / Conclusion

The federal government has supported the big data movement in various domains, including healthcare, through a number of policies. Federal government has also supported the culture of open data and empowered decision makers to share big data while assuring privacy and security concerns. Probably the most impactful federal policies and legislations that have spurred the use of big data in community-wide population-health are initiatives that have no direct aim at big data. Federal government can empower the use of big data for improved population health by: (a) Funding dedicated projects and/or open calls for translational population health research using big data; (2) Expanding the current funding opportunities such as CMMI to include the use of big data for community-wide population health; (3) Establish an 'Office of Big Data' with various committees, including population health, at ONC to coordinate big data efforts in healthcare; (4) Incorporating community-wide population health measures in the future stages of Meaningful Use.

States have less control on the overall big data policies; however, various healthcare delivery and payment models can change the landscape of big data in population health. For example, in Maryland, the All-Payer model and state-wide PCMH have empowered providers to work with the HIE to improve population health. Other State initiatives such as Health Enterprise Zoning and Community Integrated Medical Home can stimulate the use of big data in population health as well. Results of such big data initiatives will eventually change the outcomes of population health research in each State which in turn can potentially lead into changes in local healthcare policy, statues and regulations.

Academic medical centers have the opportunity to lead the Big Data movement in the population health domain.

Creating, Maintaining and Publishing Value Sets in the VSAC

Emir Khatipov[1], Maureen Madden[1], Pishing Chiang[1], Philip Chuang[1], Duc Nguyen[1], Ivor D'Souza[1], Rainer Winnenburg[1], Olivier Bodenreider[1], Julia Skapik[2], Rob McClure[1,2], Steve Emrick[1]

[1]National Library of Medicine (NLM®), National Institutes of Health (NIH), Bethesda, MD, USA; [2]Office of the National Coordinator for Health Information Technology (ONC), Washington, DC, USA

The Value Set Authority Center (VSAC, https://vsac.nlm.nih.gov/) is developed by the NLM in collaboration with ONC and Centers for Medicare & Medicaid Services (CMS). VSAC provides access to value sets that are used to define concepts used in clinical quality measures and to support effective health information exchange and many other biomedical informatics applications and programs. VSAC has fulfilled the immediate need for a comprehensive resource that supports the creation and maintenance of value sets used by data elements in 2014 electronic Clinical Quality Measures (eCQMs). It currently continues to expand its repository beyond the Meaningful Use (MU) domain into such areas, as Patient Assessment Instruments, Common Data Elements for research, public health, the ONC S&I Framework, and other clinical modeling efforts.

In October 2013, NLM launched the VSAC Authoring Tool that allows authors to create, edit, clone, update and publish value sets. VSAC also provides data integration with the CMS Measure Authoring Tool (MAT) via a REST Application Programming Interface (API) that uses the *Integrating the Healthcare Enterprise (IHE) Sharing Value Sets (SVS)* specification.

The VSAC Authoring Tool features the following major characteristics, functionalities and capabilities:

- Robust workflow with Authors and Stewards as major players that perform specific functions via sets of tiered permissions and check points. Essentially, Authors perform editing functions, whereas Stewards approve the work of the Authors and submit value sets for publication.
- Users can create extensional value sets, which are sets of codes and terms derived from a single code system, and grouping value sets that represent one or more extensional value sets grouped together based on a common purpose of use.
- Users can search codes, as well as value sets containing specific codes and keywords, and add those to the value sets. Codes can also be imported in batch or through manual input, in which cases the system automatically validates them and provides user feedback when it detects problems.
- Definition of value sets are captured via specific metadata elements describing purpose of use and inclusion/exclusion criteria for the values.
- Value set definition can be applied to any version of a code system in the VSAC. This process is called expansion. VSAC generates and makes available to users expansion profiles that are used to create and publish code system version-bound value sets.
- VSAC can create program-specific expansion profiles that include mandated code system versions, built-in code validation rules and lists of excepted legacy and provisional codes. These profiles are used, e.g., for packaging annual MU value set releases in support of eCQM development process.

In the future, VSAC will provide an interface for authors to create intensional value set definitions. Such definitions will contain the rules that will allow the system to automatically remove or replace values to reflect code system updates, as well as to discover new concepts that are appropriate for inclusion into value sets. This will significantly ease the maintenance burden on value set authors.

Access to VSAC requires a free Unified Medical Language System® (UMLS®) Metathesaurus License, due to usage restrictions on some of the codes contained in the value sets. To author value sets, users need to obtain permissions by contacting VSAC's Support Group (http://www.nlm.nih.gov/research/umls/support.html). To receive important VSAC announcements, users can subscribe to the VSAC Updates e-mail list (https://list.nih.gov/cgi-bin/wa.exe?A0=nlm_vsac_updates).

Acknowledgments: This work was supported in part by the Intramural Research Program of the NIH, NLM.

A Direct Query Mechanism for Exchanging Quality and Performance Measures: A Proof of Concept with California Health Plans and Physician Organizations

Katherine K. Kim, PhD, MPH, MBA[1], Brian Goodness[2], Holly C. Logan[3], MA, David A. Minch, BS, FHIMSS[4], Dolores Yanagihara, MPH[3]
[1]University of California Davis, Sacramento, CA; [2]Integrated Healthcare Association, Oakland, CA; [3]San Francisco State University, San Francisco CA; [4]Healthshare Bay Area, San Francisco, CA

Introduction: Clinical data has become a necessary foundation for enabling new models of healthcare delivery that support efficient, high quality, and patient-centered care. These data are equally important for aligning reimbursement and incentives and claims data may be inadequate for these purposes [1]. Physician organizations (POs) and health plans (HPs) have access to different types of clinical data and they need an efficient mechanism for sharing these data with their partners. Integrated Healthcare Association (IHA) which manages the statewide California pay for performance incentive program, initiated a proof of concept project (POC) to assess whether a direct query architecture could be used to automate the process of gathering and analyzing quality improvement/performance measurement data from POs and HPs periodically, when they are timely and relevant, rather than many months after the fact.

Methods: The goal of the POC was to utilize a direct query system to collect and exchange data for CMS Medicare Stars outcome measures for analyses involving two use cases: 1) Compiled measure results (numerators and denominators) from POs and HPs to IHA for the purpose of performance measurement and public reporting, and 2) Compiled patient-level data to allow tracking of individual patient status for quality improvement and provide input to the aggregated Medicare Stars quality measures, including pharmacy and readmission data from HPs to POs and laboratory, vital signs and other clinical data from POs to HPs. Two POs and two HPs participated with IHA in a five-month long POC. The Global Information Network Architecture (GINA), an advanced object-oriented data management technology solution for cross-entity interoperability on a peer-to-peer basis from Creek Technologies, Inc. was used to securely exchange data between multiple configurations of participant organizations [2, 3]. The evaluation included construction of the data model, performance testing, and exchange results.

Results: Patient-level lab results, vital signs, and prescription fills were successfully exchanged between one PO-HP pair. Test data were exchanged between other PO-HP pairs. Measures were delivered from two POs and two HPs to IHA. Key lessons were learned regarding: 1) the lengthy lead time required to build confidence in the privacy and security assurance mechanisms as well as to sign legal agreements, 2) slow performance encountered when transmitting large identified data sets due to limitations in the technology stack used, 3) improving the indexing of source databases to hasten record identification, and 4) improving system documentation and technical project management tools for larger scale rollout.

Discussion: This project was championed by executives at each organization and resources were dedicated to accomplishing it in on a very truncated schedule. We successfully demonstrated a proof of concept that the direct query infrastructure was operational, that the data model was implemented, and that verified exchange occurred between the appropriate paired partners. However, a limited number of records were exchanged. Continued effort is required to implement a fully operational and scalable system for statewide exchange that demonstrates value as a mechanism for exchanging both calculated performance measures and patient-level data for quality improvement.

References

1. Robinson JC, Williams T, Yanagihara D. Measurement of and reward for efficiency In California's pay-for-performance program. Health Aff (Millwood). 2009;28(5):1438-47.
2. Tudor R. The Solution to Providing Information Technology to Improve Healthcare for Americans 2013 [updated March 21, 2013; cited 2014 March 12]. Available from: http://www.creek-technologies.com/.
3. Busalacchi F. The Solution to Providing Truly Integrated Solutions to Large Scale Interoperability 2013 [updated April 21, 2013; cited 2014 March 12]. Available from: http://www.creek-technologies.com/.

Medication Prescription Status Classification in Clinical Narrative Documents

Youngjun Kim, MS[1, 3], Jennifer Garvin, PhD, MBA[2, 3], Julia Heavirland[3], Jenifer Williams[3], Stéphane M. Meystre, MD, PhD[2, 3]
[1]School of Computing, [2]Department of Biomedical Informatics, University of Utah; [3]VA Health Care System, Salt Lake City, Utah

Abstract: *Classifying what the status is of medication diagnosed for heart failure patient is one of the key tasks needed to assess whether a given patient's clinical record substantiates the proper treatment. To determine the prescription status of each medication (i.e., active, discontinued, or negative), we implemented a SVM classifier with lexical features and achieved good performance, reaching 95.49% accuracy, in a five-fold cross validation evaluation.*

Introduction: Heart Failure (HF) is a common but serious medical condition and patients with a decreased systolic function should be treated with medications such as Angiotensin Converting Enzyme Inhibitors (ACEI) or Angiotensin II Receptor Blockers (ARB) to improve symptoms and prevent progression of HF. To help assess HF treatment at the VHA (Veterans Health Administration), the ADAHF (Automating Data Acquisition for Heart Failure) project included the development of an application to automatically extract HF treatment performance measure information from clinical notes. This application (called CHIEF) included the extraction of ACEI and ARB medications, as well as the classification of their prescription status: *active* (patient currently takes the medication), *discontinued* (patient remains off the medication or is temporarily taken off the medication), and *negative* (the medication does not pertain to the patient or is negated). Such status is crucial for accurate performance measure classification.

Methods: As part of the ADAHF project, we randomly sampled 3,000 clinical notes from our training corpus. These notes were manually annotated and included 6,007 medication annotations (4,911 ACEIs and 1,096 ARBs). Among annotated medications, 4,491 (74.76%) were *active*, 1,191 (19.83%) were *discontinued*, and 325 (5.41%) were *negative*. We built a linear Support Vector Machine classifier based on earlier work by Kim and colleagues [1] with lexical features (medication name, five words preceding it, and two words following it) to automatically determine the prescription status of ACEI and ARB medications.

Results: We used a five-fold cross validation with annotated medications to measure performance of medication prescription status classification. The overall accuracy reached 95.49%. Precision of each status was above 90%, and recall of the *discontinued* status was 86.23%. Recall was higher than precision with the *negative* status, even though they corresponded to only 5.41% of the annotated medications in our corpus. A total of 230 (71 + 159) *active* or *discontinued* cases were misclassified as the other class.

Table 1. Medication prescription status classification results

	Classified as			Count	Recall (%)	Precision (%)	F_1-measure (%)
Status	Active	Discontinued	Negative				
Active	4403	71	17	4491	98.04	96.26	97.14
Discontinued	159	1027	5	1191	86.23	92.94	89.46
Negative	12	7	306	325	94.15	93.29	93.72
Overall	4574	1105	328	6007	95.49	95.49	95.49

Conclusion: Our work shows that the prescription status of each medication can be successfully classified with lexical features extracted from the words surrounding medication names. The prescription status determined by our classifier can also be used for extracting reasons not to administer ACEI or ARB medications and evaluating if HF patients have benefited from appropriate care.

Acknowledgments: This research was supported by VA HSR&D IBE 09-069 (ADAHF) and by HSR&D HIR 08-374 (Consortium for Healthcare Informatics Research) and HIR 09-007 (Translational Use Case – Ejection Fraction). The views expressed in this article are those of the authors and do not necessarily represent the views of the Department of Veterans Affairs or those of the University of Utah.

References

1. Kim Y, Riloff E, Meystre SM. Improving classification of medical assertions in clinical notes. In Proceedings of the 49th ACL/HLT: short papers. 2011; 2:311-316.

Automated Phenotype Detection Facilitates Data Warehouse Analysis; an Example Using Structured Query Language and Exudative Pleural Effusion

John Kimbrough, MD, PhD[1], Vojtech Huser, MD, PhD[2], James J. Cimino, MD[1,2]
[1]National Library of Medicine, National Institutes of Health, Bethesda, MD
[2]Laboratory for Informatics Development, NIH Clinical Center, Bethesda, MD

Abstract
To explore automated screening methods that could be used in healthcare data warehouse analysis, we modelled a clinical phenotype in Structured Query Language (SQL). As a case study, diagnostic criteria for an exudative pleural effusion were used as the basis of an SQL modelled phenotype to screen a dataset of intensive care unit patients. This approach detected cases beyond those documented in provider notes, and may be incorporated into clinical decision support and similar projects.

Introduction: Data warehouses have created opportunities to pursue studies with advantages such as reduced risk and cost savings. Studies of warehouse data usually require analysis of information originally collected for another purpose, requiring the data to be screened for the condition of interest to the current project. We chose to model exudative pleural effusion because explicit diagnostic criteria are available and because identifying an effusion as an exudate suggests diagnoses such as malignancy (whereas effusions that are not exudative are more common in conditions such as heart failure). We used the Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II)[1] database because it is accessible to external researchers and has more than 32,000 patients.
Methods: A meta-analysis that reviewed methods of pleural effusion classification reported that multiple rules were effective, but there were no overall best criteria[2]. Therefore, we used a 3-test rule supported by the meta-analysis that includes pleural fluid cholesterol, lactate dehydrogenase and protein measurement to create an SQL query that maps the rules of the test to the terms of the database. Any fluid with a lab value exceeding the test's threshold is classified an exudate. In order to evaluate our SQL-based phenotype, we created an evaluation subset of 100 randomly selected patients that had the keyword "exudate" in provider notes. We used automated review of lab results combined with manual review of provider notes to determine all true cases of exudative pleural effusions.
Preliminary Results: In the evaluation set, the manual review of provider notes detected 13 cases, recall was 81% and precision was 100%. In the same subset, the SQL phenotype detected 15 cases and was associated with a recall of 94% and a precision of 100%. The lack of a traditional gold-standard limits the interpretation of the calculations, and therefore discrepancies between automated and manual approaches were examined. There were three cases detected by the SQL phenotype that were not diagnosed in the provider notes, because these cases satisfy laboratory criteria one interpretation is the diagnosis were absent from provider notes because they were clinically missed (i.e. the provider notes had 3 false negatives). In the one case documented in provider notes and not detected by SQL phenotype, the providers based their conclusion on a serum (not pleural) laboratory result.
Discussion and Conclusions: Our approach has the advantages of automation and explicit, updatable rules. We uploaded the phenotype to PheKB (development status) to facilitate cross-institutional sharing. The motivation for our study was to demonstrate feasibility of clinician-authored SQL phenotypes. This approach may be used in circumstances beyond clinical studies such as responding to drug safety alerts. As newly discovered issues, these alerts often address problems not included in previous encounters, confounding searches that rely on provider note content. Manual review may lack the efficiency desirable in responding to a safety alert. These warnings often include a succinct list of explicit concerns, and therefore may be potentially adapted into queries that screen coded information in a relation database. Approaches such as the SQL phenotype we wrote may also serve as triggers for clinical decision support, quality studies and screening for point-of-care enrollment in clinical trials.

References

1. Saeed M, *et. al.,* Multiparameter intelligent monitoring in intensive care II (MIMIC-II): A public-access ICU database. Critical Care Medicine 2011;39(5):952-960.
2. Heffner JE, *et. al.,* Diagnostic Value of Tests That Discriminate Between Exudative and Transudative Pleural Effusions. Chest 1997; 111:970-980

Identification of Inflammatory Bowel Disease Patients with Steroid-induced Diabetes Mellitus Using an Electronic Health Record

Sivan Kinberg, MD, Lyudmila Ena, MA, Herbert Chase, MD, Carol Friedman, PhD
Department of Biomedical Informatics, Columbia University, New York, NY

Abstract

Glucocorticoid toxicity is one of the most common causes of iatrogenic illness associated with chronic inflammatory diseases, such as inflammatory bowel disease (IBD). The aim of this study was to use an informatics approach to determine the prevalence of and risk factors for steroid-induced diabetes mellitus (S-DM) and steroid-induced prediabetes (S-PDM) in patients with IBD.

Introduction

Glucocorticoids are commonly used to treat patients with IBD, often for prolonged periods of time. Although glucocorticoids have potent anti-inflammatory effects, they have been implicated in the development of S-DM and hyperglycemia in patients with various conditions.[1] Hyperglycemia, even transient, has been associated with adverse events.[2] Conversely, control of hyperglycemia during acute illness has been associated with improved outcomes.[3] Despite this, S-DM and S-PDM are often under-diagnosed and under-treated. In patients with IBD, the current knowledge of S-DM and S-PDM is limited. Using automated extraction of data in the electronic health record (EHR), we determined the prevalence of and risk factors for S-DM and S-PDM in patients with IBD.

Methods

This was a retrospective cohort study that examined patients with IBD who were treated with glucocorticoids between 2004 and 2012, excluding patients with pre-existing diabetes mellitus. Notes from the EHR were parsed using a natural language processing system[4] to encode the information, and the coded output was stored in a structured database where the relevant data were queried. We identified patients with IBD based on the intersection of ICD-9 codes and coded information from the notes. The diagnoses of S-DM and S-PDM were based on ICD-9 codes and on laboratory data consistent with the American Diabetes Association diagnostic criteria.[5] Patients with IBD treated with glucocorticoids were compared with IBD patients who never received glucocorticoids.

Results

A total of 1,719 patients with IBD were identified based on the intersection of ICD-9 codes and coded information in the notes, compared with 7,126 patients based only on ICD-9 codes and 10,812 patients based only on the notes. We found that 140 (20.1%) of 698 patients with IBD treated with glucocorticoids developed S-DM compared with 21 (5.8%) of 363 patients not treated with glucocorticoids (OR=7.42, 95% CI: 4.41-12.48). Prediabetes was identified in 192 (27.5%) IBD patients treated with glucocorticoids and in 67 (18.5%) patients not treated with glucocorticoids (OR=2.25, 95% CI: 1.59-3.17). Multivariable analysis determined increasing age, obesity and parenteral nutrition as risk factors for S-DM, and male sex and parenteral nutrition as risk factors for S-PDM.

Discussion and Conclusion

To our knowledge, this is the first study to determine the prevalence of and risk factors for S-DM and S-PDM in patients with IBD. Determining which patients had IBD was challenging. Initial manual review of patients identified based only on ICD-9 codes or only on coded information in the notes showed that some were inaccurate. Therefore, in order to increase precision, we identified IBD patients based on the intersection of ICD-9 codes and codes from the notes. We found that patients treated with glucocorticoids were 7.42 times more likely to develop diabetes mellitus and 2.25 times more likely to develop prediabetes than patients treated with non-glucocorticoids. Using the EHR for automated detection of high-risk patients is possible and could result in earlier diagnosis, timelier treatment, and possibly improved outcomes.

References

1. Perez A, Jansen-Chaparro S, Saigi I, Bernal-Lopez MR, Miñambres I, Gomez-Huelgas R. Glucocorticoid-induced hyperglycemia. *J Diabetes*. 2014;6(1):9-20.
2. Umpierrez GE, Isaacs SD, Bazargan N, You X, Thaler LM, Kitabchi AE. Hyperglycemia: an independent marker of in-hospital mortality in patients with undiagnosed diabetes. *J Clin Endocrinol Metab*. 2002;87(3):978-82.
3. van den Berghe G, Wouters P, Weekers F, et al. Intensive insulin therapy in critically ill patients. *N Engl J Med*. 2001;345(19):1359-67.
4. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. *J Am Med Inform Assoc*. 2004;11(5):392-402.
5. American Diabetes Association. Diagnosis and classification of diabetes mellitus. *Diabetes Care* 2013.36(Suppl. 1):S67–S7.

User Experiences of Speech Recognition Technology (SRT) by Physicians: A cross-sectional survey study

Joshua L. King, MHA[1], Martina A. Clarke, M.S.[2], Min Soon Kim, PhD[1, 2]
[1]Department of Health Management and Informatics, University of Missouri, Columbia, MO;
[2]Informatics Institute, University of Missouri, Columbia, MO

Abstract

In order to evaluate the user experience and utilization of speech recognition technology (SRT) by physicians for medical documentation, 71 physicians, in two Missouri hospitals, completed a 10 item survey, followed by an in-depth interview with a CMO. Survey on user experience showed 42 (59%) of participants experienced spelling and grammatical errors and 15 (21%) encountered clinical inaccuracy. CMO agreed that SRT somewhat disrupts clinical workflow and produced many errors that needed to be addressed.

Introduction

SRT implementations have been being actively sought after by healthcare organizations for reduced report turnaround time and financial savings. However, it will not be an effective tool if the reports created by SRT are inaccurate, poorly organized, or untimely. The objective of this project is to evaluate the user experience and utilization of (SRT) by physicians for medical documentation in two hospitals in Missouri.

Method

A quantitative survey tool was utilized to collect data related to a physician's area of practice, electronic equipment utilized for documentation created after providing care and overall experience and satisfaction of SRT. The sample of physicians was selected from one urban and one rural, tertiary care hospitals because of the institutions' pro-active adoption of SRT and willingness to participate in the study. In addition, a qualitative in-depth interview was conducted with a Chief Medical Officer (CMO) to discuss issues in implementation, training, choice of SRT, and perspective outcome from leader's perspective.

Results

Seventy-one (60%) of the anticipated 125 surveys were returned. Sixteen (23%) of participants were involved in Internal Medicine, 9 (13%) in Family Medicine and the additional 46 (64%) were spread throughout many varying specialties in medical practice. Fifty-six (79%) of participants utilized a desktop and 14 (20%) used a laptop. Windows operating system was utilized by more than 58 (82%) of the survey respondents. Nuance (Burlington, MA) products are the dominant brand, providing continuous SRT that is utilized by the participants with 59 (83%). With regard to user experience, 42 (59%) of participants experienced spelling and gramatical errors and 15 (21%) encountered clinical inaccuracy, which could potentially affect patient care. Accuracy is defined as the percentage of correct transcription;if user needs are not met in a timely manner, overall user satisfaction may decline as a result. When physicians were asked how satisfied they were with SRT, some respondents were satisfied 34 (48%) rather than dissatisfied 20 (28%) with SRT overall.

Recommendations made by participants on improving user experience include: 11 (15%) better interpretation of the spoken word, such as, foreign accent, unusual words, etc.; 9 (13%) wanted improved SRT speed because they found SRT to be too slow; 8 (11%) desired improved accuracy; 7 (10%) needed more extensive training of the system, such as voice training. During a qualitative interview with the CMO, he commented on issues in implementation, training, choice of SRT, and perspective outcome from leader's perspective. He noted that current SRT was somewhat disruptive to the clinical workflow and produces many errors that need to be addressed. The CMO finally suggested from the leader's perspective that , although not perfect, SRT has a potential for succesful and economical electronic record keeping when and if the performance of products is improved to achieve reliable service.

Conclusion

Even though SRT is used in healthcare settings to assist in the completion of medical record documentation, this study identified critical issues of inconsistency, unreliability, and dissatisfaction in functionality and usability of SRT. Although SRT has yet to reach the accuracy rate desired by physicians, it is a promising system that could potentially improve the clinical workflow by supporting the efficient creation of EHR documentation. Additionally, future use of SRT will be impacted by the ability to improve accessibility through integration into mobile platforms. Limitations of this study include a small sample size in two hospitals which limits transferability, which merit further attention be provided to the improvement of SRT's functionality and usability within varying healthcare settings.

Enabling Patient-Centric Comparative Effectiveness Research in i2b2

Jeffrey G. Klann, PhD[1,2,3]; Lori C. Phillips, MS[1]; Kenneth D. Mandl, MD, MPH[2,4]; Shawn N. Murphy, MD, PhD[1,2,3]

[1]Partners Healthcare, Boston, MA; [2]Harvard Medical School, Boston, MA; [3]Massachusetts General Hospital, Boston, MA; [4]Boston Children's Hospital, Boston, MA

Background - Informatics for Integrating Biology and the Bedside (i2b2) is an open-source clinical data warehousing and analytics platform funded by the National Institutes of Health. [1] It is used at over 100 sites nationwide, and multiple federated networks exist across i2b2-equipped institutions. It is also used by a third of the sites in the Patient Centered Outcomes Research Institute's Patient-Centered Clinical Research Network (PCORnet). PCORnet is a national effort to instantiate a national 'network of networks' that supports large-scale comparative effectiveness research. [2]

i2b2 uses a flexible database schema and user-defined information models, designed to create minimal friction for implementers when importing data from various sources. However, this flexibility impedes interoperability; exchange of data across i2b2 sites and across PCORnet networks requires agreement on an information model. PCORI envisions a highly interoperable network that will enable nationwide disease surveillance and data exchange between networks.

Therefore we have developed an approach for creating standardized i2b2 Information Models and mapping local data to them, all using existing ontology services in the i2b2 system. Our approach complements the existing i2b2 networking system (the Shared Health Research Information Network - SHRINE) [3], and we have implemented our PCORnet network's SHRINE with this new approach. Additionally, we have developed a tool to physically transform conformant i2b2 instances into the schema structure PCORnet requires for cross-network queries. More information and the tools themselves will be made available through our network's blog. [4]

System Description – We developed a common i2b2 ontology (Information Model) that instantiates the PCORnet Common Data Model. To map local data to this common ontology, we relied on two insights about i2b2's design. First, i2b2 ontology elements define both a unique pathname (e.g., \\PCORNET\DIAGNOSIS\DX\09\250\) and a unique code (e.g., ICD9:250). Pathnames are consistent across all implementations, but the codes are allowed to vary to match local data. Second, i2b2's query approach automatically includes ontology elements' children, so additional local codes can be included as children of a standard term and are automatically gathered when querying standard codes. We updated our mapping tools to support this approach. Our sites have successfully used these tools to map their data, and we have validated their correctness through comparisons of query results to database content.

To transform data into the PCORnet data format, information from the path is used to determine the destination location (e.g., table DIAGNOSIS, column DX) and 'virtual queries' are performed to populate the target database.

Discussion – This i2b2 mapping approach enables the popular i2b2 clinical data analytics platform to support standard, interoperable information models without altering the local data. We are using this at the sites in our PCORnet network to enable information exchange and comparative effectiveness research, and it will soon be used to create data marts in the PCORnet schema format for nationwide research. In the future, we hope to extend our tools to support transformation of i2b2 data into other standard formats for exchange with patients and other PCORnet networks.

References

1 Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, Kohane I. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Inf Assoc* 2010;**17**:124 –130. doi:10.1136/jamia.2009.000893

2 Patient Centered Outcomes Research Institute. The National Patient-Centered Clinical Research Network. PCORnet. Dec 17, 2013. http://pcornet.org (accessed 7 Jan2014).

3 McMurry AJ, Murphy SN, MacFadden D, Weber G, Simons WW, Orechia J, Bickel J, Wattanasin N, Gilbert C, Trevvett P, Churchill S, Kohane IS. SHRINE: Enabling Nationally Scalable Multi-Site Disease Studies. *PLoS ONE* 2013;**8**:e55811. doi:10.1371/journal.pone.0055811

4 SCILHS Team. Scalable, Collaborative Infrastructure for a Learning Health System. 2014. http://scilhs.org/ (accessed 13 Mar2014).

Electronic Pharmacovigilance: Calling for Earlier Detection of Adverse Reactions (CEDAR)

Elissa V. Klinger, SM[1], Alejandra Salazar, PharmD[1], Japneet Kwatra, SM[1], Jeffrey Medoff, BS[1], Patricia Dykes, RN[1], Jennifer S. Haas, MD, MSPH[1,2], Mary Amato, PharmD[1], David W. Bates, MD, MS[1,2], Gordon Schiff, MD[1,2]

[1]Divison of General Internal Medicine and Primary Care, Brigham and Women's Hospital, Boston, MA; [2]Harvard Medical School

Introduction: The safe and effective use of prescription drugs in the outpatient setting is an ongoing concern that may be monitored by the development of novel automated approaches to obtain patient self-reported outcome information related to potential adverse drug reactions (ADRs) and respond to these concerns in real-time. To address the growing importance of integrating information about safety and effectiveness we developed a patient-reported, EHR-integrated interactive voice response system (IVRS) to actively monitor the safety and effectiveness of treatment for patients taking FDA-approved medications for one of four common chronic conditions (diabetes, hypertension, insomnia, depression), with integrated management support by a clinical pharmacist.

Methods: Using an integrated EHR we developed an algorithm to identify any adult primary care patient at one of 9 participating clinics who had been started on an oral agent of interest within the previous 2 weeks. All eligible patients were sent a letter with study information and an opportunity to opt-out of participation. Using automated IVRS phone calls, patients were called 4-6 weeks and again at 6-8 months following their medication start date and prompted to confirm that they were taking the drug and that it was prescribed for the condition of interest. A broad screen and then targeted questions documented both adherence and symptoms commonly associated with ADRs. Any patient with a positive symptom was transferred to a pharmacist in real-time at the conclusion of the data-capture call; these data fired a real-time alert to the pharmacist containing relevant clinical information including possible ADR and pre-populated a database to drive and document the intervention. Patients reporting no symptoms could also opt to speak with the pharmacist in real-time or at their convenience. Any patient counseled by the pharmacist had a note filed in the EHR, with more urgent clinical concerns triaged accordingly.

Results: IVRS calls began in June 2013. To date we have identified and called 2,810 eligible patients prescribed one of 103 target medications. Of these, 414 (14.7%) have participated in the IVRS interviews, 228 (8.1%) have actively declined participation, 1464 (52.1%) have answered but not completed the call, and 704 (25.1%) have not answered any call attempt. Of those participating (n=414), 143 patients (5.1% of all patients called and 34.5% of patients participating in the screening) have reported symptoms consistent with ADRs and 231 (8.2% of all patients called) have completed successful transfers to the pharmacist and had notes filed in their charts and their PCPs alerted accordingly.

Discussion: We demonstrate that an EHR-integrated IVRS that leverages real-time data capture and live transfer to a pharmacist can be an effective tool to reach out in a proactive manner to patients starting new medications. Such a system allows a pharmacist to simultaneously access the EHR and speak with a patient to troubleshoot symptoms that may be consistent with ADRs, and to alert the care team in real time about possible clinical symptoms that may necessitate a change in dosage or a drug discontinuation. For patients not experiencing symptoms, the platform may provide a service to individuals who have questions about their medications or struggle with adherence.

Relating Health Concerns and Goals in Interprofessional Care Planning

Stephanie Klinkenberg-Ramirez[1], Kira Tsivkin[1], Perry L. Mar, PhD[1,2,3], Hari Nandigam MD, MSHI[1], Dina Iskhakova[1], Roberto A. Rocha MD, PhD[1,2,3], Sarah Collins, RN, PhD[1,2,3]

[1]Partners Healthcare System, Wellesley, MA; [2]Brigham and Women's Hospital, Boston, MA; [3]Harvard Medical School, Boston, MA

Abstract

Coordinated care planning is associated with better patient outcomes. We investigate how interprofessional care planning can move from indexing to associating clinical concepts salient for care planning across the continuum of care and between health professions. This work evaluates terminology management approaches and a conceptual schema for interprofessional care planning by developing and validating clinical care planning scenarios with subject matter experts. We identified "goals" as a terminology construct commonly used among clinicians.

Introduction

Team-based care planning requires individual team members to establish common ground and goals with each other to reconcile their respective plans of care and treatment plans into one patient-centered care plan. This reconciliation of plans is challenging to achieve via shared documentation tools used by multiple clinical professions because of their distinct and overlapping terminology needs.[1] Goals are inconsistently documented and lack linkages indicating relations to each other. The aim of this poster is to understand the common care planning concepts and their associations, with a particular focus on goals that may be leveraged for care planning reconciliation by analyzing 4 clinical scenarios validated by a set of health professionals that work as a team in an outpatient setting.

Methods

We developed four clinical scenarios for the outpatient setting, representing care for patients with the following conditions: uncontrolled diabetes mellitus type I, congestive heart failure related to myocardial infarction, diabetes mellitus type II and uncontrolled depression, and immune-mediated kidney failure. Clinical scenarios contained assessment data, past medical history, and health concerns, interventions and goals identified by a physician, care coordinator, social worker and pharmacist. We conducted individual 1-1.5 hour interviews with subject matter experts, including a primary care physician, two care coordinators, a psychiatrist, and a pharmacist, during which they validated the scenarios for clinical accuracy with emphasis on care planning concepts. Requirements for a corresponding conceptual model and terminology needs were identified.

Results

Health professionals identified distinct concerns and interventions related to their profession or specialty area. For example, the physician identified the concern of major depressive disorder and prescribed antidepressants and a referral to a social worker. The social worker's concerns included inadequate social support and requested a referral to psychotherapy. Within the team, individuals communicated to facilitate interventions identified by another clinician. For a patient struggling with adherence, the pharmacist recommended a new diabetes medication regimen for the physician to prescribe. Despite identification of distinct concerns and interventions, the health professionals identified common goals for the clinical scenarios that were either synonyms or hierarchically related as goals and sub-goals (see Figure 1). An information model was constructed to support goals and their relationships.

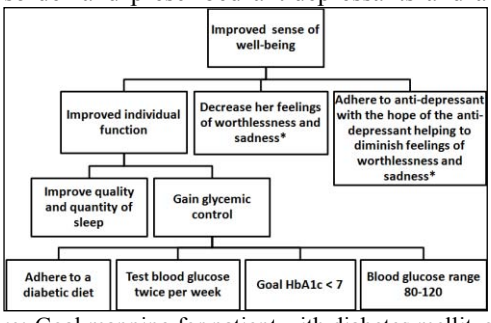


Figure: Goal mapping for patient with diabetes mellitus and uncontrolled depression. *synonymous goals from different clinical professions.

Conclusion

Our data indicate that related goals are a common construct for care planning in a team of health professionals who identify similar concerns and execute related interventions. Further work should validate this observation in other care settings and clinical scenarios and investigate how to link goals to concerns and interventions.

Acknowledgements: Project funded by a Partners-Siemens Research Council Grant (title: "*Knowledge Management Terminology Infrastructure to Support Interprofessional Plans of Care*").

References

1. Tsivkin, K. & Collins, S. Terminology Infrastructure to Support Interdisciplinary Plans of Care. in *AMIA Annu. Symp. Proc.* 1376 (2013).

Barriers and Benefits to CHT Use in Overweight and Obese Adolescents

Amy Knoblock-Hahn, Ph.D[1], Cynthia LeRouge, Ph.D[1], Shashank Jain[1], Kate Dickhut[1], Toree Malasanos, M.D.[2]
[1]Saint Louis University, St. Louis, MO; [2]Vheda Health, Baltimore, MD

Abstract

Adolescent overweight and obesity is a major public health concern, with many long-term implications for patients and the healthcare system. Consumer health technology (CHT) has the potential to support adolescents in their efforts to achieve and maintain a healthy weight. However, barriers and benefits to use of CHT must be addressed in the design of CHT, in order for the technology to actually be adopted by adolescents.

Introduction

Overweight and obesity in children and adolescents has reached an all-time high in the United States[1]. Multiple levels of influence from the socio-ecological model manifested in the Chronic Care Model (CCM) are thought to contribute to the increased prevalence of adolescent obesity. CHT tools may help close identified gaps in the socio-ecological ideal, but research in the use of CHT for obesity management in adolescent populations is lacking. The objective of this study is to examine the potential barriers and benefits to overweight and obese adolescents using CHT to self manage their condition within the context of existing CCM components.

Methods

A qualitative, multi-perspective, non-randomized research method was used. Focus groups and in-depth interviews with overweight and obese adolescents, pediatricians, and parents of overweight and obese adolescents were performed. The Unified Theory of Acceptance and Use of Technology (UTAUT) was applied as a means to categorize barriers and facilitators; the UTAUT model has been shown to account for 70% of the variance in intention to use technology[2].

Results

Adolescents, parents, and pediatricians expressed intent to co-use CHT applications. Primary barriers and facilitators to the use of CHT applications in overweight and obese adolescents were identified (Table 1).

Table 1: Intent to Use Adolescent Obesity Self-Management Applications by User Group

Construct	Barriers/Facilitators		Adolescent	Parent	Provider
Performance Expectancy	Food records & track physical activity	Facilitator	X	X	
	Meal planning & Portion Size Guidance	Facilitator	X	X	X
	Security/Privacy	Barrier		X	
Effort Expectancy	User-friendliness	Facilitator	X	X	
	Design optimized for adolescents needs	Facilitator	X		X
	Time constraints	Barrier	X		
Social Influence	Peer Networking	Facilitator	X		
Facilitating Conditions	Information on healthy behaviors	Facilitator	X		
	Multiple Technology Platforms	Facilitator	X	X	X

Conclusion

CHT tools for adolescent obesity self-management may help close identified gaps in ongoing treatment for obesity, but various barriers and facilitators need to be addressed for CHT to be effective.

References

1. Ogden, C. L., Carroll, M. D., Curtin, L. R., Lamb, M. M., & Flegal, K. M. (2010). Prevalence of high body mass index in US children and adolescents, 2007-2008. *Journal of the American Medical Association*, *303*(3), 242-249.
2. Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, *27*(3), 425-478.

Designing clinical models of EHR (Electronic Health Records) for long-term care providers to elderly persons

Shinji KOBAYASHI, MD, PhD[1], Naoto KUME, PhD[1], Tomohiro KURODA, PhD[2], Hiroyuki YOSHIHARA, MD, PhD[1]
[1]EHR Research Unit, Kyoto University, [2]Division of Medical Information Technology and Administration Planning

Abstract

Clinical information management for elderly people involves various stakeholders in local community, and sharing it among them is proposed as a promising way. We analyzed and designed standard XML formalism based on openEHR clinical archetypes models and MML (Medical Markup Language) to construct integrated EHR.

Introduction

To manage long-term care for elderly persons is one of the emerging issues in aging countries. Because elderly person tends to have multiple physical/social dysfunctions, they need various services mainly in health care for their lives. Sharing information among caregivers is considered as an effective way, so that we examined over our existing EHR implementation, and standardized information model were able to apply for this purpose. However, our EHR was targeted for mainly for hospitals/clinics in regional health care. We need to extend our EHR and standardize information models for long term care for elderly persons.

Methods

We had discussed the use cases with care providers and figured out it on mindmaps. The figured mindmap revealed that use cases had much diversity and included complexed information models. After we explored our development schedule of EHR and the priority, we determined to standardize flow sheet in daily care. While we review over flow sheet information model, vital sign needed to be a separated particle to utilize other modules. Flow sheet module was composed with four categorized items, such as 1) vital sign, 2), intake, 3) bodily output, 4) personal demographics. And the next step, we defined W3C XML schemas[1] on existing base, such as MML(Medical Markup Language)[2] and other health care standards. The openEHR clinical knowledge manager was helpful to examine clinical concepts[3]. During these processes we had implemented a sample application to show how they worked.

Results

Flow sheet and vital sign clinical information models had been published as candidates of official release as additional modules of MML. The models will be established as new modules in 2014. A group of care providers adopted these modules experimentally to their care houses. While discussion has been going on still now, we are implementing EHR modules for long-term care providers on existing regional EHR system. The open source development process has been helpful for discussion in detail, because we could prove the behavior of artefacts at the same time. However, we need more information models to implement EHR system for elderly persons. Evaluation for activity and dysfunction in daily life is a next problem, but the evaluation standards, such as WHO ICF include many subjective items hard to implement computational/interoperable metrics. More discussion with nursing professionals will be needed.

Conclusion

We had designed standard information modules for long-term care providers to elderly persons. For their use case, to record vital sign and daily flow-sheet are important. To make standard information model, mindmap and sample application were helpful to discuss on them. More clinical models need to be designed to evaluate activity or dysfunction of person in the next step for long term care EHR.

References

1. Extensible Markup Language (XML) 1.0 W3C Recommendation 10-February-1998, http://www.w3.org/TR/1998/REC-xml-19980210
2. Medlcal Markup Language version 3, http://www.medxml.net/E_mml30/mmlv3_E_index.htm
3. The openEHR project, Clinical knowledge manager, http://openehr.org/ckm/

A General Propensity Matching Algorithm to Control for Potential Confounders in Observational Studies using Outcomes Miner

Julianna Kohler MHS,[1] E Manigandan MPhil,[1] Gerardo Soto-Campos PhD,[1] Tanmay Gupta,[1] Santosh Narayanan,[1] Elisabeth L Scheufele MD, MS,[1,2] Matvey B Palchuk MD, MS[1,2]

[1]ConvergeHEALTH by Deloitte, Deloitte Consulting LLP, Newton, MA; [2] Harvard Medical School, Boston, MA

Abstract

We propose a general propensity matching algorithm to accommodate multiple outcomes observed in Outcomes Miner, a comparative analytics solution for observational studies. We categorized the approximately fifty available outcomes into groups of similar overarching concepts and used Maximal Information Coefficient to identify the variables most closely correlated with each group. After consolidating the variables across outcome groups, stepwise regression identified the final set of variables used for the propensity calculation.

Introduction

Outcomes Miner is a real world evidence-based solution that provides the user with observational study data in the form of a series of integrated and interactive reports that generate detailed insights and qualified hypotheses from comparative clinical studies. Users can explore hundreds of demographic and health-related variables and sub-groups based on those variables and their relationships to 50 previously established outcomes. One limitation is that, as with manually curated observational studies, the Outcomes Miner may also suffer from confounder bias. The typical solution to address this limitation is to develop a propensity matching algorithm, which equally distributes potential confounders across the treatment arms. To account for this limitation in the Outcomes Miner research platform, we propose a propensity matching algorithm that is developed and programmed into the research platform.

Methods

To develop the propensity matching algorithm, first, we sorted the 50 possible outcomes into five groups: “mortality outcomes,” “health system utilization outcomes,” “disease-specific health system utilization outcomes,” “laboratory biomarker outcomes,” and “clinical outcomes.” We selected a subset of outcomes from each group (e.g. number of emergency room visits from the health system utilization outcomes group), and we used MIC (Maximal Information Coefficient) to identify the 20 variables most highly correlated with each outcome from each group. Those 20 variables identified for each group were then consolidated into a single regression model. In turn, the single regression model was performed against the outcomes to identify the set of variables associated with the majority of outcomes, which are used in the final propensity model.

Discussion

With the availability of over 300 independent variables in the Outcomes Miner, the above-described propensity matching method produces a more robust algorithm that will more accurately distribute confounders across treatment groups while still taking into account the multiple outcomes of interest. Additionally, this approach takes into account unique relationships between variables and outcomes within the patient population. While the Outcomes Miner platform provides the end user with the ability to analyze up to 50 outcomes in a comparison study across a wide variety of sub-groups, the propensity matching algorithm remains constant regardless of the outcome considered. For the initial tests on the Outcomes Miner reports, we focused efforts on applying the platform to the clinical realm of diabetes. We established a first round propensity matching algorithm with 15 variables identified in the literature as associated with general morbidity and mortality for diabetes, which we have replaced with the above-described model.

References

1. Brookhart et al. Variable selection for propensity score models. Am J Epidemiol. 2006 June 15; 163(12): 1149–1156.
2. Onur Baser, PhD. Too Much Ado about Propensity Score Models? Comparing Methods of Propensity Score Matching. Value in Health. 2006; 9(6): 377-385.
3. http://www.converge-health.com/solutions/miner/outcomes-miner/. Accessed 02/28/2014.

Development of an Automated Cirrhosis-Associated Symptom/Finding Detection Tool

Jejo Koola[1,2], Robert Cronin[2], Ruth Reeves[1], Jason Denton[1] Samuel B. Ho[3,4], Michael Matheny[1,2]
[1]Tennessee Valley Healthcare System, Dept. of Veterans Affairs, Nashville, TN
[2]Vanderbilt University, Nashville, TN
[3] VA San Diego Healthcare System, San Diego, CA
[4] University Of California San Diego, San Diego, CA

Introduction

Chronic liver disease, whose prevalence reached 15% in the United States, causes 44,000 deaths annually. Despite emphasis on early detection and guideline appropriate management, twenty percent of cirrhotic patients are not identified until late complications occur. When advanced disease was identified, patients only receive 30-90% of guideline recommended measures. Informatics solutions, which may improve cirrhosis care, rely on structured data. Unstructured aspects, such as homelessness, alcohol use and psychiatric problems, are known to inform liver disease progression, but studies which incorporate them invariably use hand-curated records. This study pilots annotation of a cirrhosis document corpus.

Methods

We obtained 60 gastroenterology History and Physical (H&P) from a patient cohort hospitalized at the Dept. of Veterans Affairs (VA) MidSouth Health Network anytime between 2002 and 2011. Two physicians, with a third acting as an adjudicator, identified clinical characteristics known to be correlated with cirrhosis: hepatitis, alcohol abuse, drug abuse, marginal housing and homelessness, encephalopathy, ascites, palmar erythema, spider nevi, jaundice, hepatomegaly, splenomegaly, firm or nodular liver, and caput medusa. Additionally, annotators marked whether the document stated anything about cirrhosis itself. Annotators also marked level of assertion (positive, negative, and maybe) and time course of finding (past or present). We report prevalence of these features in our corpus and Inter-annotator Agreement.

In order to estimate information density for an information retrieval task, we used Yale cTAKES EXtension (yTEX version 0.8 using cTAKES version 2.5.0) to perform automated concept-level annotation and generate a collection of concept unique identifiers (CUI), mapped to the Unified Medical Language System (UMLS).

Results and Discussion

The prevalence of clinical characteristics within this document corpus ranged from 0.00 to 0.60. Three concepts (nodular liver, caput medusae, and spider nevi) were not identified. Clinicians most commonly documented presence or absence of organomegaly (0.43), encephalopathy (0.60), and social risk factors (0.37 for drug abuse and 0.25 for alcohol). Because only three of the notes came from patients with advanced liver disease, we are limited on identifying relevant physical exam findings. Though the VA strives to assist veterans with substance abuse issues, the corpus poorly documents alcohol and drug use. Cohen's Kappa ranged between 0.7 and 1.0 for most of the concepts. "Marginal Housing" served as the notable exception, with a Kappa of 0.3. Given the low prevalence of the target classes, simple percent agreement (which was > 95%) maybe more indicative than Cohen's Kappa. Processing with yTEX generated 8206 CUIs across the sixty documents. Of these, 1648 were distinct.

Several opportunities exist for improvement. First, we will need to enrich the document corpus with more clinical notes from cirrhotic patients. Second, CUIs overlap significantly. We are researching dimension reduction measures to cull redundant concepts. Nevertheless, this pilot suggests that key cirrhosis related concepts can be reliably identified within a clinical corpus.

Getting Past 10%: Employing a Successful Bar Code Environment for Patient Safety

Terese Kornet, MSN, RN; Denise Gilanelli, MSN, RN; Sean R Sarles, BA, BSN, RN, CCRN; David Stabile, MSN, RN; Gregory Watson, CPhT; Colleen Mallozzi BSN, RN, BSIS; Paul Miranda RPh, MBA; Michael Motto, MBA; Nishaminy Kasbekar, BS, PharmD, FASHP

Description

Our health system piloted a multi-hospital Bar Code Medication Administration (BCMA) project using our existing electronic medical record. Multiple disciplines, including Nursing, Respiratory Therapy, Pharmacy, and Information Services collaborated to create new workflows and training structures to support the BCMA project. The pilot project was deemed a success and plans to go 'enterprise-wide' were launched within months of the initial BCMA go-live.

Penn Medicine implemented BCMA as part of an initiative to meet Meaningful Use core measure requirements. The Meaningful Use (MU) Stage 2 objective is to "Automatically Track medications from order to administration using assistive technologies in conjunction with an electronic medication administration record" (CMS, 2012); however, only looking at the BCMA process in terms of MU requirements undermines the importance of patient safety as the primary motivator for implementing this technology. The multidisciplinary collaboration of the project team was the deciding factor in the successful implementation of BCMA.

Background

- Early studies quantified the extent to which errors occur at each of the stages of the medication-use process; one of the most troubling steps in the process is the administration phase, when 26% to 38% of the errors occur.[1]
- Further, it has been described that errors at the administration phase are likely to reach the patient since there are few safeguards in place to intercept the error before it is passed onto the patient. [2]
- A recent published study by Eric Poon and his colleagues at Brigham and Women's Hospital demonstrated a 41.4% relative reduction in medication administration errors using BCMA system. [3]
- Knowing that a systematic process has been put into place to ensure patients are getting the right medication, at the right time, in the right dose, and by the right route has resulted in more positive survey results from staff, patients and patient families. [4]

BCMA in Practice

Basic Steps:

1. **Scan** the 'pairing' bar code near the keyboard
2. Sign into SCM and enter the eMAR
3. Click on the <Bar Code> icon
4. **Scan** patient wristband
5. **Scan** medications
6. Acknowledge any alerts, document site, rate/dose, etc.
7. Click <Scan Complete>
8. Administer Medications
9. Click <Medications Given> to chart the meds

Collaboration

Nursing • RT • Pharmacy • Information Services

- Nursing Informatics incorporated education and ownership to change the perception of what was initially feared as being a difficult and invasive process; through this guidance each pilot unit was able to adhere to the workflow and attain high success rates.
- Pharmacy pre-emptively designed workflow changes and adapted quickly; through careful barcode and order management, problems were resolved to streamline the process with patient care.
- Information Services managed hardware ordering, configuration and logistics; process and workflow education was provided to all roles. End user support teams at each hospital were provided with configuration and troubleshooting materials to allow ownership and ongoing maintenance to be readily accepted by each participating hospital.

Nursing Acceptance

New workflows • New software • New devices

Training
- Hands on user education performed either in a classroom setting or in a hospital room with real medications (when possible.) The goal was to create a simulated environment so nurses and respiratory therapists felt comfortable with the new process prior to using BCMA on a real patient.
- The didactic portion of the training included a strong emphasis on *patient safety*.

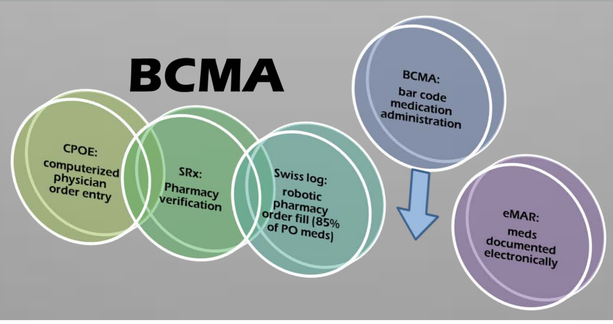
Go-live support
- Command center available on site 24 hours for the first week.

Visible Super-users
- Nursing representatives worked as super-users on their home unit.
- Resource nursing pool provided super-users, which were available for multiple units.

Daily Huddles performed
- All staff on the unit were kept informed about successful scan rates, workflow changes, and the availability of help from super-users or the command center.

Leadership participation
- Enthusiasm and visibility went a long way to drive nursing acceptance.
- Respiratory Therapy leadership's involvement improved the compliance of respiratory medications administered using BCMA.

Competency created by nursing professional development specialists.
- Competency was assessed and signed-off by nurse educators and clinical nurse specialists observing the medication administration process using BCMA.



Lessons Learned

Site visits of hospitals with the same EHR and similar hardware gave insight into workflow, potential challenges, and solutions. Based on site visits we:
- Selected bar code scanners without handles so they could be carried easily.
- Kept providers and hospital leadership 'in-the-loop' regarding the new process for medication administration.
- Set the expectation that medication administration will take longer initially.

Bar code scanning does not replace the need to do your usual safety checks.
- Must still identify the patient using two unique identifiers
- Verbally engage the patient regarding medications to be administered and any education needed.

Anticipate workarounds; especially when barcodes are available un-affixed to medication or patient wristbands. Review patient banding policy and create unique patient wristband barcodes if possible.

Design Override Reports to address non-compliant users, trouble meds, and other issues.

Communicate proper cleaning procedures for the bar code scanner hardware as these units will need to be cleaned between each (isolation) patient.

Back-up computer workstation (WOW) suggested for use when in-room devices are down.

Ongoing evaluation required to address workarounds and maintain a consistently high successful scan rate. Data sources for BCMA evaluation can include any of the following:

Nursing feedback	Override reasons
System and processes	Encourage self-report of near-misses

Outcomes

- Within several days, some locations had provided examples of what would be considered "near misses"-incorrect medications or dosages not on the patient's profile.
- Through successful adoption by each entity's pilot units, combined with the inherent patient safety value of the project, the decision was made to bring the entire enterprise up on BCMA by mid-2014.

Successful scan rate: 93% (by medication)

Reduction in Medication Errors

Resources

1. Bates, W., Cullen, J., Sweitzer, J., Shea, F., Hallisey, R., Vliet, M. V., et al. (1995). Incidence of adverse drug events and potential adverse drug events. JAMA, 274(1), 29-34.

2. Douglas, J., & Larrabee, S. (2003). Bring barcoding to the bedside. *Nursing Management*, 34(5), 36-40.

3. Poon, E. G., Keohane, C. A., Churchill, W. W., Lipsitz, S., Whittemore, A. D., Bates, D. W., et al. (2010). Effect of bar-code technology on the safety of medication administration. *The New England Journal of Medicine*, 362(18), 1698-1707.

4. Paoletti, R.D. ,Suess, T.M., Lesko, M.G., Feroli, A.A., Kennel, J.A., Mahler, J.M., Sauders, T. (2007).Using bar-code technology and medication observation methodology for safer medication administration. *Am J Health-Syst Pharm.;* 64(5) 536-543.

Contact: **Terese Kornet, MSN RN** **Terese.Kornet@uphs.upenn.edu**
Director of Nursing Systems
Sean Sarles, BA BSN RN CCRN **Sean.Sarles@uphs.upenn.edu**
Nursing Informatics Coordinator

An Agent-based Reasoning Scheme for Prioritizing Evidence Obtained from Multiple Pharmacovigilance Signal Detection Methods

Vassilis G. Koutkias, PhD[1] and Marie-Christine Jaulent, PhD[1]
[1]INSERM, U1142, LIMICS, F-75006, Paris, France; Sorbonne Universités, UPMC Univ Paris 06, UMR_S 1142, LIMICS, F-75006, Paris, France; Université Paris 13, Sorbonne Paris Cité, LIMICS, (UMR_S 1142), F-93430, Villetaneuse, France

Abstract

We develop a multiagent system capable of exploring evidence from multiple pharmacovigilance signal detection methods and data sources. Via task distribution among its agents and an appropriate interaction protocol for their coordination, the system launches diverse signal detection method implementations, filters the aggregated results with respect to novelty, and prioritizes the remaining novel indications according to various significance criteria.

Introduction & Background

Accurate and timely signal detection is a major challenge in pharmacovigilance. While diverse methods and data sources are employed for this task, current efforts illustrate bias, high false-positive rates and other limitations. In this respect, it has been argued (with supporting evidence[1]) that signal detection may be leveraged through integrated approaches. In this work, we employ the agent paradigm[2] to design a reasoning strategy that will support combinatorial signal detection using diverse detection methods and drug safety resources for filtering.

Methods

We consider detection methods suitable for observational data, spontaneous reports and free-text sources. Software agents offer an important metaphor to address problems of incomplete and distributed knowledge[2]. Delegating simple-specific tasks to agents and coordinating their actions through an agent interaction protocol enables us to: (a) launch signal detectors and aggregate their outcomes, (b) filter the results obtained with respect to known adverse drug reactions using reference sources, and (c) prioritize the remaining novel signals based on their significance. For prioritization, two types of criteria are employed: (a) computational metrics that are commonly used in the field of information retrieval, e.g. precision at *k*, and (b) domain criteria that have been proposed for implementing triage schemas for signal detection, e.g. ADR seriousness, rapidly increasing disproportionality, newer drug, etc.

Results

Our development relies on signal detection methods that are available in the PhViD R package (http://cran.r-project.org/web/packages/PhViD/) and drug safety sources like SIDER (http://sideeffects.embl.de/). For agent development and execution we use JADE (http://jade.tilab.com/). We design classes of agents according to the described tasks and work on a voting scheme to enable prioritization by combining the above ranking criteria. Agent coordination is based on message exchanges encoded in FIPA ACL (http://www.fipa.org/repository/aclspecs.html).

Discussion

We conceive signal detection within an integrative setting that accounts for multiple heterogeneous data sources and multiple detection methods. To this end, we adopt technologies like software agents and knowledge engineering[3] for the development of an integrated platform to contribute in accurate and timely signal detection.

Acknowledgement

This research was supported by a Marie Curie Intra European Fellowship within the 7th European Community Framework Programme FP7/2007-2013 under REA grant agreement n° 330422 – the SAFER project.

References

1. Harpaz R, et al. Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions. JAMIA 2013;20(3):413-9.
2. Isern D, Sánchez D, Moreno A. Agents applied in health care: A review. Int J Med Inform 2010;79(3):145-66.
3. Koutkias V, Jaulent MC. Towards an integrated framework of pharmacovigilance signal detectors through semantic mediation. AMIA Joint Summits on Translational Science, San Francisco, USA, April 7-11, 2014.

Development of an Interactive, Spatial, Web-Based Tool for Physician Workforce Planning, Recruitment, and Research

Denise D. Krause, Ph.D. [1,2], John R. Mitchell, M.D.[2], Diane K. Beebe, M.D.[1]
[1]University of Mississippi Medical Center
[2]Office of Mississippi Physician Workforce

Objective. To develop a web-based, interactive physician workforce surveillance application for the Office of Mississippi Physician Workforce (OMPW) using geographic information systems (GIS) to serve as a tool for effective workforce planning, recruitment, and health services research, to improve access to health care.

Methods. The Mississippi Board of Medical Licensure provided individual level licensure data on all active physicians practicing in the state to provide a cross-sectional view of the physician workforce. This dataset includes, but is not limited to, date of birth, gender, race, specialty, and school of graduation. From these and publicly available data from the U.S. Bureau of Census, we calculated physician-to-population ratios by county for each year of the last decade and median family income by county in order to gain an historical perspective of the demographics and distribution of the physician workforce. Data on health professional shortage areas were obtained from the Health Resources and Services Administration's website. ArcGIS Online and ESRI's geo-enrichment services supplied additional foundation layers for the application. The project was developed using ArcGIS Online and ArcGIS 10.2 Desktop. An ArcGIS 10 server application was developed in JavaScript which can run on most platforms, including mobile devices, to serve up these data based on security roles.

Key findings. The application allows users to identify and query geographic locations of individual or aggregated physicians, to perform drive-time or buffer analyses, and to explore sociodemographic population data by the geographic area of choice. Specific organizations and groups also benefit in the following ways:

- Mississippi State Department of Health officials can examine the aggregate distribution of physicians by county or public health district visualizing health professional shortage areas.
- Medical students and residents can identify areas qualifying for loan repayment programs and investigate suitable practice locations by reviewing population and area demographics and socioeconomic indicators in their areas of interest, thus identifying opportunities in health professional shortage areas early in their careers.
- Rural Physician and Dentist Scholarship Program administrators can track scholars and assess their impact on filling the needs of underserved communities. Recruitment of students from specific areas of need into the scholarship programs can increase the likelihood that they may return to those areas to practice. Conversely, health planners throughout the state can identify rural scholars from their areas for potential recruitment back to their communities.
- The OMPW can analyze patterns and trends of the health workforce over time and across the state, including attrition, saturated areas, and shortage areas.

Implications. This interactive, spatial, web-based application with analytical tools visually represents the physician workforce and its attributes, and provides access to much needed information for state-wide health workforce planning and research. The integrated and visualized data being presented improve awareness of areas in need of physicians which is useful to local and state agencies and the state legislature for making funding decisions and developing health policies. This application is an expandable tool that enables the State of Mississippi to become more proactive in addressing the needs for physicians throughout the state.

Future Directions.

- We continue to add functionality and depth to the application and intend to incorporate more advanced analysis methods, such as linear programming optimization techniques, to examine inequalities in physician availability to population demographics and determine optimal numbers needed.
- We have introduced the tool to a number of key stakeholders, including the governor's office, and have received very positive feedback. We plan to measure use statistics over time and administer satisfaction and usability surveys to user groups.

Funding source. Office of Mississippi Physician Workforce

Exploring the Content and Ranking of Diagnosis-Based Medical Problem Lists

John C. Krauss, MD, and Charles P. Friedman, PhD, University of Michigan, Ann Arbor, MI

Introduction. The diagnosis-based problem list is commonly used by clinicians to summarize the health status of a patient and organize strategies for his/her care . The problem list is a key component of the medical record; reviewing the problem list at each visit is a criterion for Meaningful Use. From an informatics perspective, the optimal strategies for problem list representation and curation are not known. Problems lists have been the focus of some, but not many, empirical studies[1]. This study assessed the generation and organization of the problem list by experienced clinicians for three complex internal medicine cases.

Methods. Eighteen physicians, averaging 19 years post-medical school graduation, reviewed three detailed patient case descriptions that have been used in previous studies of clinical reasoning[2]. For each case, they generated a problem list, transferred the problem list to cards, and then organized the cards into a list. To standardize representation of each problem as much as possible without constraining their choices, subjects generated their problem lists with reference to a provided set of candidate problems. Subjects could omit from their lists problems on the candidate set that they did not believe were important, and they could add other problems that were seen as important. After completing all three cases, the physicians were asked to explain the logic they used to organize their problem lists.

Results. The average numbers of problems included in subjects' lists were 8 (SD = 3), 12 (SD = 4), and 7 (SD = 3) for Cases A, B, and C respectively. Across all subjects, the number of unique problems listed for each of the cases was 24, 35, and 22. For each case there was substantial but far from complete agreement across subjects on the specific problems included. For Case A, the fifth most commonly listed problem was listed by 72% of subjects; for Cases B and C, the comparable percentages were 78% and 61%. The number of unique problems listed displayed more variance by case (variance component = 6.14) than by physician (variance component = 4.47). Thus there was little evidence of a consistent tendency of the subjects to be parsimonious or inclusive in the creating their problem lists. Ordering of problems within lists also was highly variable. The most commonly reported criteria for organizing problems within lists were "acuity" (n =8), "degree of life threat" (n = 4), "dangerousness" (n = 2), and need for "immediate attention" (n = 2).

Conclusions. Problem lists are highly variable and thus highly subjective. When experienced physicians were presented with identical case descriptions, a subset of problems appeared on most physicians' lists; but there was significant variance in how many problems were considered important enough to list, as well as the ordering of the listed items. Our data suggest that there is no one problem list for a patient, but rather different problem lists as generated through the lens of each clinician. This variability poses significant challenges for patient care, particularly for continuity of care in acute settings, as each physician may focus only the problems he/she prioritizes. This variability also raises a challenge for the design of EHRs to capture diverging views of the multiple physicians who care for the patient. While physicians put different labels on how they rank problems, in general the problem list was organized by all physicians based on their own interpretations, expressed in different ways, of seriousness, acuity or threat of mortality.

References

1. Wright A, Feblowtz J, Maloney FL, Henlon S, Bates DW, Use of an electronic problem list by primary care providers and specialists. J Gen Int Med, 2012;27:968-73
2. Friedman CP, Elstein AS, Wolf FM, Murphy GC, Franz TM, Heckerling PS, Fine PL, Miller TM, Abraham . JAMA. 1999 Nov 17;282:1851-6.

Exactly What Kind of Patient Data Do They Use in Research? : Analysis of Clinical Data Requests for Research

Rita Kurtz, BA, Elizabeth Bell, MPH, Hyeoneui Kim, RN, PhD
Division of Biomedical Informatics, University of California, San Diego, La Jolla, CA

Abstract
We analyzed data request logs created by the researchers at University of California San Diego (USCD) to better understand what kind of patient data items are actually utilized in biomedical research. The goal of this analysis was to inform the development of the tool that captures patient's data sharing preference. The results revealed that various types of patient data are requested and obtained for research. We need to develop a succinct yet comprehensive patient data category structure, with which patients can easily indicate data sharing preference and are sufficiently informed on their data use.

Introduction
Currently, informed consent is given by a several page long paper form, which is written in technical language and small font, and is often difficult to read and understand by the patient. Patients rarely read through the consent completely, but just sign it to receive the necessary healthcare service. Allowing a tiered manner of informed consent has been proposed as a way to better capture patient's preference on data sharing (1). Literature reported on the typical data items that are considered "sensitive" (2), however they do not take into account the data that is actually requested by researchers, and therefore cannot represent the full spectrum of sensitive patient data nor reflect the most accurate patient preference on data sharing. Therefore, to better understand the scope of the patient data actually used in biomedical research, we analyzed the data request logs collected at UCSD.

Methods
We analyzed 62 data request logs collected from September 2011 to July 2013. The request logs capture the patient data items requested by and provided to the researchers. Each requested item was organized first into "sensitive" categories suggested by the literature, comprising of demographics, diagnosis, lab and test results, medications, and medical procedures. If the requested item did not fall into any of the suggested categories, additional categories were created to encompass them, including PHI data, encounter information, insurance status, body measurements/vital signs, family history, and social history. Frequencies of items in each category were documented.

Results
While the majority of requests did ask for data covered by the sensitive data categories already suggested by the literature, more than half of requests asked for at least one patient data item that a patient might consider sensitive that was not yet covered. Examples of such patient data items are encounter information including provider name and service facility (34%), body measurement such as BMI (29%), insurance status (8%), family history (6%) and social history (6%). We developed an initial ontology for the data categories that patients can browse and indicate their preference for data sharing. The request analysis data and the initial ontology are available at https://idash-data.ucsd.edu/community/51

Conclusions and Future Directions
This study implies the need for a more comprehensive list of patient data categories to present for more accurate capture of patient's data sharing preference.

Acknowledgement
This study was supported in part by the NIH grant U54HL108460 (NHLBI).

References

1. Bell E, Ohno-Machado L, Grando MA. Sharing My Health Data: A Survey of Data Sharing Preferences of Healthy Individuals. Am Med Inform Assoc. 2014. (Accepted)
2. Meslin EM, Alpert S, Carroll AE, Odell JD, Schwartz PH. Points to consider in ethically constructing patient-controlled electronic health records. 2012. Available from: http://hdl.handle.net/1805/2936

Learning from Social Media Patient Platforms: A Framework for Exploring Mechanisms Used to Engage Patients

Claudia Lai, MSc[1]; Alex R. Jadad, MD, DPhil[1-3]; Raisa Deber, PhD[1]; Aviv Shachak, PhD[1]
[1]Institute of Health Policy, Management & Evaluation, University of Toronto, Toronto ON; [2]Department of Anesthesia and Dalla Lana School of Public Health, U of Toronto, Toronto, ON; [3]Centre for Global eHealth Innovation, University Health Network, Toronto ON

Abstract

Despite the growing interest in engaging patient in their health care process, particularly with technology, it is not clear how best to do so. Based on a narrative literature review and pilot analysis, we propose a preliminary framework to explore mechanisms used by social media platforms to engage users in health and information sharing processes, which can contribute valuable insights to support the future design of healthcare information systems.

Introduction

Patient engagement is now considered a pillar of the ideal patient centered care model, with a growing view that engaged and empowered patients are essential to improve health and decrease healthcare costs. However, despite the strong political interest in patient engagement, particularly with technology, it is not clear how best to do so. A novel approach to seek ways for better engaging and supporting patients with technology is to explore social media platforms designed for those who wish to become more involved in their health and health care decisions. Social media platforms offer rich applications, services and tools to engage users to learn how to better address their health issues. These platforms can achieve high levels of user participation outside the realm of the health care system. Thus, the objective was to develop a framework for analyzing health related social media platforms for the purpose of exploring mechanisms they use to engage patients in their health care process.

Methods

A review of the literature on patient engagement (in particular patient engagement frameworks) and patient roles in treatment decision-making was conducted to inform a preliminary framework to guide data collection and analysis. Qualitative methods were used to analyze textual and visual data (e.g., platform policies, terms of use, about us, frequently asked questions, registration process, tools, and clients) collected from a convenience sample of social media platforms.

Results

A preliminary framework was developed based on the literature review and pilot analysis. This framework draws from 1) the National eHealth Consortium patient engagement framework that attempts to align to meaningful use policies (inform me, engage me, empower me, partner with me, and support my eCommunity), 2) the Health Council of Canada's spectrum of engagement that encourages patients to participate and influence health care reform policies to improve the effectiveness of the health care system (inform, consult, involve, collaborate, empower), and 3) concepts identified by Deber et al.[1] on preferred patient roles in treatment decisions (passive, shared and, autonomous).

Discussion:

We propose a framework that can be used to analyze social media platforms that engage users who wish to be more involved in decisions relating to their health. We plan to employ this framework in a larger study of purposely selected social media platforms (e.g. sites with many registered users stratified by various characteristics representing different ownership types, funding sources and target users), and further refine it based on the findings. This novel approach to explore mechanisms to engage users in health and information sharing processes can contribute valuable insights to inform future policies and support the future design of health information systems to better engage and support patients within health care organizations.

References

1. Deber, R., Kraetschmer, N., & Irvine, J.). What role do patients wish to play in treatment decision making? *ArchIntern Med*. 1996 Jul 8; 156(13): 1414–1420.

Challenges Faced When Designing and Conducting Time Motion Studies in Health Care Environments

Barbara A. Lara MD MPH[1], Alexa Meara MD[2], Stacy Ardoin MD[2], Peter Embi MD MS[1], Po-Yin Yen RN PhD[1]
[1]Department of Biomedical Informatics, [2]Department of Internal Medicine, Division of Rheumatology, The Ohio State University Wexner Medical Center

Introduction: Clinical workflow describes a process that comprises a wide range of activities executed by health personnel to provide patient care. Understanding clinical workflow has progressively been recognized as an essential component of successful Health Information Technology implementation and adoption. Originally created as a business efficiency assessment technique, Time Motion Studies (TMS) have been adopted as a working method to describe and assess clinical workflow. While some standards about the methods used to conduct TMS have been provided[1], existing literature only provides general guidelines for designing TMS, but do not address practical challenges encountered when using this method[2]. Here we present a set of common challenges based on our group's experience.

Methods: Based on several TMS conducted by our group over the past years, we outlined crucial steps in the design and implementation process of TMS. Along with these steps we also identified a set of challenges that are distinct unique to the methodology required to conduct continuous observation TMS.

Results: We describe the most interesting challenges encountered by our group despite careful planning. We categorized them based on the phase of the study in which they were encountered. Full description and guidelines of a step by step design and implementation of TMS will be published elsewhere.

#1. **Difficult observation arrangement for sampling purposes:** To study clinical workflow that involves observing clinician-patient interaction, either clinicians or patients need to be enrolled beforehand. However, estimating the amount of observer-time, calendar time and exact calendar schedule can be challenging, as clinicians often change shifts, and patients cancel or no-show to their appointments and none of these events are predictable. Frequently, even when many potential encounters are programmed to happen sparsely on a specific day, they might take place simultaneously, preventing the observers to observe all of them.

#2. **Clinical personnel reluctant to participate:** Groups of clinicians rotate through their units throughout the year (float staff). These personnel are usually not informed of the study and are thus reluctant to participate or collaborate. This hinders data collection for that specific event which in turn compromises the sampling strategy.

#3. **Training observers and maintaining inter-observer reliability is costly and labor intensive:** Data collected for TMS is "observer-dependent", which requires consistent and highly trained personnel, to ensure study validity and reliability. Furthermore, understanding and thus being able to collect data in specific clinical environments requires subject matter expertise and/or intensive training. Moreover, because of ethical and medical issues involved when clinician-patient encounters are part of the observation event, direct training on site might not be feasible. Methods to train observers in a controlled environment are required, increasing the complexity of this process.

#4. **Selecting "reliable" and qualified observers:** Because of the high cost of training, importance of subject matter expertise and data consistency, it is optimal to select "reliable" observers that will commit to work for the full length of the study. Finding observers with the right set of characteristics is a challenge. If a high turnover rate exists, the cost of the TMS increases, prolonging the observation period, and even threatens data consistency.

Conclusion: Previously presented challenges highlight the importance of developing best practice guidelines, where pilot testing might play a significant role in accounting for unexpected variations of the idealized planning design.

References

1. Lopetegui, M., Yen, P.-Y., Lai, A., et al. J Biomed Inform. 2014 Jun;49:292-9
2. Zheng K, Guo MH, Hanauer DA. J Am Med Inform Assoc. 2011;18:704-710

Meaningful Use in Medication Reconciliation & Patient Medication Information Management Requires Patient & Healthcare Team Partnership, Community Collaboration, and Standardization: The Veterans Health Administration Story

Maureen Layden, MD, MPH Director Veterans Health Administration Medication Reconciliation Initiative

Medication reconciliation (MedRecon) and medication information management are effective patient interventions to improve care. They have the potential to reduce harm, optimize therapy, improve transitions, and create medication treatment plans that are personalized to patient's preferences and address barriers to adherence. To realize these goals requires collaboration with all stakeholders from the healthcare community, healthcare teams, and leaders in operations, policy making, informatics, quality management, and patient safety. It demands standardization in policies, monitoring, and procedures as well careful evaluation to make certain patients, their caregivers, and healthcare teams can be successful at managing medication information throughout the healthcare continuum. Ultimately, MedRecon needs to be patient centered with foremost consideration of how best to partner with patients, their caregivers, and families.

Challenges: Numerous efforts to improve the exchange of medication information and perform medication reconciliation exist at multiple levels of the Veterans Health Administration (VHA). A concerted effort is needed to establish standards, coordinate platforms, harmonize requirements, and ensure goals are not duplicative to prevent uncoordinated efforts. Lack of an of overarching strategy will increase the risk of failure in meeting external mandates from Centers for Medicare & Medicaid Services Meaningful Use Standards Incentive Programs and Joint Commission Standards as well as internal mandates such as the VHA MedRecon Quality Indicators.

Methods/Strategies: Collaborative interprofessional workgroups comprised of VHA, Department of Defense, and Indian Health Services champions were established to develop business requirements and standards in the following areas: medication information management through the VHA Essential Medication Information Standards Directive, MedRecon policies, patient and staff education programming, effective documentation strategies, performance metrics, and change management campaigns.

Results: The key elements identified deemed necessary for enterprise-wide success in medication information management and MedRecon included: leadership endorsement of a unified enterprise strategy created through community collaboration, founded in standardization and partnership with Veterans, their caregivers, and families.

Conclusion: Partnering with Veterans and their Medications Task Force will execute a colloborative overarching strategy founded in these elements. It will guide iterative development of enterprise programs and tools. Fulfillment of internal and external mandates such as Meaningful Use must be met in the setting where our patients, their caregivers and healthcare teams also experience safe, timely, accurate, understandable, and meaningful distribution of medication information.

Learning Objectives: After reviewing this poster, the learner should be able to:

1. Recognize the importance in engaging and collaborating with key stakeholders so that strategy to meet Meaningful Use MedRecon will also provide value and improvement in the day to day work flow of patient care.
2. Employ strategies that align MedRecon requirements with existing internal and external requirements as the VHA has accomplished with VHA MedRecon Quality Indicators and the Joint Commission Standards.
3. Identify three ways to standardize medication information across the healthcare continuum and manage medication information as part of the medication reconciliation process.
4. Understand the importance in emphasizing patient engagement in any strategy to meet Meaningful Use MedRecon so that it improves their experience and satisfaction with our healthcare organizations.

Considerations in Implementing Informatics Studies in Dementia Care Units

Amanda Lazar[1], Hilaire Thompson, PhD, RN, CNRN, FAAN[1], George Demiris, PhD, FACMI[1]
[1]University of Washington, Seattle, WA

Abstract
Thirty to 40% of people with dementia live in nursing or assisted living facilities and dementia care units, compared to only 2% of older adults without dementia (1). People with dementia living in memory care units are in need of interventions encouraging stimulating activities and meaningful interactions with others. Information and communication technologies (ICT) have the potential to provide opportunities for engagement for this population. We present lessons learned from conducting an evaluation study of a multifunctional technology tool designed to facilitate opportunities for entertainment, cognitive training, communication, and information access with this challenging population. The evaluation approach included both a formative and a summative evaluation of this technology and our study highlights technical and ethical implications for system designers and informatics researchers in the area of dementia care.

Introduction
As the population ages, the number of people with dementia will increase proportionally. The healthcare system is unequipped to deal with the cost of care for dementia, and needs for social interactions and meaningful engagement are often unmet for this population, especially for people living in assisted living and memory care units (2-4). Information and communication technology can provide avenues for people with dementia to interact with others, improve cognitive skills, take part in physical activity, and have fun. In light of this potential, it is imperative to design and evaluate technology tools to improve the wellness and quality of life of this population.

Methods
In a six-month longitudinal study, we used a mixed methods approach to evaluate a commercially available multi-functional technology tool designed to be used by or with older adults with dementia. The tool contained applications for communication, cognitive training, information access, exercise, and leisure activities. The technology tool was used in weekly hour-long sessions with residents and researchers as well as daily at the discretion of staff members. The authors gathered quantitative and qualitative data from the three stakeholder groups in a person with dementia's care: the individual with dementia, a close family member, and staff in the memory care unit. Data sources included interviews, standardized instruments, and system logs.

Results
Lessons learned from conducting a technology evaluation in a memory care unit involve technical and ethical aspects as well as human computer interaction challenges. Conducting a study in a memory care unit and working with a population with dementia can be challenging. An important facilitator is the approach of introducing the intervention to family members and staff to ensure greater buy-in. This was because people with dementia were not able to use the system independently and thus the system had to be integrated into unit programming and care. A common issue to overcome was the perception of family members who thought their relatives would be unable to use or benefit from the technology due to the severity of their dementia, although this was not the case. We also include considerations for planning study sessions and the administration of standardized instruments with people with dementia in a memory care unit, including how to determine the best time to conduct study procedures. These findings will help researchers plan studies working with this population in the future.

References

1. Alzheimer's Association. Alzheimer's disease facts and figures. Alzheimer's & Dementia. 2012 8(2), 131-168.
2. Hancock GA, Woods B, Challis D, Orrell M. The needs of older people with dementia in residential care. Int J Geriatr Psychiatry. 2006 Jan;21(1): 43-49.
3. Moyle W, Venturto L, Griffiths S, Grimbeek P, Oxlade D, Murfield J. Factors influencing quality of life for people with dementia: A qualitative perspective. Aging Ment Health. 2011; 15(8);970-977.
4. Wood W, Harris S, Snider M, Patchel SA. Activity situations on an Alzheimer's disease special care unit and resident environmental interaction, time use, and affect. Am J Alzheimers Dis Other Demen. 2005 Mar 1; 20(2).

Development of a Web-based Patient-Centered Discharge Checklist Toolkit

Jae-Ho Lee, MD, PhD[1,2,3], Patricia Dykes, PhD, RN, FAAN[1,2,4], Diana L. Stade[1], Frank Y Chang[5], Anuj K. Dalal, MD[1,2], David W. Bates, MD, M.Sc[1,2,5]

[1]Div. of General Internal Medicine, Brigham and Women's Hospital, Boston, MA;
[2]Harvard Medical School, Boston, MA;
[3]Depart. of Emergency Medicine, Univ. of Ulsan College of Medicine, Seoul, Korea;
[4]Center for Excellence in Nursing Practice, Brigham and Women's Hospital, Boston, MA;
[5]Partners eCare Clinical Informatics, Partners Healthcare System, Wellesley, MA

Abstract

Patients in transition of care suffer from adverse events, which results in readmission or morbidities. Patient's active engagement in the discharge process is one promising solution. A web-based patient-centered discharge checklist toolkit and summarized discharge instruction frame were developed. The development process and the result were described.

Introduction

Many discharged patients suffer from adverse events, which results in readmission or morbidities[1]. Active engagement of patients in the discharge process is one promising solution. Several kinds of discharge checklists for patients and providers have been developed for this purpose. Patients can understand the discharge plan and instructions more effectively if a patient-centered discharge checklist is provided electronically and interactively. We developed a web-based patient-centered discharge checklist toolkit (PCDCT) and described the development process and the result.

Methods

Development of our PCDCT was conducted as a subproject of a web-based patient-centered toolkit (PCTK), which was designed for hospitalized patients to participate actively in their hospital stay and treatment by using tablet devices. Our team was composed of multidisciplinary informatics researchers and hospital staff members. A modified version of the Robert Wood Johnson Foundation discharge preparation checklist was used that categorized discharge checklist items and a medication checklist. A prototype of PCDCT was developed after category names and the order of checklist items were modified. Each checklist category was displayed completely in a window without a scroll bar. The prototype was changed iteratively after meetings with care providers and a patient advocate. To give reliable information to patients completing the PCDCT, a discharge instruction frame was also developed.

Result

PCDCT and discharge instruction frame were developed and implemented into the PCTK. PCDCT was composed of five steps: 1) an introduction to the discharge checklist, 2) 'My follow-up plan', 3) 'My medications', 4) 'My self-care management', and 5) 'My discharge plan of care'. A message tool for real-time communication with the care team was also implemented. Thus, patients could ask questions if they wanted more information about the discharge plan. The discharge instruction was also provided to get detailed information without the help of the care team. Contents were retrieved from the discharge module by using an existing web service. Once patients complete the discharge checklist, a message of 'Now you can leave the hospital safely' was designed to display.

Conclusion

Our next steps are completion of the interface with the existing discharge module and development of an electronic form of the discharge instructions. We expect that this toolkit can help patient engage in discharge process actively.

References

1. Halasyamani L, Kripalani S, Coleman E, Schnipper J, van Walraven C, Nagamine J, et al. Transition of care for hospitalized elderly patients: Development of a discharge checklist for hospitalists. J Hosp Med. 2006;1(6):354-60.

A Case Study on Integrating a Genealogy Database into a Consumer-Facing Family Health History Tool

Jaehoon Lee, PhD[1], Nathan C. Hulse, PhD[1,2], David P. Taylor, PhD[2], Pallavi Ranade-Kharkar,MS[1,2]
Grant M. Wood[2], Peter J. Haug, MD[1,2], Stanley M. Huff, MD[1,2]
[1]University of Utah, Salt Lake City, UT; [2]Intermountain Healthcare, Murray, UT

Abstract

In this study we integrated family history data from the FamilySearch.org genealogy database with a consumer-facing, family health history (FHH) tool: OurFamilyHealth. A data import plug-in was developed to 1) retrieve family history data from FamilySearch.org, and 2) convert them into an internal FHH data model for clinical use in OurFamilyHealth. We successfully translated between the two models and demonstrated the feasibility of system integration. We anticipate that adding this feature will make it easier for our patients to enter rich FHH data.

Background

OurFamilyHealth is a locally-developed FHH tool used at Intermountain Healthcare[1]. Currently it provides two methods for patients to enter family members, manually or through uploading a GEDCOM (GEnealogical Data COMmunication) formatted genealogy file. FamilySearch.org is a worldwide genealogy database that provides an application programming interface (API) allowing external applications to access its family history data via the web[2]. Through OurFamilyHealth's use of this API, a patient who has already entered family history in FamilySearch can provide authorization to directly import his/her data into OurFamilyHealth as a starting point for entering FHH.

Method

The family history data in FamilySearch.org includes demographics, family relationships, events (e.g. birth, death, burial, marriage, divorce, etc.), among others. Certain relevant elements conceptually match to those of the FHH data model in OurFamilyHealth and could be consumed by the application. We utilized the Family Search APIs to retrieve data through web services, and translated between the two data models using transformations (See Table 1).

Table 1. Transformation of family history data from FamilySearch.org to OurFamilyHealth

Category	Transformation
Demographics	Name, gender, birth date, death date are transferred directly.
Uncertainty of date	If a birth date or death date has an attribute of about / after / before, it is considered as an approximate date.
Living status	If a person has a death date, he/she is considered deceased.
Adopted status	If a parent-children relationship has attribute of adopted / foster / guardianship, the child is considered adopted.
Family relationship	OurFamilyHealth consumes 3 degrees of family relationships which are meaningful for genetic risk assessment; 1st: parents, children, siblings, 2nd: grandparents, grandchildren, half-siblings, uncles / aunts, nieces / nephews, 3rd: great grandparents, great grandchildren, 1st cousins, half nieces / nephews, grandnieces / grandnephews, great uncles / aunts.
Marital status	Nine types of spouse relationships are mapped to four marital statuses. Multiple spouses and their dependent families are also transformed.

Conclusion and future work

Prototyping in an Intermountain Healthcare development environment using the FamilySearch.org sandbox database demonstrated that the proposed approach is reasonable in terms of data transformation and system integration. Our future work is to: 1) certify the application with FamilySearch.org so that we can eventually integrate with their production database, and 2) implement the application in the Intermountain production environment.

References

1. Hulse NC, Ranade-Kharkar P, Post H, Wood GM, Williams MS, Haug PJ. Development and early usage patterns of a consumer-facing family health history tool. AMIA Annu Symp Proc. 2011:578-87.
2. FamilySearch [Internet] 2014 [cited 2014 Feb 26]. Available from: https://familysearch.org/

The Development of a Mobile Nurse Shift Reporting Application based on Human-Computer Interaction Design

Mikyoung Lee, PhD, RN[1], Anna McDaniel, PhD, RN[2], Josette Jones, PhD, RN,[3] Denise Kerley, BSN, RN[4]

[1]Assistant Professor, Indiana Univ. School of Nursing, Indianapolis, IN, [2]Dean & Professor, Univ. of Florida College of Nursing, Gainesville, FL, [3]PhD, RN, Associate Professor, Indiana Univ. School of Informatics & Computing, Indianapolis, IN, [4]BSN, RN, Indiana Univ. Health University Hospital, Indianapolis, IN

Abstract

The use of mobile technologies for health care delivery is increasingly advocated. Yet little research exists about its use for nursing handoff communication. This study aimed at developing a mobile nurse shift reporting application, based on the principles of human-computer interaction design, using a focus group of practicing nurses. The new mobile application will lead nurses to capture/handover patient information in a systematic and standardized manner, and support safe, efficient, patient-centered bedside handoffs.

Introduction

Dynamic and complex clinical environments present many challenges for effective communication among health care providers. Nurse shift report (NSR) is an important communication process that plays a pivotal role in the continuity, quality, and safety of patient care. Mobile technologies are increasingly adopted into health care. Yet, little research is available about the mobile application for NSR.

Purpose

The study aimed at identifying users' needs and designing a user-friendly interface of a mobile NSR application integrating the American Association for Critical Care Nurses (AACN) Synergy model[1]

Method

With the focus group, consisting of 12 staff nurses from 6 different care units at an academic medical center, we composed and organized 175 candidate data elements, based on the analysis of current shift reporting practice, on a low fidelity prototype mobile application. The common critical and unit specific care elements necessary for NSR across the units were identified and coincided with 8 patient characteristics defined by the Synergy model. Based on human-computer interaction (HCI) design principles,[2] we asked semi-structured questions to elicit nurses' feedback on the critical interface features for shift-report repeatedly along with the demonstrations of the revised designs.

Results

Regarding interface features, there were initially different preferences by nurses from different units, but not by nurses' years of clinical experiences. Considering the nurses' preferences, HCI design principles, and usability principles,[3] the mobile NSR application was developed. Most nurses preferred one page fill-in form; however, they also wanted a navigating system. The care elements coincided with 8 patient characteristics were placed in the tab order of Vulnerability, Stability, Resiliency, Predictability, Resource Availability, Participation in Decision Making, Participation in Care, and Complexity. This order was made by the nurses' perceptions on the priority of those care components critical for capturing and reporting the patient's conditions at shift change. A rating system of each patient characteristic based on the patient condition was also created to assist nurses to judge the severity of patient conditions in total scores of patient characteristics. The terms and abbreviations were consistent with the ones of electronic health record that nurses were using. As different units deal with different amount of information on patient conditions, the data elements are set to be shown or hidden interactively according to unit characteristics. Key concerns are highlighted in red color to get more attention from nurses.

Conclusions

The focus group nurses anticipated that this new mobile NSR application would be a more focused tool for nurses to capture critical patient information, and would lead to improvements in nurse-to-nurse communication and coordination. Its usability testing will be conducted with additional staff nurses of the units in the hospital.

References

1. American Association of Critical-Care Nurses. *The AACN Synergy Model for Patient Care.* http://www.aacn.org/wd/certifications/content/synmodel.pcms?menu=certification Accessed March 1, 2011.
2. Dix J., Finlay G., Beale R. *Human Computer Interaction*, England: Pearson Education Limited; 2004.
3. HIMSS EHR Usability Task Force. Defining and testing EMR usability: Principles and proposed methods of EMR usability evaluation and rating. HealthCare Information and Management Systems Society; 2009.

Patient Safety Education using an Electronic Error Reporting System

Nam-Ju Lee, RN, PhD[1]
[1] College of Nursing, Seoul National University, Korea

Abstract

The purpose of this study was to develop an electronic error reporting system and implement it into a nursing curriculum in an effort to improve students' competencies related to error reporting. The electronic error reporting system was developed and the WHO-International Classification (WHO-ICPS) for Patient Safety was used. 76 nursing students reported 236 errors during a five-week clinical practicum. Students' knowledge, attitude and skill competencies related to error reporting all showed significant increases.

Introduction

Error reporting is an essential component in improving patient safety, but many health care providers may be reluctant to report errors due to multiple barriers. Education about error reporting as part of an undergraduate curriculum may reduce resistance to error reporting. The present study was performed as part of a project entitled the 'Development and Evaluation of a Patient Safety Curriculum for the Improvement of Nursing Students' Patient Safety Competencies'. The purpose of the study was to develop an electronic error reporting system and implement it into a nursing curriculum in order to improve students' competencies related to error reporting.

Methods

The error reporting system was designed based on an analysis of the error reporting systems used in teaching hospitals in which nursing students are trained for their clinical practicum. The electronic error reporting form was developed using Google Docs. The form consists of five parts: information about the students, the subject who was harmed or was nearly harmed, error detectors, relevant persons, and relevant events. In the event session, users select the types of events and record a detailed description of an event, its causes, the level of harm, and suggestions to prevent the reoccurrence of the event. The WHO-ICPS was used for standardized terms in the system. After pilot testing with four users, including nurses and nursing students, the form was finalized based on an iterative evaluation process. The study was approved by the institutional review board of the college of nursing and de-identified data were collected. During the clinical orientation, the students learned about general patient safety concepts and were taught how to use the error reporting system. They could access it with a wired or wireless network and document any error that they made or observed during their five-week clinical practicum. It was a voluntary, anonymous process. After the patient safety education sessions, students' error reporting competencies were measured using the questionnaires developed by the research team. The errors were analyzed by type according to the WHO-ICPS and descriptive statistics were used. Content analysis was used to analyze the data collected using focus groups.

Results

236 errors were reported by 76 students. On average, students reported about 3 errors. In 212 events, the patient was the subject who was harmed or nearly harmed. Employees and caregivers accounted for 4.2% each. 79 events were near misses, and 89 occurred without harm. 24 resulted in harm to patients. Students were involved in only 9 events. The most frequent type of error was Medication/IV Fluids (24.6%) followed by Clinical processes/procedures (21.8%), Behavior (12.9%), Clinical administration (6.2%), Patient accidents (5.5%) and Medical device/Equipment errors (5.2%). Students' knowledge, attitude and skill competencies related to error reporting were significantly increased after the implementation of the patient safety curriculum, including the use of the error reporting system. Students reported that they had an opportunity to understand errors systematically. Also, they mentioned that they realized the important role of the error reporting system in improving patient safety. Regarding suggested improvements to the error reporting system, the students noted some ambiguity in the terms used for the error types, the lack of a preview option for subtypes of errors, and limitations related to the user interface (e.g., no back buttons).

Conclusion

Patient safety education using this error reporting system may improve students' understanding of the nature and causes of errors and may therefore help students to identify errors and prevent them during their clinical practicum.

This study was supported by National Research Foundation of Korea (810-20110011, 810-20120011).

The Readability of Diabetes Patient Education Materials on the World Wide Web based on LSA and SVM technique

Ying-Li Lee, RN, MS[1, 2], Hou-Chiang Tseng[3], Yao-Ting Sung, PhD, Professor [4], Ju-Ling Chen, PhD, Assistant Professor [5], Shao-Hui Shu, RN, MSN[6]

[1]Department of Nursing, Chi Mei Medical Center, Tainan, Taiwan; [2]Institute of Biomedical Informatics, National Yang Ming University, Taipei, Taiwan; [3]Research Center for Psychological and Educational Testing, National Taiwan Normal University, Taipei, Taiwan; [4]Department of Educational Psychology and Counseling, National Taiwan Normal University, Taipei, Taiwan; [5] Center for Teacher Education, National Taitung University, Taitung, Taiwan ; [6]Faculty of Nursing, Tzu Chi College of Technology, Haulian, Taiwan

Abstract

We used latent semantic analysis (LSA) and support vector machine (SVM) to analyze the readability of education materials from the websites for diabetes patients in Taiwan. Our purposes are to evaluate the readability of these texts and to give suggestion to health professional educators.

Introduction

Readability refers the degree to which a text can be understood. The readability of patient education materials on the websites may affect the effectiveness of patient education. Traditional readability measures are primarily based on the number of words in the sentences and the number of letters or syllables per word. However, domain-specific texts contain many professional terms that should not be predicted using traditional readability formula only. LSA have been commonly used to classify texts containing domain-specific knowledge. This study aims to evaluate the readability of diabetes patient education materials by LSA and SVM technique.

Method

The readability model bases on LSA (Tseng, Chang, Chen & Sung, 2014) were used to analyze the readability of texts. Through latent semantic space, any word, text or set of texts can be transformed into a semantic vector for representing the association within domain knowledge. In the modeling stage, we applied 6230 texts from textbooks of Chinese language arts, natural science, social science, physical and health educations to develop the corpus. The Chinese Readability Index Explorer 2.0 (CRIE2.0) was used for word segmentation and tagging. The word list contained important words of each grade (G1-12) per subject was developed by LSA and was used to calculate readability indicator of each text in our study. Then, we used SVM to develop the readability model for evaluating domain-specific patient education materials. In the testing stage, we input 249 texts of diabetes patient education to the readability model and to evaluate the grade level of each text.

Result

249 patient education materials from websites were selected from 15 medical centers in Taiwan. The average coverage rate of conceptual knowledge to our corpus was .77 ± .078. Only 37.65% (n=93) of texts were the revised editions in 3 years. 29.72% (n=74) of texts did not show the published year. None of the texts marked the grade level of text. 88.4% (n=220) of diabetes patient education texts were non-specific type of diabetes. The results showed that the text with the highest readability indicator was senior high school level (G10-12) and the easiest was G1. We found that the readability indicator of 22.09% (n=55) texts graded to junior high school level (G7-9) and 34.54% (n=83) graded to senior high school level.

Conclusion

34.54% of the diabetes patients education materials on websites in Taiwan were too difficult to read for patients with an elementary school or junior high school education degree. Hence, the results of this study showed that we may underestimate the difficulty of these texts. The results supported that medical terminologies and health professional concepts should take into account as well as domain-specific knowledge when evaluate texts in patient education materials. In the future, we will integrate medical terminology and health professional concept into the readability model to improve the validity of the analysis.

References

1. Tseng, H.C., Chang T. H., Chen, B. L., & Sung, Y.T. (2014, Apr). *Analyzing textbooks by a readability model based on concepts and support vector machine.* The Asian Conference on Language Learning.

The Association between Numeracy Component of Health Literacy and Online Health Information Seeking Behavior

Young Ji Lee, PhD, RN
Department of Preventive Medicine, Northwestern University, Chicago, IL

Abstract
Numeracy is an important component of health literacy; however, its importance has been understudied among online health information seekers. The aim of the study is to explore the association between the numeracy component of health literacy and online health information-seeking behaviors (HISBs) using the Health Information National Trends Survey 2007 dataset. This study found that the numeracy component of health literacy is positively associated with online health information-seeking behaviors.

Introduction
Health literacy has been a challenge for people who seek health information from the Internet. One of the Healthy People 2020 objectives is "to increase the number of online health information seekers who report easily accessing health information". To accomplish this goal, improvement in health literacy is an obvious step. However, previous studies have focused only on the association between reading skills and online health information-seeking behaviors, although numeracy is a key component of health literacy.[1-3] Therefore, the association between numeracy and online health information-seeking behaviors (HISBs) must be identified and emphasized.

Methods
This study used HINTS 2007 data since this dataset is the only set that includes health literacy-related variables among the HINTS data sets. Statistical Package for the Social Sciences (SPSS) Version 20.0 software was used to analyze the data. First, the chi-squared test was used to examine differences in categorical variables according to online HISBs. Then, multilevel logistic regression was conducted to examine the association between the numeracy component of health literacy and online HISBs.

Results
Among the entire sample, 4,820 participants answered the online HISB-related questionnaire; 42.6% of respondents had sought health-related information from the Internet. Significant variables in the bivariate analyses were age, education level, race/ethnicity, employment status, cancer history, and numeracy. Based on the multilevel logistic regression, individuals who depend on numbers to make health decisions are likely to seek online health information. Additionally, participants who are younger, have higher education, and are employed are likely to turn to the Internet for health information. The final model was significant and demonstrated a good fit with the data.

Table 1. Multilevel logistic regression explaining respondents' online HISBs (*b*=coefficient; $^{*}p<.05$ $^{**}p<.01$ $^{***}p<.001$)

Explanatory variables		**Model *b***
Demographic & Health situation factors	Age	-0.538***
	Education level	0.633*
	Employment	0.181**
	Cancer History	0.016
Health literacy	Med stats	-0.054
	Numbers are helpful	-0.131***

Conclusion
This study showed that numeracy component of health literacy is positively associated with online HISBs. Further study with the recent data is needed to explore the trends of the association.

References

1. Stagliano V, Wallace LS. Brief Health Literacy Screening Items Predict Newest Vital Sign Scores. J Am Board Fam Med. 2013;26:558-65.
2. Benigeri M, Pluye P. Shortcomings of health information on the Internet. Health Promot Int. 2003;18:381-386.
3. Miller EA, West DM, Wasserman M. Health information Websites: characteristics of US users by race and ethnicity. J Telemed Telecare. 2007;13:298-302.

Interstate Exchange Implications for Multiple Health Information Networks

Cynthia LeRouge, Ph.D.[1], Jennifer Tillman, M.H.A.[1], Joanna Hirsch, J.D.[1]
[1] Saint Louis University, St. Louis, Missouri

Abstract
The case for health information exchange networks within states is gaining increasing growth and acceptance. Use cases also exist for interstate exchange (ISE). This study explores the latest ISE value propositions, use case scenarios, and barriers that can be found in literature review and interviews with healthcare organizations.

Introduction
Exchanging patient information through ISE among multiple state health information networks (HINs) improves communication and advances the coordination of care among a larger network of providers. In particular, ISE could prove beneficial for providers practicing in multiple states. In addition, patients could benefit, specifically those who live and receive primary care in rural and smaller communities, but cross state lines to receive specialty care in metropolitan areas. The value propositions of state-level HINs have been realized; however, there is a need for further understanding of the many benefits of ISE. Although there is a federal vision for a nationwide, public-private partnership, this national HIN is not functioning at this time and little is known about this endeavor.

Methods
Data was gathered through a literature review of ISE articles from 2009 and 2013 based on a search of common health informatics databases using the following terms: Interstate Exchange, Health Information Network, Direct, e-Health, Health IT, Western State Consortium, and Gulf Coast Task Force (contact authors for a list of literature). In addition, seven, 60-minute structured telephone interviews were conducted with leaders of state and regional HINs from five states in varying stages of implementation of intrastate and ISE to explore value propositions, use cases, and suggestions for meeting challenges. Thematic coding was completed using Dedoose © qualitative software.

Results
The value propositions related to ISE create compelling cases that center around rural patients crossing state lines to seek specialized medical care in urban centers, health systems and individual providers having a presence in multiple states, service to migrant communities, and the facilitation of care resulting from natural disasters. Direct seems to be the easiest method of exchange; however, Query-based Exchange could have a more extensive impact. Results to date (see Figure 1) indicate barriers and potential suggestions for addressing challenges to ISE.

Barrier	Possible Means to Address
Concerns relating to the need to obtain a clear understanding of other state's privacy laws	Reconcile state laws relating to purpose of use, participant eligibility identification, and content of transactions
Lack of similar patient consent policies among all states	Include patient consent status and segment sensitive data
Concerns of potential liability for improper disclosure of patient information, including in disaster situations	Address disaster policies within ISE agreements, including clearly defined scenarios for "Break the Glass" policies
Concerns that participation could lead to a loss of control over their patients' health information	Establishment of clearly defined purpose of use, auditing, and penalty policies
Concerns related to the costs to update HIN platforms, implement, and participate in ISE	Clearly delineated allocation of costs among all state HINs
Distrust towards inclusion of non-provider participants	Development of a participant authorization process
Concerns related to managing multiple agreements with differing terms	Use of one participation agreement and development of a governing model to oversee data exchange
Perception that some states will wait to join national HIN	Heighten awareness on realistic timing for a national HIN
Administrative burden to notify patients that data is on ISE	

Discussion
Though barriers exist, exchanging patient information through ISE has the potential to demonstrate improved coordination of care to provide greater convenience and better health outcomes for patients. State HIN's interviewed seem to be well aware of the benefits and barriers and are proactively seeking the means to address these challenges.

Using Natural Language Processing for Autism Trigger Extraction

Gondy Leroy, PhD, Margaret Kurzius-Spencer, PhD, Sydney Pettygrove, PhD

University of Arizona, Tucson, AZ

Abstract

The Arizona Developmental Disabilities Surveillance Program collects and abstracts special education and clinical records on 8-year old children with behaviors suggestive of autism. Natural language processing (NLP) tools are developed as part of a feasibility study of information extraction. The NLP algorithms combine standard components with customized gazetteers and finite state automata. The algorithms were evaluated on 36 records from which 19 triggers were extracted with 68% precision.

Introduction

Autism Spectrum Disorder affects about 1 in 88 children in the U.S. and its prevalence increased more than two-fold since 2000. Whether this increase is due to a true increase or better knowledge and reporting of characteristics by educators and clinicians is debated. The Arizona Developmental Disabilities Surveillance Program is part of the CDC Autism and Developmental Disability Monitoring Network and has been tracking children biennially in Maricopa County (AZ) since 2000. Using ARCHE database software, special education and clinician records of children who turned 8-years old and showed signs or symptoms of autism (i.e., 'triggers') are abstracted for further evaluation. We are developing NLP tools to extract information automatically and report on a feasibility study.

Text Mining of Abstractions

All records were processed using GATE[1]'s tokenizer, sentence splitter, and Parts-of-Speech tagger. Custom gazetteers were created to recognize trigger words. Only words that appear in the ARCHE Abstraction Manual were included (Table 1). Eighteen Java Annotation Pattern Engine (JAPE) rules were developed to annotate diverse combinations of trigger words as 'triggers'. JAPEs are implemented as finite state automata (FSA) in GATE.

Thirty-six abstracted, de-identified records were selected at random from two school districts in Arizona (Table 2). In these, a total of 19 triggers (5, 5 and 9 in resp. 2000, 2006, and 2010), e.g., "lack of social or emotional reciprocity", were extracted from the text with 68% precision. Errors were generally due to too general extractions, e.g., "not play imaginative games" is not sufficiently specific to ASD.

Table 1. Words in Customized Gazetteers

Gazetteer & JAPE	**N**
Social Connectedness / Nonsocial Connectedness	33 / 21
Social Object / Nonsocial Object	41 / 17
Social Gesture / Nonsocial Gesture	85 / 27
Inappropriate (manner)	18
Excessive (manner)	26
Absent (manner)	65
JAPE rules	18

Table 2. Text characteristics.

	Study Year			
Avg. Per Record	**2000 (N=12)**	**2006 (N=12)**	**2010 (N=12)**	**All (N=36)**
Words	589	396	481	**488**
Sentences	32	23	26	**27**
Trigger Words	39	28	32	**33**
Triggers	.42	.42	.75	**.53**
Precision	80	80	56	**68**

Conclusion and Future Work

The approach shows feasibility of extracting triggers automatically from text. Future work will include gazetteer expansion using the UMLS and developing additional JAPE rules to increase recall of triggers. Support vector machine (SVM) classification will be used to increase precision. Once sensitivity and specificity have reached acceptable levels, more records will be processed to compare study years and for further text mining.

References

1. Cunningham H, Maynard D, Bontcheva K, Tablan V. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In: 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02); 2002 July; Philadelphia; 2002.

Title: Automating Identification of Multiple Chronic Conditions in Clinical Practice Guidelines

Authors and Affiliations: Tiffany I Leung, MD, MPH[1,2], Hawre Jalal, MD, PhD, MSc[1,2], Donna M Zulman, MD, MS[1,3], Douglas K Owens, MD, MSc[1,2], Mark A Musen, MD, PhD[4], Michel Dumontier, PhD[4], Mary K Goldstein, MD, MS[1,2]
[1]Department of Veterans Affairs, VA Palo Alto Health Care System, Palo Alto, CA; [2]Center for Primary Care and Outcomes Research, Stanford University, Stanford, CA; [3]Division of General Medical Disciplines, Stanford University, Stanford, CA; [4]Center for Biomedical Informatics Research, Stanford University, Stanford, CA.
Views expressed are those of the authors and not necessarily those of the Department of Veterans Affairs.

Introduction: Two-thirds of Medicare beneficiaries have multiple chronic conditions (MCCs), or two or more chronic conditions, but clinical practice guidelines (CPGs) generally focus on single conditions. Providing guidance on managing patients with MCCs is critical for practicing clinicians. However, it is unclear to what degree disease-specific CPGs provide guidance about comorbid conditions. We developed and evaluated an automated method to determine the extent to which disease-specific CPG recommendations mention comorbid conditions.

Methods: This study focuses on guidelines for the 15 most prevalent Medicare chronic disease diagnoses, excluding *cancer* given the breadth of the term, and adding *obesity* due to its high prevalence and clinical significance. We compiled a corpus of text from CPG summaries available in the National Guideline Clearinghouse (NGC). CPG summaries are extracted from full text clinical practice guidelines published by guideline-authoring committees and organizations. CPG summaries were included if (1) at least one of the 15 diseases was mentioned in the title and (2) the target population was the general adult, non-pregnant population. To identify the set of CPG summaries for analysis, we searched their titles for the 15 chronic diseases of interest using both text search and manual verification of the titles against inclusion criteria. To complete the text corpus, we then combined CPG summaries that had been extracted from the same full guideline.

To develop our algorithm, first, we retrieved a list of relevant ICD-9 codes from the Medicare Chronic Disease Warehouse for 14 of the 15 chronic diseases, and utilized three ICD-9 codes for *obesity*. Then, we queried the Bioportal ontologies in the National Center for Biomedical Ontology for synonyms mapped to the set of ICD-9 codes. Next, we developed a text matching algorithm to search for the comorbid disease synonyms in the CPGs' Recommendation sections. Finally, we tabulated the proportion of CPGs mentioning comorbid disease terms, and the number of comorbid disease terms mentioned in each CPG. To evaluate the automated approach, 30 CPGs (two for each chronic disease) were selected using stratified random sampling, and two annotators with medical expertise performed sentence-level manual annotation to generate a reference standard.

Results: We obtained 2,503 unique CPG summaries from the NGC; 314 met inclusion criteria, and after processing to build the text corpus, 273 were available for analysis. Guidelines for concordant diseases (diseases that are part of the same pathophysiologic risk profile), such as hypertension, diabetes, and hyperlipidemia, mentioned one another most frequently. Hypertension was the only disease mentioned across all CPGs, while Alzheimer's disease and osteoporosis were mentioned the least. Annotators agreed on 96.95% of the 2,891 sentences in the pilot reference standard.

Conclusions: We developed an automated method that identifies comorbid disease terms in CPG recommendations. Evaluation of the method against the reference standard is in progress. Also, an annotation guide to improve the reference standard is in development. This method may be useful to describe important comorbidity relationships, inform gaps in guideline recommendations regarding comorbid diseases and therefore identify opportunities for guideline improvement. Further investigation is needed to understand the context and variation of comorbid disease mentions in CPGs.

The Challenges of Disparate Data Formats: Analysis and Visualization in the SLIDES Project

Ang Li[1], Steven Chall, MS[2], Allison Vorderstrasse, DNSc[1], Constance M. Johnson, PhD[1]

[1]Duke University, Durham, NC; [2]Renaissance Computing Institute, Chapel Hill, NC

Abstract

Type 2 Diabetes (T2D) is a chronic disease epidemic in the U.S. Innovative interventions that empower patients in diabetes self-management (DSM) are needed to decrease morbidity and mortality associated with this disease. We developed and tested a virtual environment (VE) for adults with T2D and collected multidimensional data over a period of 6 months. We employed visualization techniques to reveal thematic patterns among our data.

Introduction

Visualization of multidimensional data provides multiple views of the data beyond the traditional 2D perspective. The use of color, shapes, volume, and animation for example allows a greater exploration of the data through visual images. These dynamic drawings can uncover evolutionary paths of change over time. Using animation allowed us to combine multiple types of data and to highlight aspects of our temporal data. As the interactive web becomes a popular mode of delivery of health information, we need interactive research tools to better understand the outcomes of these studies.

Methods

Using data from the SLIDES (*eHealth: Second Life Impacts Diabetes Education & Self-Management)* study, we developed an Internet visualization tool that allowed dynamic data processing with an easy-to-use and implement user interface to evaluate the preliminary effects of VE participation on metabolic outcomes. The kinds of data generated included: text transcribed from audio recordings of participant conversations; numerical avatar positions within the VE; biometric measurements; survey results; images and video. Initially, visualizations were created using the capabilities available in the R environment, an open-source data analysis and statistical computing software. These capabilities were extended by incorporating R-Shiny libraries, which include a larger repertoire of visualization approaches plus web deployment. We next applied the D3.js JavaScript libraries, which allowed us to animate our results and added more versatility and power.

Results

There were a total of 20 participants in this pilot study with T2D. This visualization (Figure 1) allowed us to see through animation how the participants weight, HbA1c (indicator of metabolic control) and corresponding VE activity level changed over the course of the study (6 months). Figure 1 shows screen shots from a browser-based visualization at baseline, 3 and 6 months with color indicating individual participants. The horizontal axis is weight and the vertical axis is HbA1c. The size of each circle is determined by the corresponding subject's activity level.

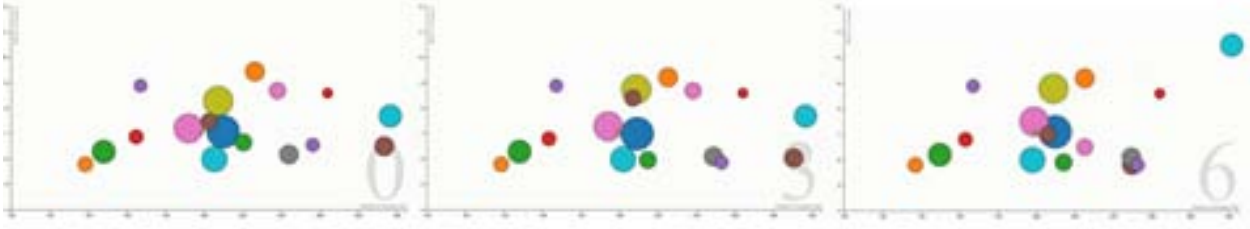


Figure 1. Visualization of participant weight, HbA1c and activity level.

Conclusion

Spatially transforming these different types of related data revealed trends and phenomenon that were not possible to see with traditional graphs. Visually representing these data with animation not only made salient information more apparent, but also allowed us to enhance our understanding of how self-management of T2D can be impacted by time and group dynamics. Major challenges remain in integrating the variety of data gathered in the SLIDES study including problems with rendering raw audio data into a format suitable for analysis, extracting meaning from the resultant text, and especially performance. Finding solutions to these problems could be beneficial to researchers using mixed methods approaches to evaluate healthcare big data. These visualizations help to show how data clusters and in turn helps researchers to see new dimensions to these data.

Reducing the Screening Burden of Systematic Review with a Multiple-level Relevance Ranking System

Dingcheng Li, PhD[1], Zhen Wang, PhD[2], Feichen Shen[1], Mohammad Hassan Murad, MD[1], Hongfang Liu, PhD[1]
Mayo Clinic, Rochester, MN[1], University of Missouri, Kansas City, Missouri[2]

Abstract. *Systematic review (SR) requires time-consuming manual work in screening potentially relevant articles retrieved from multiple databases. In this paper, we propose a multiple-level SR supporting framework including three components: i) the use of relevance ranking to assist the screening process aiming to increase the efficiency without compromising the validity, ii) topic analyses for distributed semantics discovery, and iii) network analyses on relation extracted for comprehensive semantic summarization. The sensitivities on a case study reached above 80% while the screening burdens were lowered to 25% even based on a crude relevance ranking approach. Topic analyses based on Latent Dirichlet Allocations (LDAs) showed the high consistency among the domain experts' selection of articles and relation network analyses displayed clearly the important predicate relations between medical concepts.*

Background and Introduction.

There is an increasing recognition that healthcare providers, researchers, and government agencies should use the body of research evidence to inform their decision making, from the care of individual patients to country-level decisions [1, 2]. As the cornerstone of evidence-based medicine, systematic review (SR) identifies, appraises, and syntheses all literatures regarding to a question of interest in a transparent and systematic way. However, high-quality SRs follow strict procedures (Figure 1) and require significant resources and time. Allen et al estimated that a SR with 1000 potential studies retrieved for abstract screening was predicted to take 952 working hours to complete. A recent evaluation of 63 SRs with 114 reviewers found on average a reviewer spent 0.9 minutes, 7 minutes and 53 minutes on abstract screening, full text screening and data extraction. Thus, methods to increase efficiency of conducting SRs without compromising validity are essential and strongly required.

Methods.

Inspired by our prior work on automated reference assignment [3] , which explores methods for assigning reference automatically to expert-written content, we propose the use of diverse relevance ranking metrics for accelerating the screening process of SR. In addition, we include deep semantic analyses into the SR framework. The first component functions as the screen process in traditional SR, composed of five relevance-ranking metrics: Semantic Relevance Ranking, Journal Relevance Ranking, Citation Relevance Ranking, Publication Type Relevance Ranking and Mesh Concepts Relevance Ranking, plus the Protocol Screening, which form a pipeline workflow. The second component is topic analyses based on Latent Dirichlet Allocations (LDAs) [4] and the third component is the extraction of predicate relations from the semantic MEDLINE.

Results and Discussion.

We piloted this framework in one SR studies conducted by our group. The SR evaluated the effects of pharmacological interventions to prevent or delay the onset of type 2 diabetes. After the manual screening, 38 articles were selected. For DM Prevention shown in Table 1, the total number of articles is 700, the total screening burden, which can still keep 100% of sensitivity using 350 articles and the screening saved is 50%. After we further decreased the screening burden, the sensitivity started to decrease. But even after we reduced the screening burden to 200, the sensitivity reached more than 80% for each case. We found no significant difference on combined effect sizes between different screening burdens and manual screening. However, we noticed the accelerated decreasing speed of screening sensitivity when lowering the screening burden from 200 to 100, which requires further investigation. In addition, the topic analyses demonstrate that the documents after screening focus on a small of topics and majority of the topics are irrelevant to the target SR. Such analyses suggest an alternative way of performing SR screening where we can perform topic analyses first and then screen based on topics. Finally, we performed the network analyses on the final list of articles for SR, which shows clearly the TREAT and PREVENT relations and gives us comprehensive information. All of those results demonstrated that our information retrieval system based on diverse relevance ranking can reduce the labor of SRs to a large degree while keeping comparably high recall.

Table 1. Performance in retrieving relevant articles for DM Prevention

Ranking threshold	False negatives	Sensitivity	Screening saved in percentages	Combined effect size (RR) and 95% confidence interval	Difference between Diverse Relevance Ranking and manual screening (p value)
700	0	100%	0	0.92 (0.86, 0.99)	1.00
350	0	100%	50%	0.92 (0.86, 0.99)	1.00
200	7	82%	71.5%	0.94 (0.88, 1.02)	0.68
100	14	61%	85.7%	0.96 (0.87, 1.05)	0.53

Reference:

[1] D. L. Sackett, W. M. Rosenberg, J. Gray, R. B. Haynes, and W. S. Richardson, "Evidence based medicine: what it is and what it isn't," *BMJ: British Medical Journal*, vol. 312, p. 71, 1996.

[2] L. Chambers, "Evidence-Based Healthcare: How to Make Health Policy and Management Decisions," *CMAJ: Canadian Medical Association Journal*, vol. 157, p. 1598, 1997.

[3] D. Li, L. H, C. CG, and J. S, "Towards Assigning References Using Semantic, Journal and Citation Relevance," in *The IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Shanghai, 2013.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003.

Automated Heparin Nomogram System

M. Li*, MD; N. Ou Pharm.D**; R. Wendt***, BA; M. Foley*** MS; P. Daniels****, MD;
V. Herasevich*, MD, PhD; P. Messner, *****, DNP, RN, CAN-BC; L. Oyen**, Pharm.D;
*Multidisciplinary Epidemiology and Translational Research in Intensive Care (METRIC), **Pharmacy Department,
***Population Management Systems, ****General Internal Medicine, *****Nursing Administration Department
Mayo Clinic, Rochester, MN

Abstract: To facilitate and improve the safety, accuracy, and efficiency of heparin management we designed and developed the Heparin Nomogram System (HNS) – a web based computer application which provides the capability to calculate IV heparin doses, electronically order activated partial thromboplastin time (aPTT) tests and generate reminders through the use of built-in safety check logic. HNS has been used for nearly 50,000 episodes of heparin therapy at Mayo Clinic Hospital Rochester Campus.

Introduction: Heparin has been widely used in a variety of settings to prevent and treat thromboembolic events for over half a century and the level of the systemic anticoagulation is a key determinant of the clinical outcome. The aPTT has been used to monitor the anticoagulant response of heparin, but it is extremely difficult and error prone to establish and maintain the therapeutic level of anticoagulation for an individual patient due to the complicated pharmacodynamics, patient-specific characteristics and the nature of heparin administration which creates many situations where medication errors and patient mismanagement can occur [(1).] The data from the United States Pharmacopeia MEDMARX database shows a total of 59,316 medication errors related to anticoagulants were reported from 2003 to 2007 and more than 17,000 involved heparin. Of the 556 errors (3.1%) resulted in harm to patients including seven deaths [(2).] Studies [(3)] have shown that a computer based system can increase the effectiveness of heparin administration.

To overcome the challenges and improve heparin management we have designed and developed the HNS system which automatically calculates IV heparin doses based on the patient's weight and aPTT results, electronically orders aPTT tests, and generates reminders to nurses through the use of built-in safety check logic to flag patients with abnormal pertinent lab results. HNS also checks drug interactions, duplicate orders and heparin allergy.

The objective of this project is to improve the efficiency and accuracy of the heparin administration through automating and monitoring the process by an in-house developed computer system.

Methodology: HNS is a web based computer application implemented in classic ASP technology that gets data from the Hospital Rules-Based System (HRBS) data warehouse, a relational Microsoft SQL database that integrates a near-real-time copy of clinical and administrative data from heterogeneous and distributed Mayo Clinic administrative and medical record systems.

HNS provides functionalities to allow the users to start or terminate the heparin treatment following prescribed goal aPTT and indication, and then periodically scans the data in the HRBS warehouse using predefined algorithm rules based on the weight-based heparin nomogram and selection criteria to determine if actions are needed.

Results: HNS has significantly reduced the incidents of heparin events more than10 fold over past 14 years. It has overcome the challenges of heparin management by using the following process:

- Allow the authorized users to start the heparin treatment by selecting the patient and a standard nomogram (high, intermediate or low intensity) with dose checking to the pharmacist order profile
- Calculates doses with decision-support for erroneous variables
- Initiates aPTT electronic order at 6 hours after starting or changing doses until a therapeutic dose is achieved (two consecutive values in range if within first 24 hours of HNS), then once daily
- Tracks and displays the aPTT order history and status
- Displays the anticoagulation and bleeding test results, such as patient's aPTT, hemoglobin and platelets in concise form to managing nurse
- Employs decision-support for spurious labs, potential heparinized lab samples, incorrect patient weights used, and potential side effects of heparin to the nurses
- Allows systematic analysis of recent changes and the three separate goal-based intensities of heparin used
- Serves as the interface of the quality data repository for analytics and quality improvement activities

Conclusion: HNS facilitates systematic support for and analysis of heparin therapy management. It provides an efficient workflow and safety solution to improve the efficiency and accuracy of heparin management by automating the dosing and lab ordering. It significantly reduces the opportunities for the errors to occur through the use of built-in automatic aPTT test ordering, nurse-reminders, abnormal lab test flags and decision support mechanism.

References:

1. Bauer SR, Ou NN, Dreesman BJ, Armon JJ, Anderson JA, Cha SS, Oyen LJ. Effect of body mass index on bleeding frequency and activated partial thromboplastin time in weight-based dosing of unfractioned heparin: a retrospective cohort study. *Mayo Clin Proc. 2009 Dec;84(12):1073-8. doi: 10.4065/mcp.2009.0220.*
2. http://www.automedicsrx.com/publications/The_Joint_Commission_Sentinel_Event_Alert_2008.pdf
3. Oyen LJ, Nishimura RA, Ou NN, Armon JJ, Zhou M. Effectiveness of a computerized system for intravenous heparin administration: using information technology to improve patient care and patient safety. Armon JJ, Zhou M. Am Heart Hosp J. 2005 Spring;3(2):75-81.

Alerting System for Patients with Advanced Multiple Sclerosis

Yihua Li[1], Dorothy W. Curtis, M.Sc.[1], Esteban J. Pino, D.Sc.[2], Pablo Aqueveque, D.Sc[2],
[1]Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA,USA;
[2]Department of Electrical Engineering, Universidad de Concepción, Concepción, Chile

Overview

This abstract presents the design and implementation of the software components of a new non-invasive monitoring system for people with Advanced Multiple Sclerosis. The proposed system addresses issues in advanced MS such as low mobility and heat sensitivity, and consists of a central monitoring station for nurses and individual monitoring systems for each patient. It gives suggestions and generates alarms to the patients and aims to help caregivers in assisted living facilities provide better supervision. Our poster will show preliminary results from experiments testing the effectiveness of our tilting alerts and our heat exposure alerts.

Introduction

Multiple sclerosis (MS) is an autoimmune disease that affects the central nervous system. Patients often suffer from problems such as fatigue, low mobility and pressure ulcers. In addition, they might experience a temporary worsening of their symptoms when they are outside in hot and humid weather for long duration. To encourage patient independence and comfort and also help caregivers to improve patient care and supervision, a multi-sensor monitoring system for people with advanced Multiple Sclerosis has been designed and implemented[1]. Several non-invasive sensors are deployed on an electric wheelchair and they constantly collect Ballistocardiogram (BCG) signals, pressure signals, accelerometer data, ambient temperature and relative humidity. The acquired data is stored in PostgreSQL database in a laptop mounted on the wheelchair. We further develop the system and build software components that utilize the data captured by the sensors to generate alarms related to the MS patient health status.

System Description

The monitoring system consists of three main parts: (1) the laptops in individual wheelchairs; (2) a central monitoring station; (3) and a webserver that links the two and transmits the data from the laptops to the central station via a wireless network. The software program on each individual wheelchair analyzes the sensor data and generates audio alarms for the patient. These audio alarms need to have some variety since the patient might become accustomed to a particular alarm and ignore it. The programs can work in a standalone mode for alerting the patients and we also implement a GUI with PyQt that allows caregivers to quickly assess patients' degree of exposure to heat.The central monitoring station is designed mainly for the caregivers to monitor patients' health condition remotely. The user interface (UI) displays information relating to patient's health status, and gives visual alarms to the nurses when necessary. We make the UI web-based to provide real-time display and develop it with techniques and technologies such as html, javaScript and jQuery library.

Alert Generation

The system generates alerts to prevent excessive heat exposure. The detection parameter is a cumulative heat index, computed from temperature and humidity over time. The alarm generation threshold is chosen by the caregiver. Initial alerts tell the user how much heat exposure they have had and how long they can stay outside. Later alerts remind the user that he/she should return indoors. Alerts are also used to remind patients when they should tilt their wheelchairs to prevent pressure ulcers. The schedule of these alerts is determined by the caregiver.

Conclusion

We have designed and developed the software components of the monitoring system that generates alerts relating to MS patient's health status from the sensor data. These alerts, together with the relevant health information, will help prevent risk conditions for the MS patients and assist the nurses to give better patient care. We plan to deploy and test the final complete system at The Boston Home, Dorchester, MA.

References

1. Pino E, Arias D, Aqueveque P, Vilugrón L, Hermosilla D, Curtis D. Monitoring technology for wheelchair users with advanced multiple sclerosis. Conf Proc IEEE Eng Med Biol Soc 2013; 2013:961-4.

Identification of Common Concepts for Clinical Decision Support and Mapping to the Health Level 7 Virtual Medical Record Data Model

Yanhua Lin, PhD, Brandon Welch, PhD, Tyler Tippetts, BS, Polina Kukhareva, MS, David Shields, MS, Catherine Staes, BSN, MPH, PhD, Vijay Kandula, MD, Bruce Bray, MD, Kensaku Kawamoto, MD, PhD

Department of Biomedical Informatics, University of Utah, Salt Lake City, UT

Abstract

Standard terminologies are a critical resource for enabling interoperable clinical decision support (CDS). However, the enormous quantity of concepts available in these terminologies makes it imperative to identify concept subsets relevant for CDS. To address this problem, a domain expert-driven approach was used to systematically identify concepts relevant for CDS, and these concepts were mapped to relevant attributes of the HL7 Virtual Medical Record CDS data model. The resource developed is freely available through www.opencds.org.

Introduction

A key challenge for clinical decision support (CDS) is the difficulty of scaling CDS across institutions.[1] Standard terminologies can facilitate the achievement of interoperable, cross-institutional CDS by allowing the mapping of local data representations to standard concepts used by the CDS knowledge base. Currently, a standardized data model for CDS is available in the form of the HL7 Virtual Medical Record (vMR) standard. Also, many standard terminologies such as SNOMED-CT, RxNorm and LOINC are available for use in CDS. However, the sheer volume of concepts in these terminologies can make it challenging to ensure that different CDS implementers choose the same concepts in their respective implementations. For example, SNOMED-CT alone includes over 300,000 concepts. The objective of this effort was to identify a subset of concepts from standard terminologies likely to be relevant for CDS, so as to form the foundation of a more manageable corpus of concepts used commonly across CDS implementations and mapped to relevant attributes of the HL7 vMR data model.

Methods

We systematically analyzed all coded data elements in Release 1 of the HL7 vMR standard and then developed a strategy for identifying the concepts most relevant for CDS in each of these areas. This strategy involved two steps: the manual identification of candidate concepts by the project team followed by concept review by a physician informaticist. For example, in the first step, we identified candidate problems from the CORE Problem List subset of SNOMED-CT[2] and candidate laboratory results from LOINC's "top 300" orders.[3] We also reviewed value sets available in the Public Health Information Network Vocabulary Access Distribution System (PHIN VADS) and included value sets for Encounter, Problem Importance, Severity, Gender, and Race, for example. In the second step, a physician informaticist (VK) reviewed each concept in the candidate set and identified those that had a reasonable likelihood of being useful for CDS purposes in a typical clinical care setting. In addition, we included concepts from quality measurement specifications, such as Surgical Care Improvement Project (SCIP) and Healthcare Effectiveness Data and Information Set (HEDIS) measures. Each of these concepts were then uploaded into an Apelon DTS terminology server as CDS-related concepts managed by the OpenCDS initiative (www.opencds.org).

Results

Approximately 2,200 clinical concepts were identified spanning 74 vMR attributes, including Adverse Event Type, Encounters Type, Goal Type, Observation Focus, Problems, Procedures, Medications, and Supplies. A terminology mapper was also developed to map concepts from various standard terminologies (e.g. SNOMED-CT, ICD9, and LOINC) to this CDS concept taxonomy. These concepts and tools are now being used in the OpenCDS initiative and are freely available to this collaborative open-source community.

Conclusion

A CDS concept taxonomy can facilitate the use of standard terminologies to enable scalable and interoperable CDS.

---

[1] Osheroff JA, Teich JM, Middleton B, Steen EB, Wright A, Detmer DE. A roadmap for national action on clinical decision support. JAMIA. 2007; 14:141-5.

[2] National Library of Medicine. The CORE Problem List Subset of SNOMED CT. http://www.nlm.nih.gov/research/umls/Snomed/core_subset.html.

[3] LOINC. Common LOINC Results and Order Codes. http://loinc.org/usage.

Iterative Participatory Design of Health Information Technology for Underserved Populations

Elizabeth A. Linton[1], Tina Kurtz, R.N[1], Kim M. Unertl, PhD[1]
[1]Vanderbilt University, Nashville, TN

Abstract

We describe how researchers working with the Vanderbilt-Meharry-Matthew Walker Sickle Cell Center for Excellence (VMMW) applied iterative participatory design (IPD) of health information technology to transition paper-based sickle cell health management tools into electronic tools within an EHR. IPD proved labor intensive, but yielded high usability and stakeholder satisfaction.

Introduction

(IPD) is "a process of mutual learning by designers and domain experts (users) ... [aimed] at changing the users' work practices ... [by introducing] information systems" [1]. Research shows that health information technology (HIT) that interrupts workflow or is confusing to use suffers from low uptake. End users also develop ad-hoc workarounds with unintended consequences, and systems may fall short of achieving intended purposes [2]. An iterative participatory design (IPD) approach to product development has the potential to help avoid these pitfalls.

Methods

We conducted a qualitative study of HIT's place at the Vanderbilt-Meharry-Matthew Walker Sickle Cell Center for Excellence (VMMW), documenting over 250 hours of interactions between patients, healthcare providers, and technology. Members of the research team shadowed health care providers during patients' appointments and in the doctors' office as notes were documented. We paid special attention to issues arising around the electronic health record (EHR), workflow, communication, and decision-making. Healthcare team members identified translating paper forms into the electronic medical record as a key area for improving workflow decreasing EMR workarounds. Figure 1 summarizes the four-stage IPD process used during the project.

1) Communally ***[RE] EVALUATE*** strengths and weaknesses of patient care current tools	2) ***[RE] DESIGN*** working model of tool to address weaknesses and incorporate pre-existing strengths
4) ***[RE]TEST*** privately, during team meetings, and during real clinic visits to verify and validate the tool	3) ***[RE] BUILD*** tool only after establishing clear but reasonably flexible usage criteria and requirements

Figure 1. Iterative Participatory Design with VMMW

Results

Clinic observation revealed that the sickle cell healthcare providers heavily employed paper forms like a Sickle Cell Pain Action Plan as EHR workarounds. VMMW providers identified these forms as a necessary aspect of patient care that could be improved if available within the EHR. IPD informed the successful transition of seven sickle cell health management tools from paper forms to computerized tools in the EHR. Resources needed to facilitate IPD include: buy-in from both the informatics and healthcare teams, an individual assigned to build the tools, and a regular venue where drafts are communally tested and edited. The electronic forms addressed difficulties associated with paper versions including: multiple editions in use simultaneously, illegible scanned versions in the EHR, and difficulty accessing care plans for patients outside the clinic setting. Computerized documents were used ~400 times; feedback from revision sessions (about 3 per form) indicated they were well received.

Conclusion

Iterative participatory design allows patients, medical team members, and informatics teams to partner in improving HIT monitoring and delivery. However, IPD's labor intensive process is a potential limiter of widespread use. Finding routes to refine the efficiency of IPD in practice (especially in limited resource settings) could make this approach valuable in future biomedical informatics research.

References

1. Simonsen J, Hertzum M. Iterative participatory design. In: Simonson J, Baerenholdt JO, Buscher M, Scheuer JD, editors. Design research: Synergies from interdisciplinary perspectives. London: Routledge; 2010. p. 16-32
2. Flanagan M, et al. Paper- and computer-based workarounds to electronic health record use at three benchmark institutions. *JAMIA* ; 2013. p. 59-66 http://jamia.bmj.com/content/20/e1/e59.long

Developing Clinical Decision Support for Direct Use by Patients: Challenges and Lessons Learned from Symptom Management in Cancer Patients

David F. Lobach, MD, PhD, MS[1]; Janet L. Abrahm, MD[2]; Donna L. Berry, RN, PhD[2]; Michael S. Rabin, MD[2]; Ilana Braun, MD[2]; Manan Nayak, MA[2]; Mary E. Cooley, RN, PhD[2]
[1]Religent Health and Duke University Medical Center, Durham, NC; and
[2]Dana-Farber Cancer Institute, Boston, MA

Abstract

One approach for engaging and empowering patients is through the development of clinical decision support (CDS) tools designed for direct use by patients. In this project we assessed the feasibility of creating CDS tools for the self-management of symptoms in cancer patients. Patient safety and health literacy issues were addressed. Additionally, in order to provide suggestions for symptom relief, recommendations for medication therapy were structured to reference existing clinical management practices of the overseeing clinician.

Introduction

Relatively little is known regarding CDS for patients with cancer and few studies have described CDS tools for direct use by cancer patients[1,2]. In these studies, the focus of the CDS was to identify the presence or absence of a specific symptom through an algorithm with a single decision node followed by general recommendations for managing that symptom[1] or through heavy reliance on interactions with clinical personnel[2]. The literature provides no precedents for developing more complex self-management algorithms for cancer patients. The goal of this project was to understand the needs of patients with regard to CDS, along with identifying effective mechanisms for delivering patient guidance for self management of their symptoms and direction for when to call their clinicians.

Methods

The first phase of the project was to understand patient and clinician perspectives regarding symptoms and quality of life experiences during cancer treatment and the second phase was to develop resources such as CDS tools for direct use by patients. In phase 1, patient and clinician perspectives were ascertained through focus groups and semi-structured interviews. Participants included a total of 57 adult patients who had received cancer treatment within the past six months and 51 clinicians who provided care in ambulatory oncology settings. For phase 2, oncologists, palliative care specialists, and nurse scientists drew from evidence-based resources and worked with CDS experts to develop computable algorithms that would enable patient self-management for pain, constipation, and nausea and vomiting. The clinical algorithms were converted to patient friendly language and an interactive format in order to assess usability and comprehensibility through focus groups, interviews, and patient-directed think aloud sessions.

Results: Challenges and Lessons Learned

Significant challenges were identified in developing CDS tools for symptom management in cancer patients. A primary concern was maintaining patient safety by identifying potential serious or life-threatening causes for symptoms and directing patients to seek contact with their clinicians. Accordingly, all three patient CDS algorithms began with the identification of "red flags" that would necessitate that the patient exit the CDS tool to seek clinician guidance. A second challenge was to support users across a wide breadth of health literacy. To address this challenge, we developed content at a fifth grade reading level and provided descriptions and educational information that users could elect to review or skip based on perceived information needs.

In order to provide medication-related advice, we sought to ensure that recommendations were evidence-based and grounded in therapies approved by the clinician overseeing a patient's care. Accordingly, we inquired whether a patient was using a particular therapy or if he had been prescribed that particular therapy. If he was not using the therapy, he was advised to use it. If he had not been prescribed the recommended therapy, he was advised to contact his clinician and inquire if this therapy could be appropriate.

Conclusion

Through this project, we have shown an approach for providing CDS for patient self-management in cancer care including provisions for patient safety and recommendations for medication use for symptom control.

Acknowledgements

This project was funded in part by The Patient Centered Outcomes Research Institute Grant PI-12-001.

References

1. Head BA, Keeney C, Studts JL, Khayat M, Bumpous J, Pfeifer M. Feasibility and acceptance of a telehealth intervention to promote symptom management during treatment for head and neck cancer. *J Support Oncol.* 2011;9(1):e1-e11.
2. Weaver A, Young AM, Rowntree J, *et al.* Application of mobile phone technology for managing chemotherapy-associated side-effects. *Ann Oncol.* 2007;18 (11): 1887-1892.

Towards a Representation Format for Sharable Self-Monitoring Data

Guillermo H. Lopez-Campos, PhD, Manal Almalki MSc, Fernando Martin-Sanchez, PhD
Health and Biomedical Informatics Centre, The University of Melbourne, Melbourne, Australia;

Abstract

Self-monitoring devices and apps are becoming increasingly popular for research purposes. In the context of biomedical research, metadata associated with experimental studies has become a key element to make sense of heterogeneous and very complex sets of data. In this work we discuss the need for a reporting guideline to describe data and metadata generated in self-monitoring experiments in a standardized way. This data representation format could enable the reproducibility of experiments and improve the interpretation of data collected using "self-monitoring" strategies.

Introduction

In recent years advances in technology have enabled the development of a vast amount of increasingly miniaturized sensor devices and mobile apps that can be used for monitoring an increasing number of features such as physical activity, physiological parameters, behaviour or exposure to environmental factors following a "self-tracking" or "self-monitoring" approach. These approaches have been successfully used in different scenarios in clinical research (i.e. self-tracking of physical activity[1]). Although they represent an opportunity to gather highly valuable patient-generated data in a continuous manner, also pose challenges for biomedical informatics due to the need to integrate new data sources with different formats from different devices. Another important issue that needs to be addressed is the interpretability of those data. By providing the appropriate metadata associated with the measurements we can increase the usefulness of these data for clinical research

Methods

Based in previous successful experiences in bioinformatics with the development of Minimum Information Guides for different research areas[2,] we propose the development of a similar guideline for the use of self-monitoring methodologies for research purposes.

The proposed guideline is structured around the ISA (Investigation-Study-Assay) concept and uses five major axes to describe the monitoring activities. The first axis covers a experimental description of the experiment design. The second axis focuses on the description of the "sample" used in the study, either a whole individual (i.e. for physical activity measurements) or an element of the body. The third axis describes the annotation and description of the device/s used during the monitoring process. The fourth element captures the "measurement", that is what is measured and the procedures related with how "sample" and "device" interact with each other. Finally the fifth and last axis focuses in the collected data and their annotation. During the annotation of each of these axes, the use of appropriate existing ontologies and controlled vocabularies is encouraged and required.

Conclusion

This representation format should be firstly adopted by the research community to ensure that data collected through "self-tracking"/"self-monitoring" can be compared and easily shared. As it happened before with other minimal information guidelines its adoption would then enable and support the development of public data repositories for these studies ensuring data accessibility and enabling meta-analyses. Device manufacturers could be involved in the development of the guideline making their data formats compliant with the proposal to ensure compatibility across platforms. Finally, the existence and adherence to such minimal information guideline could facilitate data exchange between the self-trackers community and researchers in a structured way and providing enough metadata to the researchers to apply those data enlarging annotated datasets for research.

References

1. Takacs J, Pollock CL, Guenther JR, Bahar M, Napier C, Hunt MA. Validation of the Fitbit One activity monitor device during treadmill walking. J Sci Med Sport. 2013 Oct 31, pii: S1440-2440(13)00472-6
2. Taylor CF. Standards for reporting bioscience data: a forward look. Drug Discov Today. 2007 Jul;12(13-14):527-33

Physician Satisfaction With Computerized Order Entry

Paul Loubser; MB, ChB; University of Texas Medical School at Houston, Houston, Texas
Val Hooper; PhD, MBA, Hoer Bibl Dipl, BA; Victoria University of Wellington, New Zealand

Introduction: Computerized e-ordering is rapidly becoming the norm in many healthcare institutions. The Department of Anesthesiology implemented such a system one year ago. We elected to determine the success of this system by surveying user satisfaction, i.e. the faculty and residents, with the system. The factors that contributed to user satisfaction were assessed to ascertain which exerted the greatest influence on the users. The relative importance of each attribute of the system would provide an indication of where developers of such systems would need to focus their attention in order to ensure greatest satisfaction with the system in the future.

Methods: A 10 question 'satisfaction' survey was conducted amongst 128 faculty and residents of which 39 faculty and 38 residents responded. Structural equation modelling (SEM) was used to analyse the data. All the respondents' data were analysed together first and then the residents and faculty's data were analysed separately. The research model was based on the Technology Acceptance Model (TAM) (Davis et al., 1989) and extensions thereof (Venkatesh & Bala, n.d.; Venkatesh & Davis, 2000; Venkatesh et al, 2003).

Results: The R^2 in each instance represents the extent of the variance in the relevant dependent variable attributable to the independent variables. Thus, 'reducing medication errors' was only responsible for 1.2% of the ability to render patient care in the combined sample, 3.5% amongst faculty only, and a much larger 13.4% amongst residents. Together, the ability to render patient care, its user friendliness, and the available IT support, contributed to a large 76.9% of the variance in the satisfaction of the combined sample of the users, 78.5% amongst faculty and 75.7% amongst residents. In the combined sample, the user satisfaction, together with the perceived receptivity to the system by others, accounted for 19.8% of the variance in the extent to which expansion of the e-ordering system to other services was welcomed; 18.4% of the variance amongst faculty; and 17.6% amongst residents. The paths from IT support to satisfaction, and from satisfaction to welcoming expansion to other services were not significant, even though the paths between IT support and satisfaction for the faculty, and between satisfaction and welcoming expansion were strong enough. For the combined sample, the paths between rendering patient care and satisfaction, between user friendliness and satisfaction, and between receptivity and welcoming expansion, were particularly strong and significant. The attributes of the e-ordering system that had been identified as contributing to a large percentage of the impact on satisfaction with the system were user friendliness and the ability to render patient care. The user friendliness of the system was the overwhelming attribute that influenced the satisfaction with the e-ordering system. With a similar impact on satisfaction amongst the combined sample, and slightly less so amongst the faculty and residents, the ability to render patient care was minimally influenced by the perceived reduction in medication errors.

Conclusion: The importance of the user friendliness and the ability to facilitate rendering of patient care emerged as the main attributes sought from such a system. For the system to be rolled out to other services, these are the two main attributes that influence system acceptance and adoption. IT support did not seem to be a significant factor; it may be taken for granted by users. Although the satisfaction did not demonstrate a strong influence on the welcome of expansion of the system to other services, certainly the perceived receptivity of the system by colleagues played an important role in influencing that welcome to expansion. In addition, there were probably many other considerations such as the type of system and the type of work, plus the available resources, and the mindsets of the users of other systems, which would play a role. However, future research will explore the preparedness of the users of other services to adopt such an e-ordering system.

Using Element Words to Generate (Multi)words for the SPECIALIST Lexicon

Chris J. Lu, Ph.D.[1, 2], Destinee Tormey[1], Lynn McCreedy, Ph.D.[1], and Allen C. Browne [1]
[1]National Library of Medicine, Bethesda, MD; [2]Medical Science & Computing, Inc., Rockville, MD

Abstract
The SPECIALIST Lexicon has been distributed annually by the National Library of Medicine (NLM) since 1994. Lexical records are used for Part-of-Speech (POS) tagging, indexing, information retrieval, concept mapping, etc. in many Natural Language Processing (NLP) projects, such as Lexical Tools, MetaMap, SemRep, UMLS Metathesaurus, and ClinicalTrials.gov. This paper describes a new systematic approach to identify single words and multiwords from MEDLINE through the use of element words. Element words are lowercase single words without punctuation and are not stopwords. Results show an accelerated growth of the Lexicon, particularly an increase in multiword records. Hence, improvement in recall or precision can be anticipated in NLP projects using the SPECIALIST Lexicon and its applications.

1. Introduction – The NLP SPECIALIST Lexicon and LexBuild

The Lexicon is built by linguists through a web-based computer-aided tool, LexBuild [1]. Element words are a resource used by linguists to 1) add new Lexical records if no exact/close match is found in LexBuild; 2) update existing lexical records if related records are found by close match. Multiwords that contain these new element words are reviewed through the Essie search engine [2], Google Scholar, dictionaries, etc. during the LexBuild process.

2. (Multi)words by New Element Words from MEDLINE

For (multi)word inclusion in the Lexicon from MEDLINE, we developed this system: 1) retrieve element words through tokenization (lowercase, remove punctuation, and use space as word boundaries) from MEDLINE titles and abstracts; 2) categorize these element words by type: single words in the Lexicon (e.g. diabetes), not a single word but parts of multiwords already in the Lexicon (e.g. mellitus), numbers (e.g. five), digits (e.g. 5), non-words (e.g. 3h), and new element words (e.g. cdh); 3) calculate word count (WC). New element words with high frequency (WC >= 1500) are retrieved automatically for review to cover single words (97.58%) and multiwords from MEDLINE. For example, the new element word "cdh" (9983 WC) leads to 44 new lexical records with base forms in 78 single words (e.g. cadherin1) and 23 multiwords (e.g. "chronic daily headache").

3. (Multi)words by Existing Element Words from MEDLINE

This system also retrieves candidates of new multiwords from MEDLINE for (existing) element words: 1) Generate high frequency n-grams of length 1-5 from MEDLINE. The low frequency n-gram terms are filtered out if the associated (n-1)-gram terms have low WC of normalized form (NWC). 2). N-grams are normalized by abstracting away from genitive, punctuation, and case so that different forms of a same term are grouped together for further analysis. 3). Generate new candidate multiwords by applying a rule-based system to filter out invalid multiwords from n-grams. These rules exclude (normalized) n-grams that exist in the Lexicon, start/end with a preposition/auxiliary/modal/conjunction, end with determiner/acronym in a parenthesis, etc.. Document count, WC, and NWC are also used to filter out low frequent error prone n-grams (e.g. typos). Further development of these rules is intended to increase the precision of candidate multiwords. 4). These new candidate multiwords are reviewed by linguists, who add grammatical and lexical variant information, yielding completed Lexicon records. For example, the element word "mellitus" was identified in 24 multiword lexical records in the previous Lexicon release (2014). In our new approach, a candidate list (532) is retrieved automatically from 1304 n-gram terms containing "mellitus" after filtering out ~60% of invalid words. This list is then mapped into 390 normalized forms to ease the final review process in linguistic contexts. As a result, 36 new lexical records with base forms in 9 single words (including, actually, "mellitus") and 41 multiwords (e.g. "diabetic mellitus") have been added, a 150% growth from Lexicon.2014. In addition, 7 other associated existing records with 10 multiwords have been updated for spelling variants and acronym expansions. Please refer to the Lexicon web site for details [3].

4. Conclusion

There are 477K lexical records with 1.69M forms in the 2014 Lexicon release. About 47.7% (418K) of unique forms (875K) are multiwords. Multiwords are an essential ingredient and play a key role in the success of NLP tasks. This new system enhances the Lexicon's coverage, especially on multiwords. We expect the growth of multiwords in future Lexicon releases to reach the estimated value (50%) through this system [4]. This new system encourages rapid growth of single words as well as multiwords in the Lexicon, which will ultimately provide better NLP results.

References

1. C.J. Lu, L. McCreedy, D. Tormey, and A.C. Browne., "A Systematic Approach for Automatically Generating Derivational Variants in Lexical Tools Based on the SPECIALIST Lexicon", IEEE IT Professional Magazine, May/June, 2012, p. 36-42
2. N.C. Ide, R.F. Loane, D.D. Fushman, "Essie: A Concept-based Search Engine for Structured Biomedical Text", JAMIA, Vol. 14, Num. 3, May/June, 2007, p.253-263
3. http://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/lexicon/2015/docs/designDoc/UDF/medline/index.html
4. I.A. Sag, T. Baldwin, F. Bond, A. Copestake, D. Flickinger, "Multiword Expressions: A Pain in the Neck for NLP", Computational Linguistics and Intelligent Text Proc., Lecture Notes in Computer Science, Vol. 2276, 2002, p. 1-15

De-identified clinical research data warehouse in a Korean tertiary hospital

Yong-Man Lyu[2], Soo-Yong Shin, PhD[1,2], Yongdon Shin[2],HyoJoung Choi, RN[2], Jihyun Park[2],Eun-Ae Kang[2] , Woo-Sung Kim, MD, PhD[1,3],Chang-Min Choi, MD, PhD[23],Yu Rang Park[1], JaeHo Lee, MD, PhD[1,2,4]

[1]Department of Biomedical Informatics; [2]Office of Clinical Research Information; [3]Department of Pulmonary & Critical Care Medicine; [4]Department of Emergency Medicine, Asan Medical Center, Seoul, Korea

Abstract

Research environments are emphasized toward provisions of ethics in Korea. To comply with governmental regulations, all identifiers of patient must be removed from Electronic Medical Record. Asan Medical Center (AMC) have developed a de-identification clinical research system, called ABLE (Asan BiomedicaL research Environment), including cohort discovery, de-identified chart review, and data extraction tool. ABLE has about 4M registered patients with over 600M orders, 715M lab results, 257M clinical notes, and 4M DICOM images.

Introduction

According to the revised regulations in Korea, research environments are emphasized toward provisions of ethics in order to protect patients' privacy. This is in response to growing interest from researchers who want to use clinical data for study purpose. To comply with Korean regulations, AMC has developed a de-identified clinical research system named ABLE in accordance with domestic and international research regulations.

Method

Based on the previous studies (1, 2), we have developed ABLE in collaboration with inbrein company. We defined 19 internal identifiers and developed de-identification methods. The de-identification system achieved 99.87% accuracy and 96.25% recall for validation dataset which composed 5,000 clinical notes of 33 different note types (2). We tried to remove identifiers in PACS images as well as EMR. For the convenience of researchers, we developed three tools such as cohort discovery, de-identified data review and data extraction. All tools were developed using Microsoft.NET framework and C#. MS parallel data warehouse version 2 was chosen as data warehouse appliance with MS SQL Server 2012.

Result and Conclusion

ABLE system includes the about 4M patients' data since 1989. As a result, ABLE can search over 600M physicians' orders including over 191M medication orders. Also it can search over 715M laboratory results. The de-identification chart review tool can review over 257M clinical notes of 892 different kinds of clinical notes, over 3M CT images and over 1M MR images. The data extraction tool also supports keyword search within clinical notes. By developing an anonymising system, we have greatly increased the opportunity to use clinical data for study purposes under the new research regulations. The system described herein represents an important reference toward provisions of research ethics and protection of patient privacy in Korea.

References

1. Shin SY, Lyu Y, Shin Y, Choi HJ, Park J, Kim WS, Lee JH. Lessons learned from development of de-identification system for biomedical research in a Korean tertiary hospital.*Healthc Inform Res*.2013; **19**(2): 102-109.
2. Shin SY, Shin Y, Choi HJ, Park J, Lyu YM, Lee MS, Choi CM, Kim WS, Lee JH. De-identification method for bilingual clinical texts of various note types, *J Korean Med Sci.* 2014 (under review)

Use of Natural Language Processing in Terminology Coverage Analysis

Sina Madani, MD, PhD[1], Dean F. Sittig, PhD[2], Michael M. Riben, MD[1]
[1]The University of Texas MD Anderson Cancer Center, Houston, TX
[2]The University of Texas School of Biomedical Informatics at Houston, Houston, TX

Abstract
Interface terminologies designed by vendors have been used to facilitate provider friendly data entry at the point of care within healthcare organizations. While these terminologies with backend mappings to standard vocabularies may be advertised for data integration and decision support benefits, the content coverage should be evaluated against existing healthcare provider vocabulary usage. We investigated two interface terminologies against our clinical narrative repositories using natural language processing in order to evaluate the term coverage.

Introduction
Standard terminologies are rigid by nature and designed to be consumed by computer applications. On the other hand, interface terminologies are created to facilitate provider friendly term entry at the point of care while maintaining back end mappings to standard vocabularies. Many third party vendors are now producing such terminologies. However, health care organizations are unique in terms of their clinical service and documentation contents. Therefore, third party interface terminologies should be evaluated, in terms of content coverage, and customized or extended before and after acquisition respectively. We used Natural Language Processing (NLP) methods in order to evaluate two interface terminologies against our clinical narrative contents.

Methods
Clinical notes were extracted from MD Anderson Cancer Center repositories and preprocessed for identification of section headers related to the Past Medical History and Problem List. An NLP pipeline was developed and contents of the extracted section headers were processed by MetaMap with restriction to the "Disorders" semantic group, including "neoplastic" semantic type for cancer related concepts. XML outputs were converted to Resource Description Framework (RDF) and loaded into a local instance of AllegroGraph® triple store. The 50 most frequently occurring concepts were extracted using semantic queries and evaluated by a trained physician for cancer and non-cancer related terms against two interface (provider friendly) terminologies from Health Language, Inc. and Intelligent Medical Objects, Inc.

Results
A total of 291,139 Past Medical History (group 1) and 232,154 Problem List related section headers (group 2) were identified from the corpora of 420,557 clinical narratives. Under these sections, 2,525,527 non-cancer related terms (27,321 concepts) from group 1 and 158,392 cancer related terms (676 concepts) from group 2 section headers were extracted. An average of 6.18 and 6.36 synonyms per concept were calculated for group 1 and 2 respectively. Content coverage evaluation results within top 50 concepts (by frequency) are summarized in table 1 below.

	Non-Cancer Items		Cancer Items	
Total terms	662,217 (50 concepts)		148,278 (50 concepts)	
Unique terms	318 (50 concepts)		309 (50 concepts)	
Content Coverage	Terminology 1	Terminology 2	Terminology 1	Terminology 2
Number of unavailable unique terms	59	82	72	54
Percentage of total term coverage	88.67%	81.61%	92.97%	90.45%
Unavailable terms in *both* terminologies	39 (9.72% of the total terms)		32 (3.4% of the total terms)	

Table1. Term coverage analysis between the two provider friendly (interface) terminologies

Discussion
Although we didn't find any significant difference in content coverage between the two interface terminologies, the result of this study can be used for gap analysis and enable terminology providers to enrich their vocabularies. Such an approach could also enhance the usability of third party content sets that are prepared more toward consumer needs and create an opportunity for participation in the domain-specific and/or provider-friendly terminology development process.

Workflow-based modeling of cancer care trial protocols

Aisan Maghsoodi[1,2], Anca Bucur, PhD[2], Paul de Bra, Prof, Dr[1], Norbert Graf, Prof, Dr[3]
[1] Technical University of Eindhoven, Netherlands; [2]Philips Research Europe, Netherlands;
[3] Saarland University, Germany

Abstract

A large number of cancer patients are treated in the context of clinical trials, for which free-text evidence-based protocols aiming at standardization of care delivery exists. We propose a workflow-based representation model for cancer trial protocols, so that information relevant to each active process can be directed to the intended role. This modelling approach enables monitoring patients with respect to the protocol-based care path, as well as retrospective analysis of care process to understand deviations and sources for delays and bottlenecks.

Introduction

Research on computer interpretable clinical guidelines has largely focused on individual tasks rather than complex processes of care which are extended in time[1]. CDS systems have mainly focused on using guideline knowledge for decision making at a time point. We have focused on cancer care trial protocols (CCTP) which contain detailed description of care methods like complex treatment plans over a period of time delivered in different care settings, along with study-related processes to be adhered to, such as reporting. CCTPs are meant to standardize trial-based care processes across participating institutions. Our modeling methodology enables generation of CCTP-based systems for monitoring the patients' progress and providing next step guidance for clinical actors in different setting by aggregating and presenting only the relevant role, activity and process-specific information spread in the document. Analysis of the execution of such a system will lead to better comprehension of the problems. The frequent deviations can be indicators of problems in protocol that can trigger new hypothesis for research.

Methodology

We investigated protocols from two different domains: Wilms Tumor (SIOP), Leukemia (AIEOP-BFM ALL) in order to find similarities of CCTPs and to test the model's ability to be generalized. For each cancer domain, common processes and the critical path for all centers are identified. Our model includes an entity model, such as medical named entities, trial entities and their specific attributes. The second part includes care and study-related processes, activities and their context attributes. Domain-specific processes, their order, dependencies, resources and events were represented using business process modeling notation (BPMN) and checked with clinicians for validation.

Using BPMN CCTPs can be intuitively represented as collection of linked processes in their context such as resources, communications, events that can occur such as errors, exceptions, delays, local adaptations and activity or entity states.

To provide link back to original text as evidence for activities in the model, CCTP was processed using NLP and techniques together with SNOMED and UMLS annotations.

Conclusion

This methodology enables directing tasks, pre and post conditions, and communications as well as evidence to the intended role in the care process. Also, it caters for locating the patient in the care process.

The CCTPs models generated can be shared at the level of domain elements and main processes. By identifying the critical paths that need to be followed by all canters in all countries, local adaptations can be addressed by allowing non-prohibited changes. The modeled CCTP is adaptable to updates as the changes can be easily assigned to its related process and affected process elements be tracked using the model. Inter-operability with EHR has not been the focus since many participating trial centers do not obtain an EHR, but using standard vocabularies in entity model provided hooks for mapping the data model. Colleting the event logs after the execution of enables retrospective analysis of deviations and delays. The methodology can be communicated to the authoring committees to be used in authoring templates which can further facilitate the automatic model element extraction.

References

Patkar, V., South, M., & Thomson, R. From guidelines to careflows: modelling and supporting complex clinical processes. Computer-based medical guidelines and protocols: a primer and current trends.2008;139, 44.

A Usability Analysis of a Fuzzy Match Search Engine for Physician Directories

Mahnke, Andrea[1], MS; Baker, Kate[1], MA; Krause, Tim[2], PhD; Christian, Courtney[2]; Hale, Katy[2]; Kautz, Karissa[2]; Mueller, Elizabeth[2]; Majid Rastegar-Mojarad[1], MS; Waltonen, Stuart[3], PhD, ABPP; Kortenkamp, Sarah[3], PhD; Lin, Simon[1], MD
[1]Marshfield Clinic Research Foundation, Biomedical Informatics Research Center, Marshfield, WI
[2]University of Wisconsin Stevens Point, Stevens Point, WI
[3]Marshfield Clinic, Neuropsychology, Marshfield, WI

Background:
A search engine to find physicians' information is a basic but crucial function of a healthcare provider's website. Inefficient search engines returning no results or incorrect results can lead to patient frustration and potential customer loss. A search engine that can that handle the misspelling and spelling variation of names is needed, as the United States has culturally, racially, and ethnically diverse names. This search engine could also have broader implications for the other many types of searches necessary in the medical setting.
Objective:
The Marshfield Clinic web site provides a search engine for users to search for a physicians' name. The current search engine provides auto-completion function but requires exact match. We observed 26% of the searches yielded no results. The goal is to design a fuzzy match algorithm to aid users to find sought physician more effectively and efficiently. A second goal is to then conduct usability testing on both the current web site and fuzzy match search engines to see if differences exist and where.
Methods:
Instead of an exact match search, we used a fuzzy algorithm to find similar matches for searched term. In the algorithm, we solved three types of search engine failures: "Typographic", "Phonetic spelling variation", and "Nickname". To solve these mismatches, we used a customized Levenshtein distance calculation that incorporated Soundex coding and a lookup table of nicknames derived from US census data. Usability testing was conducted in collaboration with local university students as part of their semester independent study project. Staff neuropsychologists were also consulted regarding study design. A pilot usability study was conducted with a convenience sample of undergraduate students, n=10 for the existing web site search engine and n=9 for the fuzzy match search engine (total n=19). After a verbal consent, participants were given a scenario where they had an injury and wanted to consult a physician they had seen in the past. Participants were shown a head shot of a provider with the name, specialty and office phone number listed next to the photo for 10 seconds. An audio pronunciation of the name was also provided. The physician name chosen was considered a relatively harder name to spell. They then viewed a 10 minute video categorized as entertaining and distracting. They were assigned one of the two search engines and asked to find the physician they saw at the beginning of the session and verbally state their appointment phone number. Participants were given five attempts to correctly find the physician. Attempts, rather than time on task were used to compare because the two search engines reacted at different speeds. Audio, screen and participant face recording were captured using Morae software. Participants were entered into a raffle for an iPod shuffle for their time. A system usability scale (SUS) survey was completed after the test session.
Results:
Participants were measured on task success (0=completed successfully, 1=completed with difficulty, 2=failed to complete), recall (did they provide an answer) and precision (did they provide the correct answer). Task success Marshfield Clinic/Fuzzy Match: 0, 30%/78%; 1, 30%/11%; 2, 40%/11%. Recall Marshfield Clinic/Fuzzy Match: 60%/88.89%. Precision Marshfield Clinic/Fuzzy Match: 83.33%/100%. SUS score Marshfield Clinic/Fuzzy Match 78.50/79.25.
Conclusion:
The results of the pilot test suggest that the Fuzzy Match search engine is better at helping people find health care providers than the existing search tool featured on the Marshfield Clinic Web site. An additional study with a larger sample size will provide better information regarding the search recall and accuracy of the Fuzzy Match search engine.

An Evaluation of Computerized Medication Alert Override Behavior in Ambulatory Care

Nivethietha Maniam[1]; Sarah P. Slight, MPharm, PhD, PGDip[2,3]; Diane L. Seger, RPh[1]; Mary Amato, PharmD, MPH[2,7]; Julie M. Fiskio,[1]; Dustin McEvoy[1]; Karen C. Nanji, MD, MPH[4,5]; Patricia C. Dykes, PhD, RN, FACMI[2,4]; David W. Bates MD, MSc.[1,2,4]

[1] Partners Healthcare, Wellesley, Boston, MA, USA; [2] Division of General Internal Medicine, Brigham and Women's Hospital, Boston, MA, USA; [3] School of Medicine, Pharmacy and Health, The University of Durham, Stockton on Tees, Durham, UK; [4] Harvard Medical School, Boston, MA, USA; [5] Department of Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital, Boston, MA, USA; [7] MCPHS University, Boston, MA

Abstract: *Alert fatigue may compromise patient safety during the prescribing process by overburdening providers with unnecessary alerts. Our goal was to identify providers with a high inappropriate override rate and conduct academic detailing sessions regarding their override behavior. This study identified factors that influenced physicians' overriding behavior and identified areas where alert functionality could be improved to create a safer, more efficient computerized decision support system.*

Background: While evidence suggests that computerized decision support (CDS) increases safety and quality of care, understanding how physicians respond to CDS alerts is a critical factor in achieving meaningful use of electronic health records (EHRs). Application of the CDS alert functionality is variable among providers and we continue to observe a high level of medication alert overrides for many prescription domains. While many overrides are justified clinically, some are not, and it is important to be able to reach out to those providers who are not prescribing optimally and understand their reasons for overriding alerts.

Methods: All Level 2 alert overrides that required providers to give a coded reason for overrides at the time of prescribing were downloaded between January 2009 and December 2011 in the outpatient primary care setting. We limited our sample to providers who had received 20 or more alerts (opportunity to override) in each of the prescribing domains (drug-drug interaction, drug-allergy interaction, renal suggestion, age-based, duplicate drug, and formulary substitution alerts) and calculated the number of times each provider overrode these alerts. Of the 725 providers eligible for the study, those with a high inappropriate override rate (above 75% within a specific domain or overall) were targeted for academic detailing sessions. The sessions were conducted by a research pharmacist or physician trained in effective counter-detailing techniques and tailored to each provider's overrides. Graphical materials including performance level data, provider-specific inappropriate overrides and supporting evidence-based summaries were used as the basis for a two-way discussion. A robust analysis of the data was carried out and general views on alert functionality and specific prescribing behavior were identified.

Results: We conducted 23 academic detailing sessions across primary care clinics affiliated with Brigham and Women's Hospital and Massachusetts General Hospital. We identified seven high level content categories: clinical satisfaction, clinical utility and relevance, variant user knowledge, impact on clinician reviewing process, patient preferences, current alerting tool challenges, and considerations for the future. Overall, providers were generally favorable towards the alerts and felt they were helpful in identifying possible adverse interactions. Many providers found that the clinical relevance of the alerts could be improved by including more detailed data such as the magnitude of risk, suggesting alternate treatments, providing recent laboratory values, and up-to-date reference material. Many participants reflected on how they failed to provide a valid override reason because they found the alerts time-consuming and disruptive. Creating a more patient specific alerting tool was recognized as a future area of focus. Limitations of the existing EHR infrastructure such as inaccurate medication and allergy lists were identified as contributing to unnecessary alerts.

Conclusion: Many providers were unaware of their relatively high rate of overriding and this study allowed providers the unique opportunity of objectively assessing their prescribing behavior. Key issues that emerged from the sessions included the perceived risk to physician autonomy in decision-making, the increase of clinically irrelevant alerts leading to alert fatigue, and the lack of supplementary information. By incorporating provider preferences, customizing alerts to the context of the visit, and offering additional clinical data, providers felt that alerts would be less likely to be overridden providing more effective and efficient care.

References:

1. Nanji KC, Slight SP, Seger DL et al. Overrides of Medication-Related Clinical Decision Support Alerts in Outpatients. JAMIA 2013;0:1-5.

This study was funded by grant #U19HS021094 from the Agency for Healthcare Research and Quality (AHRQ)

Analysis of Content Coverage for Informed Consent Concepts

Frank J. Manion, MS[1], Elizabeth R. Eisenhauer, RN, MLS[2], Alla Karnovsky, PhD[3], Yongqun He, PhD[4], Yu Lin, PhD[4], Marcelline R. Harris, PhD, RN[2]

[1]University of Michigan Comprehensive Cancer Center
[2]Division of Systems and Effectiveness Science, School of Nursing
[3]Department of Computational Medicine and Bioinformatics, School of Medicine
[4]Department of Microbiology and Immunology, School of Medicine

Introduction

Knowledge embedded within informed consent for research is of increasing interest in light of efforts to establish large strategic and distributed research networks. Information models and terminology systems can be used to model such knowledge within a domain; well-formed models can inform model driven approaches to developing systems and software that incorporate the model-derived functional and semantic interoperability specifications and ultimately execute on specific implementation platforms. The informed consent domain however lacks comprehensive models of the necessary interactions of information and attributes arising from the informed consent processes and artifacts. Ideally, coherent, canonically-linked formal models of the informed consent for the research domain, incorporating information models and terminology, would allow for improvements in data sharing, secondary use of data, and facilitate IRB review in multiple sharing environments such as consortium and research networks.

Objectives

We report on the first phase of a larger project focused on modeling this domain. In this study we evaluated terms and concepts derived from three different informed consent templates, each representing a different perspective: biomedical clinical research, health and behavior science research, and biorepositories. Specific aims were to 1) identify unique concepts across three IRB templates; 2) map concepts to Concept Unique Identifiers (CUIs) within the Unified Medical Language System (UMLS), and 3) examine content coverage across 2 information models (BRIDG, HL7 RIM) and 4 terminology systems, NCI thesaurus (NCIt), Consumer Health Vocabulary (CHV), Ontology of Biomedical Investigation (OBI), and the University of California San Diego (UCSD) informed consent permission ontology.

Methods

Manual review and annotation of informed consent templates from two University of Michigan IRBs was performed covering 1) biomedical clinical research, 2) health and behavior science research, and 3) the prospective biobanking of DNA. Terms were collected from the 6 terminologies and information models named above. Using the UMLS Terminology Services (UTS) Metathesaurus Browser, each term was entered into the manually reviewed to identify a CUI, terms and definitions consistent with the meaning of the term in the consent template. BioPortal was used to search for OBI and UCSD terms.

Results

324 terms were initially identified from the three templates. Removal of synonyms and duplicates resulted in 296 unique terms; 11 could not be mapped to CUIs. Of the remaining 285 terms, the NCIt provided the most extensive coverage (200 terms), and the UCSD consent permission ontology the least coverage (12 terms).

Discussion & Conclusion

We have identified a set of concepts, CUIs, definitions and preferred terms from key source systems that are needed to more fully model the domain of informed consent. Next steps include an analysis of the binding of the relevant aspects of the information models with the terminology systems, and validation using informed consent templates from other organizations.

Systems Informatics and Information Modeling in Healthcare

Perry L. Mar, PhD[1,2,3], Oliver James[1], Sarah Collins, RN, PhD[1,2,3], Margarita Sordo, PhD[1,2,3], Saverio Maviglia, MD, MS[1,2,3], Li Zhou, MD, PhD[1,2,3], Priyaranjan Tokachichu, MD[1], Hari Nandigam, MD, MSHI[1], Howard Goldberg, MD[1,2], Roberto A. Rocha, MD, PhD[1,2,3]

[1]Clinical Informatics, Partners HealthCare System, Boston, MA; [2]Brigham and Women's Hospital, Boston, MA; [3]Harvard Medical School, Boston, MA

Abstract

Due to the importance of systems thinking in healthcare and the usefulness of information modeling in clinical informatics, best practices of systems design and management (SDM) and its application to information modeling were compiled and considered. A resulting set of best practices was arrived at for a systems approach to informatics and information modeling.

Introduction

Systems design and management (SDM) has been increasingly recognized as an important capability in healthcare[1-5], which can be regarded as a system of many intricately interacting parts. In this work, best practices of SDM and their application to information modeling were considered by the Information Modeling Center of Excellence at Partners HealthCare System[6]. A resulting set of high-level best practices was arrived at for a systems approach to clinical informatics and information modeling.

Methods

As input knowledge for consideration, several key references on SDM were reviewed[1-5]. The input of members with systems thinking or engineering background was also obtained and reviewed. The best practices were extracted and organized to arrive at recommendations for systems informatics methods.

Results

Many widely known best practices were confirmed, including well-known methods for information system design and implementation. Other important, less frequently discussed practices included iterative problem definition, adopting rigorous engineering methodology, dependency management, challenging previous assumptions, and global optimization. For example, relaxing an assumption of only simple relationships in a patient's family history results in a model that was structured to allow for a more complete and rigorous pedigree for better decision support.

Conclusion

In order to address the increasing complexity of informatics needs in healthcare, a systems approach provides access to solutions of greater comprehensiveness and optimality.

References

1. Plsek PE, Greenhalgh T. Complexity science: the challenge of complexity in health care. BMJ. 2001 Sep 15;323(7313):625-8.
2. Wilson T, Holt T, Greenhalgh T. Complexity science: complexity and clinical care. BMJ. 2001 Sep 22;323(7314):685-8.
3. Plsek PE, Wilson T. Complexity science: complexity, leadership, and management in healthcare organizations. BMJ. 2001 Sep 29;323(7315):746-9.
4. System design and management [Internet]. Cambridge, MA: System Design and Management Program, Massachusetts Institute of Technology; c2010. MIT SDM: Systems Thinking Conference 2010: SDM Systems Thinking Conference 2010 Presentations and Videos. 2010 [cited 2014 Mar 3]. Available from: http://sdm.mit.edu/systems_thinking_conference_2010/presentations.html
5. Grossmann C, Goolsby WA, Olsen L, McGinnis JM. Engineering a learning healthcare system: a look at the future: workshop summary. Washington, DC: National Academy of Sciences; 2011. 313 p.
6. Mar PL, Rocha RA, Goldberg HS, Einbinder JS, Middleton B. An information modeling center of excellence. In: AMIA Annu Symp Proc; 2010 Nov 13-17; Washington, DC. p. 1166.

Comparing Accuracy, Efficiency, and User satisfaction of Two EMR interfaces

David T. Marc, MS[1], Charat Thongprayoon, MD[2], Andrew Harrison, BS[2], John C. O'Horo, MD, MPH[2], Ronaldo A. Sevilla Berrios, MD[2], K. Harder, PhD[3], Brian Pickering[2], MD, Vitaly Herasevich, MD, PhD[2]

[1]Institute for Health Informatics, University of Minnesota, Minneapolis, MN; [2]Multidisciplinary Epidemiology and Translational Research in Intensive Care (METRIC), Mayo Clinic, Rochester, MN; [3]Center for Design in Health, University of Minnesota, Minneapolis, MN

Abstract The purpose of this study is to compare the accuracy, efficiency, and cognitive load of completing various clinical tasks in a critical care setting and overall user satisfaction between two patient information displays: AWARE (Ambient Warning and Response Evaluation) and a traditional EMR. When compared to a traditional EMR display, the AWARE interface lead to lower cognitive task load and higher efficiency of task completion.

Introduction EMR platforms continue to proliferate, often with insufficient attention to aspects important to ensure that meaningful information is conveyed to users. A novel patient information display, AWARE, was designed to provide an easy-to-use and easy-to-learn electronic patient information display that presents meaningful information to clinicians to improve their decision-making and reduce errors. AWARE extracts and combines textual and graphical elements to convey specific clinical patient information on one screen in contrast to multiple pages in a traditional EMR. The goal of this study is to determine if the AWARE interface leads to greater accuracy and efficiency of completing various clinical tasks when compared to a traditional EMR.

Methods A randomized crossover design was used to assess the performance of 5 research fellows (all of whom trained outside the US) without extensive experience using the AWARE patient information display or the traditional EMR display. The subjects completed global and localized tasks. Global tasks required subjects to answer questions pertaining to all patients on a specific patient unit and were completed once per electronic display (AWARE or traditional EMR). Localized tasks required subjects to answer clinical questions about three randomly selected patient cases for each display. The National Aeronautics and Space Association Task Load Index (NASA-TLX) was used to measure task workload on a 1-100 dimensionless scale, where higher numbers correspond to greater cognitive demand. An electronic timer was used to measure the time subjects needed to perform each task. Task accuracy was determined based on the correctness of the subject's responses to each task. At the conclusion of the experiment, each subject completed a user satisfaction survey.

Results When comparing the results of the NASA-TLX for the two systems, the traditional EMR (mean: 72.1, SEM: 6.8) had significantly higher weighted scores than the AWARE interface (mean: 27.8, SEM: 6.2, p=0.001). The AWARE interface took significantly less time to complete tasks when compared to the traditional EMR for both the global tasks (mean: 65 sec vs 0.257 sec, p<0.05) and localized tasks (mean: 37 sec. vs 59 sec., p<0.05). There was not a difference in the accuracy of the clinical tasks completed between the two interfaces. Additionally, subjects reported that AWARE was easier to use, lead to greater productivity, and was an overall improvement over the functionality of the traditional EMR display.

Conclusions The AWARE patient information display led to lower cognitive load and greater efficiency when completing various clinical tasks and improved user satisfaction than a traditional EMR in an experimental setting.

A New Visual Navigation System for Exploring Biomedical Patents

Christopher Markson, MS, Songhua Xu, PhD
Department of Information Systems, College of Computing Sciences,
New Jersey Institute of Technology, Newark, NJ 07102, USA

Abstract

This research develops a new visual navigation system for exploring key technologies and trends along with their temporally evolving relationships exhibited by biomedical patent literature. The proposed system is implemented as a web-based graphical browsing environment (http://web.njit.edu/~crm23/AMIA) through which users can examine key technologies together with their dynamics and inter-relationships as revealed by the patent literature.

Introduction

US Patent and Trademark Office (USPTO) biomedical patents are valuable intellectual assets of the community. In particular, the *Claims* section of a patent document provides rich information disclosing technological advancement[1]. We therefore focus on this central element to algorithmically discover and graphically present valuable technologies and their trends in biomedicine.

Our Method: Data Collection, Manipulation, and Visualization

Four years of Google Bulk Patent[2] data are used in this work to allow for a temporal analysis of biomedical patents. The aim is to computationally detect and visually reveal the dynamic relationships between technological concepts and the technological advancements relevant to a user's interest. For each patent document, its claims section is parsed for *keywords* and *salience weights* using the RAKE algorithm[3]. In total, 88,493 patents are processed and 2,501,012 keywords are extracted. To identify frequent co-occurrences of keywords among all patents, the Apriori algorithm is used to extract frequent keyword item sets[4]. For each graph representation, nodes represent frequent keywords occurring in the claim data. Edges represent a collection of patents that share certain relationships between their keywords. The Sigma.js platform was chosen as the underlying visualization platform due to its scalability[5]. We also modify the platform to enable the removal of keywords from the displayed graph to adjust the amount of information presented.

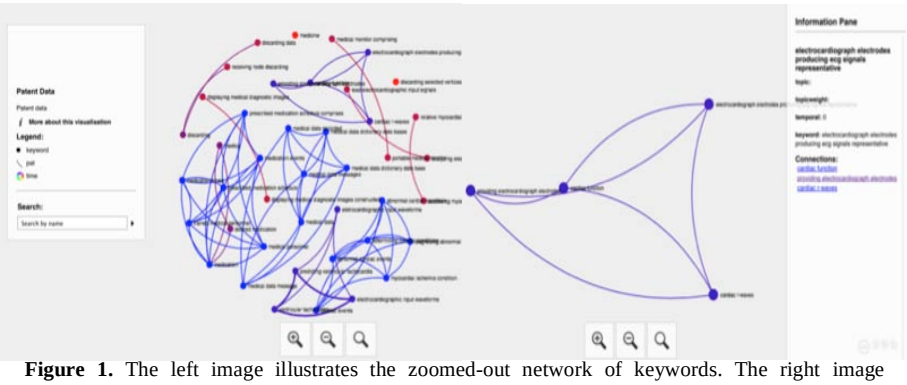


Figure 1. The left image illustrates the zoomed-out network of keywords. The right image illustrates a close-up on a particular set of connected keywords.

Conclusion

This work presents a new visual navigation system built on an open-source platform for identifying key technological concepts in biomedical patent claims as well as these concepts' dynamic relationships and development trends. The visualization platform introduced in this study renders the results of patent claims analysis through visually informing graphs.

References

1. Tong X, Frame JD. Measuring national technological performance with patent claims data. Research Policy. 1994;23(2):133-41.
2. Google. USPTO Bulk Downloads: Patent Grant Full Text: Google; 2013, 2012, 2010, 2005. Available from: http://www.google.com/googlebooks/uspto-patents-grants-text.html.
3. Rose S, Engel D, Cramer N, Cowley W. Automatic keyword extraction from individual documents. Text Mining. 2010:1-20.
4. Agrawal R, Imieliński T, Swami A, editors. Mining association rules between sets of items in large databases. ACM SIGMOD Record; 1993: ACM.
5. Bastian M. HS, Jacomy M. Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media. 2009.

From computer scientists to research practitioners: lessons learned when implementing use of mobile devices in a cancer research setting

Stephanie L. Martch, MS, RD[1], Karen Basen-Engquist, PhD, MPH[1], Wendy Demark-Wahnefried, PhD, RD[2], Alexander V. Prokhorov, MD, PhD[1], Kevin Patrick, MD, MS[3], Eileen H. Shinn, PhD[1], Emilia Farcas, PhD[3], Chaitan Baru, PhD[3], Ingolf Krueger, PhD[3], Kai Lin, PhD[3], Phil Rios[3], Yan Yan[3],Viswanath Nandigam[3], Susan K. Peterson, PhD, MPH[1]

[1]MD Anderson Cancer Center, Houston, TX; [2]University of Alabama at Birmingham, Birmingham, Alabama; [3]The University of California, San Diego, La Jolla, CA

Mobile health and internet-based technologies can provide effective tools for managing the healthcare of cancer patients, although a dearth of research exists to provide evidence of clinical impact.[1] CYCORE (CYberinfrastructure for COmparative effectiveness REsearch) is a software-based prototype of a user-friendly CyberInfrastructure (CI) designed to support collection and analyses of cancer prevention and treatment data from multiple domains using home-based mobile sensors. A collaborative effort between CYCORE's computer scientists and engineers at the University of California, San Diego (UCSD), and behavioral scientists from two cancer centers (MD Anderson [MDACC], Houston, Texas; and University of Alabama at Birmingham) enabled requirements gathering from over 100 cancer center stakeholders. This work informed the prototype design to address four real-world use cases that assessed (1) physical functioning in colorectal cancer patients, (2) dehydration risk in head and neck cancer (HNC) patients undergoing radiation therapy (RT), (3) adherence to swallowing exercises in HNC patients undergoing RT, and (4) tobacco use in cancer survivors who completed an in-house tobacco treatment program. Devices for the four use cases were, respectively: (1) a heart rate monitor, two accelerometers, and a blood pressure (BP) and global positioning system (GPS) device; (2) a BP device, two accelerometers, and a weight scale; (3) a smartphone-based video application (app); and (4) a carbon monoxide monitor and the smartphone-based video app. All device suites included a miniature computer base station for home use to handle both wireless and wired data acquisition and CI transmission, and a smartphone that featured in-house-developed apps to record/transmit self-reported data and applicable videos. The UCSD group encoded devices and modes for data collection, transmission and storage; as well as CI integration and display. The research dissemination group at MDACC developed protocols for conducting four use-case-related feasibility trials that included participant user guides designed to minimize concerns about device functionality[2]; staff guides to set up, maintain, and inventory devices and to track and troubleshoot device problems; and protocols for addressing mobile device service provider issues that arose. This poster will present several lessons learned by the research dissemination group during incorporation of mobile devices into a cancer research project including (1) the need for research staff to become site technology experts by delving into the manufacturers' usage manuals to create staff troubleshooting guides as well as simple need-to-know participant device usage booklets, (2) how to "tame" the smartphone, so that its extensive features do not interfere with research goals, and (3) considerations for the simultaneous use of multiple devices that collect continuous data.

References

1. eHealth Initiative. A study and report on the use of eHealth tools for chronic disease care among socially disadvantaged populations: issue brief on eHealth tools for cancer patients. California HealthCare Foundation, December 24, 2012. Grant #16920.
2. Byrne D, Kelly L, Jones GJF. Multiple multimodal mobile devices: lessons learned from engineering lifelog solutions. In: Alencar P, Cowan D, editors. Handbook of research on mobile software engineering: design implementation and emergent applications. Hershey, PA: Engineering Science Reference; 2012. p. 706-724.

Point of Care Intake Tool and Clinical Decision Support to Achieve Meaningful Use

Maryanne E. Mathiowetz, M.P.H., Jennifer L. Horn, M.D., Samanthie I. Epps, M.B.A., Paul J. Johnsen, B.S., Pedro J. Caraballo, M.D.
Mayo Clinic, Rochester, Minnesota

Abstract
Capture of discrete data during routine clinical practice is critical for the Meaningful Use of the electronic health record (EHR). We developed a smart form integrated in the EHR. After implementation, structured data capture for specific MU measures improved from 0% to >95% without affecting workflow efficiency. We concluded that standard EHR functionality can be enhanced to support an efficient workflow and improve standardization and reporting.

Introduction
The Meaningful Use (MU) program provides incentive payments to ambulatory care facilities based on specific criteria. Capture of discrete data is a critical MU requirement. Most of these data are captured during routine clinical practice at our institution, but the existing documentation was poorly standardized, available only in a textual format, outside of the certified EHR, or absent. Our aim was to use certified EHR functionality to develop a smart form for intake documentation that would comply with MU without negatively impacting clinical practice.

Methods
Our institution has GE Centricity Enterprise as the main component of the EHR. We used standard functionality including FlowForm, rule engine and scripting to customize an intake smart form to be used by clinical assistants and nurses to gather and document information needed for the outpatient visit. The main requirements for the smart form and the functionality we used to achieve those requirements were: 1) Support intake standardization (best practice) with synchronous decision support to prompt documentation at the appropriate intervals and visit types. 2) Maintain efficiency of the intake process while changing what is documented and how by a) eliminating redundant documentation, with data sharing across EHR components, b) providing visual cues (color, icons, etc.) to focus documentation where needed and ensure completeness, c) incorporating pick lists, check boxes, radio buttons, and automated calculations for ease of charting in discrete fields, and d) facilitating navigation within and between the EHR and other electronic tools including direct access to printable resource materials for the patient. 3) Connect the smart form and the provider's note to transfer the discrete intake data to the designated note sections in the appropriate format for the provider. 4) The development strategy also required that the smart form be scalable to support future phases of MU as well as other external reporting requirements and practice changes.

Results
The intake smart form was implemented in six primary care areas with approximately 300 clinical assistants and nurses and is used for approximately 3500 to 4000 visits per week. There were three MU categories for which a complete set of elemental data was not available for reporting from the certified EMR in the baseline workflow (0%): Adult Weight Screening and Follow-up, Tobacco Use and Cessation Intervention, Weight Assessment and Nutrition/Exercise Follow-up for Children and Adolescents). Following the implementation, our institution's MU Reporting Team found that the compliance rate for the measures in these categories did increase to above 95%. No difference was found in the duration of the intake process when comparing a pre-implementation baseline time period with multiple post-implementation time periods. Following the original implementation, additions and enhancements have been made in support of new external reporting requirements and practice changes.

Conclusion
Creative documentation process redesign based on existing functionality of a certified EMR can produce new tools that support MU and other reporting requirements, as well as help to standardize best practice and support efficient workflow. Particularly notable with this implementation was the lack of impact on the providers in the clinical practice, as demonstrated by no increase in pre-visit intake time and no decrease in the availability of intake documentation in the provider's note.

Using REDCap to Evaluate Clinical Decision Support Alert Appropriateness

Allison B. McCoy, PhD[1,2], Eric J. Thomas, MD, MPH[3], Marie Krousel-Wood, MD, MSPH[1,2], Dean F. Sittig, PhD[3]
[1]Tulane University, New Orleans, LA; [2]Ochsner Health System, New Orleans, LA; [3]The University of Texas Health Science Center at Houston, Houston, TX

Abstract

Effectively evaluating the appropriateness of clinical decision support (CDS) alerts and responses is critical to improving patient safety through health information technology. We describe the use of REDCap to create a data collection instrument for evaluating the appropriateness of CDS alerts and responses. While the tool is effective, further enhancements to REDCap would improve the process for future evaluations.

Introduction

Computerized alerts that warn clinicians about drug interactions or provide dosing guidance are commonly implemented forms of clinical decision support (CDS) to improve patient safety. Alert overrides can hinder provider effectiveness and result in adverse patient outcomes. Detailed evaluation of the appropriateness of alerts and clinician responses is necessary; however, few have described the processes for performing these evaluations.

Methods

Based on a previously described evaluation framework,[1] we developed an assessment tool using REDCap (Research Electronic Data Capture), a secure, web-based application designed to support data capture for research studies.[2] The tool was designed to facilitate reviews by clinicians in determining the appropriateness of CDS alerts and responses.

Results

We created a single data collection instrument within REDCap for reviews of CDS alert and response appropriateness. The instrument consists of data fields in two sections: alert log data to assist clinicians during reviews and questions for the clinicians to answer while conducting the reviews. We extracted log data from the electronic health record and imported it into REDCap to populate these fields prior to clinician reviews. For the second section, clinician reviewers indicated whether the alert and response were appropriate, why the alert was inappropriate (if applicable), and what response(s) would have been appropriate. Each alert was reviewed by multiple clinicians to allow assessment of inter-rater agreement. Use of REDCap was advantageous in that it allowed quick, easy development of a HIPAA-compliant data collection tool that is simple for clinicians to use. However, although REDCap allowed importing of data and multiple reviews of a single record (i.e., dual data entry), the two could not be used together. As such, we created data access groups for each reviewer, assigned each alert a separate, unique identifier for each reviewer, then assessed agreement and merged responses outside of the REDCap system.

Conclusion

We successfully used REDCap to create a data collection instrument to facilitate evaluation of CDS alert and response appropriateness through chart review by clinicians. Further enhancements to REDCap, such as the ability to import data with dual data entry, would improve the process for future evaluations.

References

1. McCoy AB, Waitman LR, Lewis JB, Wright JA, Choma DP, Miller RA, et al. A framework for evaluating the appropriateness of clinical decision support alerts and responses. J Am Med Inform Assoc. 2012 Jun;19(3):346–52.
2. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research Electronic Data Capture (REDCap) - A metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. 2009 Apr;42(2):377–81.

Acknowledgments: This work was supported by NLM Grant 1K22LM011430-01A1, a UTHealth Young Clinical and Translational Sciences Investigator Award (KL2 TR 000370-06A1), and NCRR Grant 3UL1RR024148.

Identifying Strategies to Promote Adoption of a Web-based Patient-Centered Communication Tool by Providers in the Acute Care Setting

Kelly McNally, [1] Diana Stade, [1] Patricia C. Dykes, RN, PhD,[1,2] David W. Bates, MSc,[1,2] and Anuj K Dalal, MD[1]

[1]Brigham and Women's Hospital, Boston, MA, [2]Harvard Medical School, Boston, MA

Abstract: *Web-based tools may improve patient-centered communication but require provider acceptance to realize its full potential. We designed a web-based communication tool to manage electronic conversations among patients and providers in the acute care setting. We conducted focus groups to explore potential uses, perceived impact on workflow, and provider concerns regarding the tool. We identified potential barriers and developed strategies to optimize adoption of the web-based communication tool.*

Introduction: Promoting patient-centered communication may favorably impact patient safety, quality of care, outcomes, and healthcare costs.[1] Web-based communication tools such as microblogs have the potential of improving patient-centered communication among patients and providers, but have not been systematically adopted in complex hospital settings. Although patients are eager to use such tools to communicate with providers, providers have expressed resistance in using them.[2] The purpose of this study is to identify barriers to implementing a web-based communication tool and establish strategies to maximize adoption by providers in the acute care setting.

Method: We conducted 3 focus groups, consisting of 4 to 6 providers (attending physicians, physician assistants, and housestaff) at Brigham and Women's Hospital. During each session we presented the intended use of and discussed provider concerns with regard to effectively using the web-based communication tool. We debriefed following each focus group to analyze and discuss the implications of our findings. We then identified optimal strategies for implementation and training.

Results: We identified two types of barriers that could impede adoption. *Clinical workflow barriers* consisted of ensuring easy access to the web-based communication tool (which varied by provider type), and integrating the tool into clinicians' demanding inpatient work schedules. *Provider "emotional" barriers* consisted of concerns regarding notifications of new messages, fear of providing poor quality patient care (e.g., by virtue of spending less time with patients at the bedside), and malpractice concerns pertaining to communication between patients and providers.

Conclusions: We developed strategies to promote adoption of our web-based communication tool based on the clinical workflow and emotional barriers identified in focus groups. Strategies to address *workflow barriers* include identifying key clinical applications used by different providers, ensuring reliable use of those applications, providing links to the web-based communication tool from those applications, and establishing a specific time within current work schedules to review electronic conversations. Strategies to manage *emotional barriers* are as follows*:* providing several user-configurable options for message notification (e.g., email vs. mobile "app" push-notification) to mitigate perception of alert fatigue; framing use of the tool as a means to supplement patient care activities rather than replace existing practices to address patient care quality concerns; and finally, educating providers regarding appropriate messaging etiquette when communicating with patients and other providers and indicating that hospital administration supports use of the tool to mitigate legal concerns related to posting messages on the web-based communication tool.

Acknowledgements: The Brigham and Women's Hospital PROSPECT project is part of the Libretto Consortium supported by the Gordon and Betty Moore Foundation

References

1. Stewart M, Brown JB, Donner A, et al. *The impact of patient-centered care on outcomes*. J Fam Pract. 2000; 49:796–804.
2. Hassol A[1], Walker JM, Kidder D, Rokita K, Young D, Pierdon S, Deitz D, Kuck S, Ortiz E. *Patient experiences and attitudes about access to a patient electronic health care record and linked web messaging. Next-generation phenotyping of electronic health records.* J Am Med Inform Assoc. 2004.11(6):505-13.

Introduction
The American College of Emergency Physicians recommends the use of Wells' Score for Pulmonary Embolism (PE) and D-dimer as objective risk-stratification of patients with suspected PE prior to ordering computed tomography pulmonary angiogram (CTPA). Implementation of these recommendations by use of algorithmic clinical decision-support (CDS) incorporated into computerized provider order entry (CPOE) has been demonstrated with mixed results. Our objective was to evaluate the acceptability of such CDS at our institution, with the hypothesis this intervention would increase positive CTPA yield for PE.

Methods:
This study was performed at a Level 1 academic emergency department with an enterprise-wide electronic health record. A CDS algorithm for guideline-compliant diagnosis of PE using Wells' Score and D-dimer was integrated into CPOE. A baseline CTPA positivity rate of 8.0% for acute PE was determined from the most recent 250 studies prior to the intervention going live. The evaluation period following CDS implementation was from March 1, 2014 until July 10, 2014. Upon placing a CTPA order, the CDS flow chart is reflected in Figure 1. Alert #1 prompted clinicians to complete a checklist for objective risk-stratification by Wells' Score. Alert #2 prompted clinicians with recommendations to discontinue the CTPA order if the patient were low- or intermediate-risk, and the D-dimer result was <0.4 μg/mL. Data were collected retrospectively on all patients who had a CTPA ordered and charts were abstracted for the presence or absence of D-dimer results, Wells' score documentation, and CTPA result.

Results:
CTPA were ordered on 262 patients during the evaluation period. The flow of these patients through the CDS is summarized in Figure 1. 49 patients had an elevated D-dimer recorded and no prompts were displayed, 3 (6.1%) of whom were diagnosed with acute PE. Alert #1 was shown 213 instances, but acceptability was poor, with only 61 risk-stratification checklists completed. Overall, guideline-compliant care was performed in 69 patients, resulting in 7 (10.1%) positive CTPA. However, the overall positive CTPA yield of our institution's entire CTPA cohort also increased to 13.8% during the observation period.

Discussion
The CDS tool had poor acceptance at our institution, as evidenced by the disproportionate number of CTPA completed without objective risk-stratification. Due to the limited sample of completed CDS checklists, no valid conclusions may be drawn regarding the effectiveness of the CDS. Causative factors for the increased yield of CTPA during the observation period are unclear. A Hawthorne effect associated with CDS implementation cannot be excluded.

Assessment of the Quality of Computerized Physician Order Entry (CPOE)-Related Medication Error Reports in a Large Medication Error Database

Amato MG, PharmD, MPH[1,2]; Seger AC, PharmD[1,]; Wright A, PhD[1,3]; Koppel R, PhD[4]; Rashidee AH, MBBS, MS[5]; Elson RB, MD[6], MS; Whitney, DL, BS[7]; Thach TT, MPH[1]; Bates DW, MD, MSc[1,3,8]; Schiff GD, MD[1,3]

[1] Brigham and Women's Hospital Division of General Medicine and Primary Care, Boston, MA, USA, [2] MCPHS University, Boston, MA, USA, [3] Harvard School of Medicine, Boston, MA, USA, [4] University of Pennsylvania, Philadelphia, PA, USA, [5] Quantros, Inc., CA, USA, [6] MetroHealth Center for HealthCare Research and Policy, Cleveland, OH, USA, [7] Baylor College of Medicine, Houston, TX, USA[8] Harvard School of Public Health, Boston, MA, USA

Abstract: *Improved awareness and monitoring of medication errors related to CPOE can help institutions implement effective prevention strategies. The objective of this study was to assess the quality of narrative reports of CPOE-related errors in the MEDMARX database and identify qualities of useful error reports.*

Background: Reports of adverse drug events and medication errors are more useful if they include well-written narrative reports in addition to structured data.[1, 2] Monitoring of medication errors related to CPOE can help institutions identify potential causes and implement adequate error prevention strategies to improve patient safety.

Methods: We assessed quality of error report narrative descriptions included in the MEDMARX reporting system from 2003-2010 that were listed as being related to CPOE. MEDMARX is a national reporting program developed by the United States Pharmacopeia (now administered by Quantros) containing >1.2 million reports from over 860 facilities. For each error narrative, we assessed "what" happened, "why", and how the error might be prevented, and ranked overall quality of the narrative report (1 = no information, 2 = poor information, 3 = some information, 4 = adequately described, and 5 = error well described). We also ranked how easily the cause of the error could be determined from the report (1 = no information, 2 = poor information, 3 = some information, 4 = adequately described, and 5 = cause well described), assessed whether the report described the specific involvement of CPOE, whether the narrative report was consistent with structured data included in the report and whether the narrative report improved overall understanding of the error when added to structured data.

Results: Of 63,040 error reports classified as related to CPOE during 2003-2010, we reviewed a stratified random sample of 10,060 (16%). Quality of the error reports was ranked as well-described in 1336 (13.3%) and adequate in 3312(32.9%). 2871(28.5%) had some information, 2265 (22.5%) were poor and 274(2.7%) had no information (example: only stating "CPOE error"). Cause of the error was rated adequate or well described in 20%, however 51.6% included no information to assess error cause. CPOE involvement was well described in 13.6%, mentioned with some details in 20.4%, lacking details in 13.2%, and implied or not mentioned in 49.8%. In a random sample of 100 reports coded by two investigators, inter-rater reliability using linear weighted kappa was good for quality of the error report (kappa 0.604), cause of the error (kappa 0.692), and CPOE involvement (kappa 0.606). A temporal trend was seen with an initial improvement in quality scores after the first two years, which leveled off for the remaining five years studied. Narrative reports were somewhat or clearly consistent with structured data in 89.1%. The narrative improved overall understanding of the error in 97%. The highest rated quality reports included clear descriptions of what happened and suggested potential error causes and contributing factors.

Conclusion: Narrative reports improved understanding of what happened in this large database of CPOE-related medication error reports, however over half of the reports did not include enough detail to assess causes and suggest prevention strategies. It is important for healthcare systems to focus on improving the quality of medication error reports in order to learn more about causes and implement health information technology strategies to prevent subsequent errors.

References

1. Brajovic S, Piazza-Hepp T, Swartz L, Pan GD. Quality assessment of spontaneous triggered adverse event reports received by the Food and Drug Administration. Pharmacoepidemiol Drug Saf. 2012;21:565-70.
2. Schiff GD, Seger AC, Amato M, Whitney DL, Boehne J, Rashidee A, Elson RB, Koppel R, Wright A, Bates DW. CPOE-related medication errors: Analysis of 10,000 error report narratives and vulnerability testing of current systems. Journal of General Internal Medicine 2012;27(Suppl 2):S136.

Application of Pediatric Dosing Rules Across a Diverse Health Care System

Paul E. Milligan PharmD[a], Nicholas B. Hampton PharmD[a], Chandana Anigolu BE[a], Kevin Heard BS[a], Melissa Heigham PharmD[a], Paul Hmiel MD[b], Keith F. Woeltje MD, PhD[a,b]
[a]BJC HealthCare, Saint Louis, MO [b]Washington University School of Medicine, St. Louis, MO

Abstract
Over time, specific pediatric dosing rules have been developed, tested, refined and applied to all patients at St. Louis Children's Hospital (SLCH), the academic pediatric hospital of the BJC HealthCare System. However, 20% of all pediatric inpatient hospital visits to BJC HealthCare were to our community hospitals in 2012. Pediatric dosing rules have not been applied consistently outside SLCH, primarily due to diverse formularies and different clinical information systems. In 2013, we built pediatric dosing rules for drugs judged to have the highest potential for adverse events based on historical data and an expert panel. These rules will alert pharmacists system-wide, as well as provide evidence-based instructions on how to communicate the proper dose to the physician.

Background
BJC HealthCare is a health system based in St. Louis Missouri comprised of one pediatric, two academic and eight community hospitals. Twenty percent of BJC HealthCare's 2012 pediatric hospital admissions (excluding standard births) were at the community hospitals. St. Louis Children's Hospital, the academic pediatric hospital, developed, tested, and refined specific pediatric dosing rules. Until recently, these rules have not been applied consistently across our system. The primary reasons for this inconsistency are the diverse formularies and different clinical information systems. Therefore, a team set out to develop and apply a starter set of pediatric dosing rules for drugs judged to have the highest potential for adverse events to children at all of our hospitals.

Methods
The BJC HealthCare Clinical Decision Support (CDS) Collaborative, a system-wide CDS steering committee, commissioned an effort to implement a starter set of pediatric dosing rules at all hospitals. To initiate this effort, system-wide inpatient pediatric medication order data were collected from the Pharmacy Expert System (PES), a BJC HealthCare developed enterprise CDS system[1]. The top 50 prescribed drugs were then reviewed by a SLCH physician and 3 pharmacists to identify medications with the greatest potential for adverse events. St. Louis Children's Hospital previously went through a similar exercise to target medications, and developed, tested, and refined dosing rules in their custom Dose Range Auditing and Checking Overseer (DRACO) application[2].
After identifying a set of 19 high-risk drugs, dosing rules were developed in PES for these drugs to act as a safety-net for the entire system. PES continually screens the electronic medical record and when rules are violated, an alert is displayed on a webpage for the local clinical pharmacist to review. As these rules will be triggered rarely, often to non-pediatric specialty pharmacists, the alert will include evidence-based instructions on how to communicate the proper dose to the physician.

Results
The high-risk drugs identified are listed below: ampicillin, acetaminophen, caffeine, cefazolin, cefotaxime, ceftriaxone, diphenhydramine, fentanyl, gentamicin, heparin, hydrocodone with acetaminophen, ibuprofen, midazolam, morphine, potassium chloride, ondansetron, oxycodone, oxycodone with acetaminophen, and phenobarbital.

Discussion and Conclusion
We identified an at-risk population of patients, then systematically identified, developed, and applied a set of CDS rules to protect them. We then gave the pharmacists tools to help communicate dosing rules to the prescriber. Until real-time applied dosing rules upon order entry become more sophisticated, asynchronous CDS dosing rules are an effective safety net to protect patients consistently from adverse events.

Author's Note: Alerts per screened order data is not currently available, but we plan to present this data on the poster.

References

1. Reichley RM, Seaton TL, Resetar E, et al. Implementing a commercial rule base as a medication order safety net. J Am Med Inform Assoc 2005;12(4):383–389.
2. Andrus, CH, Yu F, Implementing a Hierarchical Pediatric Medication Dose-Range Decision Support System. HIMMS Conference 2013.

Identifying eHealth literacy demands of health information seeking tasks

Jelena Mirkovic, PhD[1], Maria Sims, BS[2], David R. Kaufman, PhD[2]
[1]Oslo University Hospital, Oslo, Norway; [2] Arizona State University, Scottsdale, AZ

Abstract

To be able to effectively engage eHealth systems the individuals must possess a set of skills and knowledge referred to also as eHealth literacy. The presented work analyzes participants' barriers while performing health information seeking tasks on MedlinePlus. The goal is to classify task demands in terms of different literacy types and diagnose cognitive operations at varying levels of complexity. The objective is to contribute to design strategies that serve to reduce complexity for fuller consumer participation.

Introduction

eHealth systems play an increasingly important role in engaging patients as productive participants in health management and decision making process. eHealth literacy refers to a set of skills and knowledge that are required for an individual to effectively use eHealth systems. Chan and Kaufman developed a methodological framework to classify tasks demands and user performance according to different eHealth literacy types and cognitive complexity levels (e.g., analysis, evaluating, comprehending)[1]. The goal of this study is to identify and classify potential barriers in health information seeking tasks with greater precision along the dimensions specified by the framework. The ultimate objective is to provide resources to individuals of lower literacy on problems of greater complexity.

Method

Fifteen participants were recruited from a community center associated with Columbia University to perform a set of information retrieval tasks in view to solve hypothetical health problems using MedlinePlus. Participants ranged in age between 30 and 65, with most participants in their 50s and of varying levels of literacy. In this paper, we report the results of a study in which users are tasked to find information related: (1) to management of high blood pressure, and (2) causes, signs, treatment of ADHD. The participants were video-recorded and analyzed using Morae™ usability and video-analytic software. This project was approved by the Columbia University IRB.

Results

Each task presented unique configurations of complexity that presented challenges for users as measured by errors, requests for help and the need for experimenter prompts. Barriers observed during task completion were noted and the framework coding was then applied to perform classification based on literacy type and complexity level (Table 1). The analysis showed that the majority of encountered barriers were associated with participants' lack of information literacy and health literacy. Participants struggled with recognizing and understanding needed information and evaluating and applying the health-related information. Additionally, a substantial number of barriers were classified as computer literacy and numeracy issues.

Table 1. Number of barriers participant encountered mapped to literacy types

	Computer literacy	Information literacy	Traditional literacy	Health literacy
Task 1	17	30	19	15
Task 2	8	29	7	21
Both tasks	**25**	**59**	**26**	**36**

Conclusion

The study revealed barriers that health consumers encounter when using online eHealth tools according to type of eHealth literacy and the specific information-seeking demands of the task. The results can be used to inform improvement and development of eHealth tools for health consumers that vary in terms of their eHealth literacy.

Acknowledgments

This work is supported by a grant from the National Library of Medicine (1R21LM01068801) awarded to David Kaufman. We would like to thank subjects for their participation in this study.

References

1. Chan CV, Kaufman DR. A Framework for Characterizing eHealth Literacy Demands and Barriers. J Med Internet Res; 2011;13(4):e94.

Title: Methods for Patient Matching in Patient-Centered Outcomes Research

Authors: Adil Moiduddin, Prashila Dullabh, Michael Latterner, Samantha Zenlea, Michael Davern

Theme: Clinical Research Informatics

Introduction and Purpose. This paper describes opportunities for and barriers to linking patient-level data across different health data sources, a process called "patient matching." Effective patient matching creates a more complete patient record, which can improve individual-level clinical care and care coordination. It can also enhance patient-centered outcomes research (PCOR), including public health and quality-based initiatives, by uniting disparate data sources (e.g., claims, pharmacy, clinical, and disease registry data) for analysis. In assessing the necessary data infrastructure to support patient matching, we sought to answer the following questions: 1) What types of datasets need to be linked to enable PCOR? 2) What is the general process used for linking datasets and what challenges arise? 3) How could existing processes and technologies be improved? ; 4) What innovations or strategies would help achieve these improvements?

Methods. This qualitative study included; 1) an environmental scan and literature review to identify key resources; and 2) Semi-structured telephone interviews with 9 experts with relevant experience with patient matching, such as expertise in creating standards and strategies for matching individual records from disparate data sources; linking data to research the care continuum and care coordination; or background in public health.

Principal Findings

We identified the processes behind patient matching and the associated technical considerations, such as matching methods and data governance considerations, as well as the challenges faced by stakeholders in accessing the data needed for effective matching. Experts noted challenges stemming from:

- Need for consensus on core patient attributes to capture for accurate matching;
- Need for adequate standards for data quality and methods for understanding the implications of variable data quality;
- Significant error rates associated with current methods of matching;
- Pros/cons of algorithms employed for data matching;
- Restrictions on use of patient attributes in data for research;
- Resource requirements to conduct patient matching; and
- A business culture that does not facilitate open sharing of data-matching strategies across organizations

Conclusions. Large scale patient matching among disparate data sources requires advanced technical and statistical solutions, as well as the management of complex data governance issues, including patient safety and privacy and informed consent. In order to realize the potential of data matching to support PCOR, these complex issues require leadership. Here, we describe areas in which ongoing or new initiatives can help address important obstacles to patient matching through improvements in data standards, policies, and infrastructure.

A study of synonym extraction from clinical texts using semantic vector models

Sungrim Moon, Ph.D.[1], Trevor Cohen, MBChB, Ph.D.[1,2], Hua Xu Ph.D.[1]
[1]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA; [2]National Center for Cognitive Informatics and Decision Making in Healthcare, Houston, TX, USA

Introduction: Synonyms have different lexical forms but share the same sematic meanings. Automatically extracting synonyms has been researched to capture the different levels of semantic relatedness between two terms. As an informational retrieval approach, the Vector Symbolic Architectures (VSAs) has been utilized with distributional hypothesis, which is terms in the same contexts tend to have similar meaning. To overcome computation burden and poor scalability of Latent Semantic Indexing (LSI), Random Indexing[1] (RI) was successfully utilized. Moreover, RI encoding additional information shows the good performance to identify general English TOEFL synonym evaluation. This additional information includes relative orientation (Random Direction, RD) or the relative term position with order information (Random Permutation, RP) from a given term. Another variance, Term-based Reflective Random Index (TRRI) uses iterative vector generation to identify latent relationship among terms.

Methods: We investigated automatic synonym extraction from clinical documents using various distributional semantic vector models. Our clinical corpus consists of 189,099 clinical notes spanning two years extracted from a local clinical data warehouse. The corpus contains unique 82,655 unique terms with a minimum frequency of five or more occurrences, after pre-processing with Lucene and the exclusion of terms that contain non-alphabet characters. In this experiment, we used 490 unigram synonyms, which were automatically extracted from the UMLS/SNOMED-CT, and manually validated by one expert. We applied RI, RP, and TRRI to find the nearest neighbors of the cue terms in this gold standard set. The dimensionalities of the vectors used were 200, 500, 1000, and 1500 with real vector representation. For RD and RP, five different window sizes (window radius of 1, 2, 3, 4, and 5) were tested. For evaluation, we extracted the ten nearest neighboring terms from the model after post-processing (as per Henriksson's method[2]) to eliminate poor suggestions, and then calculated recall at k=10.

Results: Among 48 different models with different dimensions, window sizes, and RI variants, the best recall was 16.73 (1000 dimension, w=3, and RD). For all dimensions, RD models present the best performance with three windows as surrounding words (3 + the given word + 3). With respect to window size, window size of three shows the best recalls for all RD and RP models. All RI variants (RD, RP and TRRI) show better performance than the basic RI model. Higher dimensionalities (>= 1000 in our experience) tend to have better performance. The following Table shows the recall across different parameter settings.

Table. Recalls depending on dimensions, window sides, and variants of RI

	Dimension= 1000				Dimension= 1500			
	RD	RP	RI	TRRI	RD	RP	RI	TRRI
Window size = 1	10.41%	10.41%	1.63%	7.14%	10.20%	10.20%	2.04%	6.94%
Window size = 2	14.90%	12.86%			14.69%	13.27%		
Window size = 3	**16.73%**	13.88%			16.53%	13.67%		
Window size = 4	15.10%	12.86%			16.12%	13.27%		
Window size = 5	15.31%	12.04%			15.71%	12.45%		

* RD: Random Direction, RP: Random Permutation, RI: Random Index, TRRI: Reflective Random Index

Limitations: This study was conducted using a comparatively small English clinical corpus from a single institute with a reference standard that was not exhaustive. Therefore, comparing absolute values of our results to those obtained in Henriksson's study[2] is not meaningful. Nevertheless, our experiments confirm some of their findings (e.g. larger window and higher dimension improve results.) We plan to expand the size of our corpus and reference standard and to combine our models to improve performance in future work.

Conclusion: Variants of RI with different dimensions and window sizes were applied to clinical notes to extract synonyms automatically. RD models considering orientation with a window size of three and 1000 dimension show the best recall. That all of the variants of RIs present better performance then the original implementation of RI, suggests encoding this information is helpful for synonym detection, confirming results of previous studies. In the future, we plan to combine different models to improve performance.

1. Widdows Dominic and Kathleen Ferraro. "Semantic Vectors: a Scalable Open Source Package and Online Technology Management Application." LREC. 2008.
2. Aron Henriksson, Hans Moen, Maria Skeppstedt, Ann-Marie Eklund, Vidas Daudaravicius, and Martin Hassel. "Synonym Extraction of Medical Terms from Clinical Text Using Combinations of Word Space Models". In Proceedings of the 5th International Symposium on Semantic Mining in Biomedicine, 2012

Vocabulary Density Method for Customized Indexing of MEDLINE Journals

James G. Mork, MSc, Dina Demner-Fushman, MD, PhD,
Susan C. Schmidt, MLS, Alan R. Aronson, PhD,
Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, National Institutes of Health, DHHS, Bethesda, MD

Abstract

Automated indexing of MEDLINE citations remains a challenging problem due to the growing volume of citations and the over 27,000 MeSH indexing terms that can be assigned to them. This paper presents a corpus-based approach to improving indexing for specific journals. The Vocabulary Density approach takes into account frequencies of indexing terms previously assigned to a journal when recommending indexing terms for a new citation in that journal. After implementing the approach, we saw a 2.69 (4.44%) improvement in Precision.

Introduction

The successes in automatic indexing of MEDLINE® citations using the NLM Medical Text Indexer (MTI)[1] have led to First Line (MTIFL) indexing of several journals. MeSH terms automatically assigned by MTIFL provide the initial indexing to a MEDLINE citation which is then reviewed and completed by an indexer. As we expand the set of journals indexed via MTIFL, we continue seeking improvements to the MTI algorithms. Potential for improvement using journal-specific data was discussed by Tsoumakas et al[2]. To explore whether customizing MTIFL indexing of a journal is worthwhile, we have implemented a simple Vocabulary Density approach for all journals indexed by MTI.

Methods

We used 3,401,111 citations involving 6,606 journals from the 2014 MEDLINE Baseline that have been indexed over the last five years (henceforth referred to as *Corpus*). For each MeSH Heading (MH) used by each journal we captured the number of its occurrences (NOM) and the Number of Articles in the journal (NOA). We then normalized the frequency of each MH in the journal, computing Factor = NOM / NOA. For example, the MH "Swiss 3T3 Cells" occurred four times in the 2,231 articles of the journal "Biochemical Society (Great Britain)" in the *Corpus*. The Factor for this MH is 0.001793 (4 / 2231).

We applied the Vocabulary Density method to journals that had at least 80 citations in the *Corpus* and to MHs introduced at least a year ago. Given the Vocabulary Density information, MTI does not recommend MHs that are not used for the journal and automatically recommends MHs with a Factor > 0.74. For frequently occurring MHs, e.g., Female or Humans, the threshold for automatically recommending is 1 to reduce incorrect recommendations.

Results

On average, only 999 unique MHs of the 27,149 available in 2014 MeSH are used per journal in the 6,606 journals in our *Corpus*. 83.81% of the used MHs are found in 500 or fewer journals and 271 MHs are only found in a single journal (see Figure 1). This selective use of MHs confirms the intuition that taking into account journal-specific data can lead to improvements in MTI recommendations. Furthermore, implementing this simple approach leads to a 2.69 (4.44%) improvement in Precision, 1.36 (2.23%) increase in F_1 score, and a 0.05 (0.08%) increase in Recall.

Conclusion

The significant improvements in Precision without losses in Recall when taking into account the Vocabulary Density information for a journal shows that exploring other potential approaches to using the journal-specific data is worthwhile.

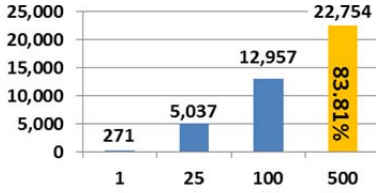


Figure 1. Cumulative MH Usage across Journals

References

1. James G Mork, Antonio J Jimeno-Yepes, Alan R Aronson. The NLM Medical Text Indexer system for indexing biomedical literature. BioASQ Workshop, Valencia, Spain, September 27, 2013.
2. Tsoumakas G, Laliotis M, Markantonatos N, VlahavasI. Large-scale semantic indexing of biomedical publications at BioASQ. BioASQ Workshop, Valencia, Spain, September 27, 2013.

Engaging Patient and Family Stakeholders in Developing Innovative Patient-Centered Care Interventions to Enhance Patient Experience

Conny Morrison,[1] Maureen Fagan, DNP, MHA, FNP-BC,[1] Priscilla Gazarian, PhD, RN,[1,2] Orly Tamir, PhD, MSc, MHA,[1] Jacques Donzé, MD, MSc,[1,3] Patricia Dykes, PhD, RN,[1,3] Diana Stade,[1] David W. Bates, MD, MSc,[1,3] Ronen Rozenblum, PhD, MPH[1,3]
[1]Brigham and Women's Hospital, Boston, MA; [2]Simmons College School of Nursing and Health Studies, Boston, MA; [3]Harvard Medical School, Boston, MA

Abstract: *The use of Patient and Family Advisory Councils (PFACs) are an important step towards patient-centered healthcare, yet the potential contribution of patients and families in developing innovative interventions has been neglected. Thus, we included patients, families and PFACs in the development of interventions aimed at enhancing patient experience. In order to do so, we conducted a focus group and interviews with patients and families to solicit their input on patient-centered care interventions. Our findings describe their contribution to intervention development that truly meets the needs of hospitalized patients and families.*

Introduction: Patient-centered care and patient engagement are part of a shift in healthcare that highlights the value of incorporating patients' needs and perspectives into care.[1] While Patient and Family Advisory Councils (PFACs) have emerged as a key means of involving patients and families in improving quality of care,[2] these stakeholders have been less involved in developing innovative patient-centered interventions. Thus, we engaged Brigham and Women's Hospital PFAC members and hospitalized patients and their families in developing and refining a structured communication model (Patient Satisf*Active*® Model[3]) and a web-based toolkit, interventions designed to improve patient experience in the Medical Intensive Care units (MICU) and Oncology units.

Methods: We selected four patient and family advisors (PFAs) with intensive care and/or Oncology experience to participate in a focus group and individual interviews. Semi-structured focus group and interview guides were developed based on systematic literature reviews and expert input. We interviewed 10 hospitalized patients and families about factors that promote dignity and respect. We also engaged patients and families in testing and refining the tools during 12 additional interviews. Data was analyzed qualitatively for themes. Finally, the patient and family input was reviewed by the MICU/Oncology leadership, to validate feasibility of proposed refinements.

Results: During the focus group, the PFAs shared their hospitalization experiences and gave input on ways to promote dignity and respect within the intervention tools. During the refinement interviews, patient and families tested toolkit usability, providing feedback on the interface and content. The participants' perspective allowed us to identify clinical behaviors and practices, as well as specific toolkit functions, that might enhance patient experience and promote dignity and respect in care; this input was integral to the refinement of the Patient Satisf*Active*® Model and web-based toolkit. In follow-up interviews, the PFAs reported that the focus group meaningfully involved them in the intervention development process. In addition, MICU and Oncology unit leadership recognized the value of these perspectives, and are now working to engage patients in other unit improvement projects.

Conclusions: Our experience validates the value of incorporating the PFAC and patient/family perspective in the development of any truly patient-centered intervention, including health information technologies. Further research is needed to develop valid models of PFAC involvement in healthcare improvement initiatives.

Acknowledgements: The Brigham and Women's Hospital PROSPECT project is part of the Libretto Consortium supported by the Gordon and Betty Moore Foundation.

References:

[1]Institute of Medicine. Crossing the quality chasm: a new health system for the 21st century. Washington, DC: National Academy Press; 2001.

[2]Johnson B. H., Abraham, M. R. Partnering with patients, residents, and families—a resource for leaders of hospitals, ambulatory care settings, and long-term care communities. Bethesda, MD: Institute for Patient- and Family-Centered Care; 2012.

User Preferences Influencing the Design of a Tailored Virtual Patient Educator in a Latina Farm Worker Community

Bijan Morshedi[1], Alexis Chaet[1], Cameron Brown, BA[2], Gloria Arroyo[2], Sara K. Proctor, DW[3], Rupa S. Valdez, PhD[1], Kristen J. Wells, PhD, MPH[2], Laura E. Barnes, PhD[1]
[1]University of Virginia, Charlottesville, VA; [2]San Diego State University, San Diego, CA; [3]Catholic Mobile Medical Services, Dover, FL

Abstract

As use of health information technology increases, the large digital divide between minority groups and the general population runs the risk of widening health disparities. This poster reports on the ongoing design progress for the development of a tailored, low-literacy Virtual Patient Educator (VPE) related to cervical cancer (CC) screening for a community of Latinas in a rural area of central Florida to be delivered in a clinic waiting room through a touch-screen tablet. The incidence of CC among US Hispanics is double that of the general population. These rates are influenced by cultural beliefs, linguistic barriers, socioeconomic status, and low levels of health literacy, which discourage CC screening. This research uses a participatory design approach, which involves integrating the end users of the technology as partners in the design process in order to guide technology development. The research reported in this abstract was supported by the National Cancer Institute and the Office of Research on Women's Health (R21CA167418).

Methodology

Six Latina women were recruited from a faith-based primary care clinic to answer open-ended interview questions on their opinion about the background format of a VPE. The development of this VPE prototype was informed by an initial pilot study consisting of twenty-six interviews within this population and a review focusing on Latino technology usage patterns[1]. Interviews were conducted in Spanish by a native-speaking patient navigator working at the clinic site. Patients were shown three images that only differed in their background color schemes on a touch screen tablet depicting a woman within a home setting to be used for the VPE. Patients were asked a series of questions including their impressions of the furniture in the setting, the appearance of the woman, the room decor, and the wall color. Patients were asked to select their favorite and least favorite settings. Patients were then presented with a panel of button icons. The panel contained five groupings of typical symbols used within American and Latina cultures to depict various button commands: "yes," "no," "next," "help," and "repeat." Patients were asked to state their button preferences for each command. Patients' responses were audio-taped and transcribed verbatim. The San Diego State University and the University of Virginia IRBs approved this study.

Evaluation Results and Conclusions

Background. The majority of participants (n=3) preferred Background 1 (light green) due to the lighter color of the walls, and disliked (n=4) Background 3 (blue) due to the dark or saddening nature of the color scheme. While only one participant liked Background 2 (beige), three patients identified this background as most calming, with two participants identifying background 1 as calming. Overall, patients associated lighter colors with happier, more calming environments. Opinion was divided on the appearance of the furniture; three participants liked the furniture and three participants identified the sofa as uncomfortable. The majority of participants recommended changing the pear painting to another image such as a flower, landscape, or the sun, stating that these images are more comforting and made the setting appear more like a home and less like an office or clinic. Four participants responded that they would feel comfortable being placed within the depicted environment, while two individuals identified the background as "clinic-like," evoking feelings of nervousness.

Buttons. The majority of patients identified the circular-shaped "thumb-up" (n=3) and "thumbs-down" (n=4) symbols as most fitting for the "yes" and "no" buttons, respectively, preference for bold-colored symbols. The colored tailed-arrow was preferred (n=2) for "next" over the white tailed-arrow (n=1) and the arrowhead without the tail (n=0). Bold-colored symbols were preferred for the "next" category as well. There was no one preference identified the question mark button representing "help." Pink was the preferred color option for "repeat" (n=4) over blue. The majority of patients correctly identified each button symbol with the intended button function.

The patient feedback gathered through this study will be used to refine background layout and functionality to design a VPE best tailored to the patients' preferences within this community. Despite the population's receptiveness to a user-centered approach to designing the VPE, not all users may agree on the preferred system attributes. Consequently, we intend to continue to engage community members as we progress with our iterative design process.

References

1. Barnes LE, Rivera M. Meade CD, Proctor SK, Gutierrez LM, Wells KJ. Feasibility Study for Technology-Based Cancer Education for Latina Women from an Agricultural Community. Cancer Epidemiol Biomarkers Prev. 2011; 20(Supplement 1): A44.

An Informatics Framework for Clinical and Translational Research: The Mizzou Approach

Abu S. M. Mosa, MS[1,2], Nate C. Apathy, BSBA[3], Kelly J. Ko, PhD[3], Jerry C. Parker, PhD[1]
[1]Institute for Clinical and Translational Science, University of Missouri; [2]Informatics Institute, University of Missouri, Columbia, MO; [3]Cerner Corporation, Kansas City, MO

Abstract
One of the main challenges in clinical and translational research is providing secure access to clinical data and tools. In this poster, we will demonstrate a workable and secure clinical research informatics framework that has been implemented at the University of Missouri (MU). This framework leverages the secondary reuse of MU Healthcare data for cohort identification and retrospective data analysis, and provides an online application for project and data management.

Introduction
Providing informatics applications to support clinical research is of high importance among the research based university hospitals. The Institute for Clinical and Translational Science (ICATS) leads the development of informatics framework (Figure 1) for clinical research at the University of Missouri (a.k.a. Mizzou) Healthcare System. It has two components: (1) a clinical data warehouse, and (2) an online project management tools.

Clinical Research Informatics Framework

Clinical Data Warehouse: We implemented a clinical data warehouse using an open-source system called i2b2 [1]. The data warehouse has two data repositories: identified and de-identified. In the de-identified data repository, protected health information (PHI) is removed following the HIPAA regulations. The users have three levels of access: basic, expanded and full. With basic access, the users can search the de-identified data repository to generate aggregated results. The users can download de-identified data with proper IRB application through expanded access for a certain period of time. The honest broker has full access to identified data repository who deposits PHI data for IRB approved projects in the research data repository.

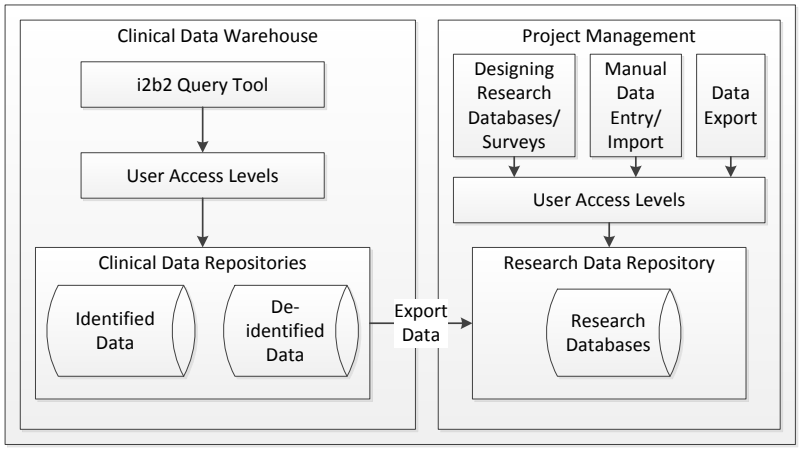


Figure 1: Clinical Research Informatics Framework

Online Project Management: We implemented the REDCap [2] system in which the research/survey databases are created and stored securely. The access levels are controlled based on the user's role in the projects. The study coordinators (e.g. principal investigator) have full access to the projects and data, the data entry users can create and modify records, and the data analysts can download de-identified data only.

Conclusion
A useful informatics framework for clinical research has been implemented as a foundation upon which other tools and resources will be developed.

References

[1] Murphy SN, Mendis M, Hackett K, Kuttan R, Pan W, Phillips LC, Gainer V, Berkowicz D, Glaser JP, Kohane I, Chueh HC. Architecture of the open-source clinical research chart from Informatics for Integrating Biology and the Bedside. AMIA Annu. Symp. Proc. 2007;11:548–52.

[2] Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J. Biomed. Inform. 2009;42(2):377–81.

Developing a Knowledge Base for Detecting Carotid Stenosis with pyConText

Danielle Mowery, PhD[1,5,6], Daniel Franc, MD[2], Shazia Ashfaq, MD[3,4], Tania Zamora, BA[4], Eric Cheng, MD, PhD[2], Wendy W. Chapman, PhD[5,6], Brian E. Chapman, PhD[6]
[1]University of Pittsburgh, Pittsburgh, PA; [2]VA Health Care System, Los Angeles, CA; [3]Veteran Medical Research Foundation & [4]VA Health Care System, San Diego, CA; [5]VA Health Care System & [6]University of Utah, Salt Lake City, UT

Abstract: *We present a pilot study to develop a semantic schema for extracting mentions of carotid stenosis and their modifiers from ultrasound reports to extend a natural language processing (NLP) algorithm, pyConText.*

Introduction: 795,000 people suffer strokes each year. Comparative-effectiveness studies can determine whether one treatment is more effective than another for a stroke cohort. However, these studies rely on manual chart abstraction which is expensive and tedious. NLP can help by automatically abstracting carotid stenosis mentions and their modifiers, applying rules to assert the severity of stenosis (no stenosis to severe) for each encounter, and tracking patient responses (severity changes) over time. We developed a model to support this use case.

Methods: In this pilot study, we selected 29 de-identified, carotid ultrasound reports from www.mtsamples.com to develop a schema and guidelines that support information extraction of carotid stenosis mentions (all authors). Two physicians (DF, SA) independently annotated each report and adjudicated each disagreement with consensus review using eHOST annotation tool[1]. For each modifier, we report an example, the frequency (proportion) of stenosis modifications, the partial IAA (F1-score for partial span, same attribute), and the distribution of lexical cues.

Results: We observed 177 **Findings** representing carotid stenosis with the following distribution of modifiers:

Modifier	*Example*	Count (%)	IAA	Lexical cue/regular expressions (Counts)
Finding	Severe *stenosis* observed	177 (--)	96%	*steno(ses|sis)* (111), *occlu(ded|sion)* (26), *carotid|caritid stenosis* (16), *plaque* (10), *wall thickening* (5), *hypoplastic* (4), *carotid|caritid artery disease|stenosis* (4), *atherosclerotic changes* (1)
Existence Indicator	*probable* stenosis	13 (7%)	96%	*no|without evidence* (6), *no* (5), *could be* (2),
Severity Indicator	*45-55%* occlusion	44 (25%)	97%	*less than \d{2}%|(<)\d{2}%* (14), *\d{2}%(-|to) \d{2}%* (10), *(hemodynamically) significant* (8), *mild* (4), *appreciable* (2), *moderate* (2), *complete* (1), *very severe* (1), *near* (1)
Neurovascular Anatomy	*internal carotid* is occluded	44 (25%)	95%	*ica|internal|vertebral|common|external carotid artery|arteries* (34), *carotid(s)* (7), *carotid artery|arteries* (3)
Anatomic Modifier	*proximal* carotid artery	5 (3%)	91%	*proximal* (3), *bifurcation* (1), *cavernous portion* (1),
Sidedness	*R* ICA occluded	43 (24%)	95%	*bilateral(ly)* (16), *r|right(-sided)* (13), *l|left(-sided)* (12), *either side* (1), *both* (1)
Exam Quality	*fully visualized* stenosis	2 (1%)	100%	*not visualized* (2)
Reason for Exam	*REASON:* R/O stenosis	26 (15%)	97%	*assess|eval(uate) for* (12), *r/o|rule out* (7), *?* (3), *reason* (3), *indication* (1)

Conclusions: Most carotid stenosis **Findings** are modified by cues for **Severity Indicator**, **Neurovascular Anatomy**, and **Sidedness**. All modifiers are annotated with high agreement. Most **Severity Indicators** are expressed as a range e.g., "60-69%" or "less than 40%" followed by qualitative expressions e.g., "significant stenosis" or "mild plaque". We are iteratively annotating data and adding new regular expressions and rules to train pyConText[2] to detect carotid stenosis mentions from the Office of Quality and Performance Stroke Study dataset.

Acknowledgements: This work was funded by VA HSR&D Stroke QUERI RRP 12-185, NIH NHLBI 1R01HL114563-01A1, & NIGMS R01GM090187 and was approved by the Department of Veteran Affairs, University of Utah Institute Review Board, & MTSamples transcription company.

References:

1. South BR, Shuying S, Leng J, Forbush, TB, DuVall SL, Chapman WW. A prototype tool set to support machine-assisted annotation. BioNLP. 2012;130-139.
2. Chapman B, Lee S, Kang H, Chapman WW. Document-level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm. J Biomed Inform 44 (2011) 728–737.

A New Alternative for Accessing Medicare Claims Data: The Virtual Research Data Center

Mallika Mundkur, MD[1], Swapna Abhyankar, MD[1], Alex Constantin, PhD[2], Leighton Chan, MD, MPH[2], Clement J. McDonald, MD[1],
[1]National Library of Medicine, National Institutes of Health, Bethesda, MD
[2]Rehabilitation Medicine Department, Clinical Center, National Institutes of Health, Bethesda, MD

Abstract
In November 2013, the Centers for Medicare and Medicaid Services (CMS) launched the Virtual Research Data Center (VRDC), a tool for accessing Medicare claims data[1]. The VRDC enables wider use of these data in a more secure and cost-effective format compared to the previous method of delivering encrypted files to researchers on a CD or hard-drive. We describe key features of the VRDC from an end-user's point of view as a reference for other researchers interested in using Medicare data.

Introduction
The Medicare claims data are a valuable resource for conducting epidemiological and health services research. Historically, CMS provided claims data in the form of encrypted files on CD or hard-drive. The VRDC provides virtual data access and obviates the need for managing physical hardware and file encryption. Interested researchers currently have the option to choose either physical or virtual access to Medicare files.

CMS Data Files Available for Request
More than ten years' worth of data, from 1999 to 2012, are available in electronic format and accessible through the VRDC. This includes claims data from Medicaid, Medicare Parts A (Inpatient), B (Outpatient), and D (Prescription Drugs), the Master Summary Beneficiary File, which includes the National Death Index, and the Medicare Provider Analysis and Review File[2]. The Long-Term Care Minimum Data Set is also available, which includes standardized assessments of physical, psychological, and psychosocial functioning of all patients in a Medicare- or Medicaid-approved long-term care facility.

Request Process and Data Cost
Requesting either the physical or virtual data is a multi-step process, which we will illustrate with a process map on our poster. The cost structure of the VRDC compared to the physical data request is one of the main differences between the two. The cost of the physical data is driven by the number of files and years of data requested; for large data samples this could reach several hundreds of thousands of dollars or more. In contrast, access through the VRDC is fixed per user per year, currently $25,000 for federal and $40,000 for non-federal users. Additional users cost $15,000 per user per year[2]. Our poster will also include a table describing the cost structure in more detail.

Technical Aspects of Data Access
The VRDC is a secure central repository that researchers access via VPN and provides 500GB of disk space per user. The Statistical Analysis System (SAS) Enterprise Guide (EG), is available in VRDC's virtual environment for researchers to conduct statistical analysis, and then download aggregate analysis results. The VRDC must be accessed via a Windows operating system.

Conclusion
We conclude that the VRDC offers a highly secure and cost-effective approach to accessing Medicare data and should be the preferred approach for researchers who want to analyze large samples of the Medicare population.

References

1. CMS.gov. *Press Release: CMS Announces New Data Sharing Tool.* http://www.cms.gov/Newsroom/MediaReleaseDatabase/Press-Releases/2013-Press-Releases-Items/2013-11-12.html. (accessed 13 Mar 2014).
2. Research Data Assistance Center. *CMS Virtual Research Data Center (VRDC) FAQs.* http://www.resdac.org/cms-data/request/cms-virtual-research-data-center. (accessed 13 Mar 2014)

Facilitating Visual Exploration of System-generated Reasoning Pathways underlying Drug-Side effect Relations

Sahiti Myneni, PhD, MSE[1], Trevor Cohen, MBChB, PhD[1]

[1]The University of Texas School of Biomedical Informatics at Houston, TX, USA

Abstract

Scalable literature-based discovery methods can infer relationships between biomedical entities of interest from large volumes of computable biomedical knowledge. However, exploring these connections is challenging given the combinatorial explosion of possible reasoning pathways connecting the entities. This work provides an interface to explore interesting inferences derived using geometrically-motivated computational models of analogical reasoning. Principles from cognitive engineering, visual analytics and discovery browsing are brought together to develop a novel interaction paradigm for information retrieval systems.

Introduction

Scalable methods of literature-based discovery (such as Predication-based Semantic Indexing) are able to infer a multiplicity of relationships between biomedical entities of interest[1]. Representing the system-generated patterns and supporting evidence in a user interpretable way plays a key role in enabling the biomedical scientists act upon the knowledge extracted by the system from the literature. However, these patterns may represent thousands of unique reasoning pathways. Therefore, new user interaction paradigms are required, to allow users to navigate through underlying evidence efficiently. Unlike conventional document-based search engines, Semantic MEDLINE[2] facilitates summarization and visualization of retrieved documents, motivated in part by principles developed to mediate literature based discovery. It allows users to construct argumentation through iterative combination of key concepts and relations. Inspired by the "discovery browsing" principle[3] that motivates Semantic MEDLINE, we develop a novel interaction paradigm for information retrieval systems that allow users to explore computer-generated reasoning pathways suggesting a plausible relationship between a drug and potential side-effect.

Methods

Techniques from visual and network analytics in conjunction with principles of cognitive engineering and discovery browsing were used. For illustration purposes, we chose to represent the pathways supporting the assertion that "rosiglitazone CAUSES myocardial_infarction". A total of 108100 unique pathways involving 390 unique middle terms were connecting this drug-side effect pair. The middle terms were grouped into 29 UMLS semantic types. We used D3 and Jquery to develop a system for inference visualization and user exploration of the pathways.

Results

As shown in Figure 1, the initial framework consists of four modules. The first module allows the user to choose a semantic type, which consists of a set of middle terms connecting the drug and side effect. The second module allows the user to sort the middle terms using inference metrics and underlying evidence from the literature. The third module allows the user to explore the system-generated reasoning pathways and connecting middle terms by means of collapsible nodes. The fourth module, at the bottom, provides an organizer such that the user can selectively explore aspects of these pathways that are interesting and unusual.

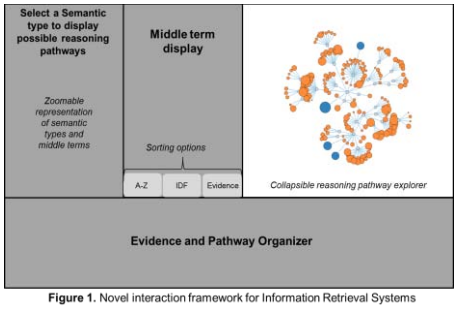


Figure 1. Novel interaction framework for Information Retrieval Systems

Ongoing work

The user interaction framework will incorporate geometrically-motivated models of reasoning within contemporary visualization approaches to provide users with the ability to explore reasoning pathways and literature that supports the hypothesis that a drug and side effect are causally related. More sophisticated user-centered approaches to enable the exploration of pathways in the proposed system will be studied and integrated.

Acknowledgement: The work was supported by the US National Library of Medicine Grant (1R01LM011563), Using Biomedical Knowledge to Identify Plausible Signals for Pharmacovigilance.

References

1. Cohen T, Widdows D, Schvaneveldt RW, Davies P, Rindflesch, TC. Discovering discovery patterns with predication-based Semantic Indexing. Journal of biomedical informatics.2012;45(6):1049-65.
2. Cairelli J, Miller CM, Fiszman M, Workman TE, Rindflesch TC. Semantic MEDLINE for Discovery Browsing: Using Semantic Predications and the Literature-Based Discovery Paradigm to Elucidate a Mechanism for the Obesity Paradox. In AMIA Annual Symposium Proceedings; 2013 Nov 16-20;Washington,DC.p.164-73.
3. Hristovski D, Peterlin B, Mitchell JA, Humphrey SM. Using literature-based discovery to identify disease candidate genes. International journal of medical informatics.2005;74(2),289-98.

Prevalence of a standardized storage format and next steps for clinicians

Masaharu Nakayama, MD, PhD[1,2]

[1]Disaster Medical Informatics, International Research Institute of Disaster Science, Tohoku University, Sendai, Japan; [2] Medical Informatics Center, Tohoku University Hospital, Sendai, Japan

Abstract

Sharing medical information among hospitals remains challenging due to differences in vendor specifications. A standardized format for storage, the Standardized Structured Medical Information eXchange (SS-MIX), has spread nationwide and has been utilized in several regional and national projects on medical information systems in Japan. The format is mainly divided into two categories: standardized storage and extended storage. The latter should be determined in detail for clinical use.

Introduction

Sharing medical information among hospitals is crucial for patient care. However, the interfacility exchange of patient data is difficult because of differences in vendor specifications. Kimura et al. developed the Standardized Structured Medical Information eXchange (SS-MIX) [1], which enables data from medical record systems developed by different vendors to be stored in a similar format. SS-MIX was modified and released as SS-MIX2 in 2012. Several projects have utilized this format to share and back up patient data in Japanese hospitals[2]. The number of hospitals with Uploader of SS-MIX2 was 216 in June 2013.

Two categories of SS-MIX

SS-MIX2 is divided into two categories: standardized storage and extended storage. Standardized storage includes clinical data in standard form (HL7 v2.5), such as the basic patient data, prescriptions, laboratory results, and disease name registration. These data are easily utilized for secondary use. Other important clinical data, however, have not been structured and should be stored in the extended storage. Therefore, these data should be structured, coded, and standardized.

IHE

The Integration healthcare enterprise (IHE) is an international association that aims to improve the way in which healthcare computer systems share information, and sets profiles for the coordinated use of established standards such as the Digital Imaging and Communications in Medicine (DICOM) and Health Level Seven (HL7). One of the IHE's activities, "connectathon," provides connectivity testing across a variety of vendor platforms and validates the participants' interoperability and compliance with the IHE profile. In Japan, IHE-Japan (IHE-J) is also active in promoting IHE profiles in accordance with the Japanese Health System through collaboration between technicians from vendor companies and clinicians from several hospitals.

An important next step should be the standardization of clinical data from various modalities, such as electrocardiogram, ultrasound cardiogram, and catheterization. IHE-J cardiology has commenced the initial steps for standardization with the help of several academic cardiology societies. This will be critical toward sharing important clinical data, which will benefit specialists who require such data for clinical use and research purposes.

Conclusion

Although SS-MIX has spread nationwide in Japan, clinicians still require extension of structured storage formats for the utilization of clinical data regardless of differences among vendors or modalities. To this aim, standardization should be spread promptly. IHE-J cardiology has begun to standardize data from several modalities in cardiology by collaboration with academic associations. This step will be important for the development of medical care.

References

1. Kimura M, Nakayasu K, Ohshima Y, Fujita, N, Nakashima N, Jozaki H, Numano T, Shimizu T, Shimomura M, Sasaki F, Fujiki T, Nakashima T, Toyoda K, Hoshi H, Sakusabe T, Naito Y, Kawaguchi K, Watanabe H, Tani S. SS-MIX: A Ministry Project to Promote Standardized Healthcare Information Exchange. Methods Inf Med 2011;50:131–9.
2. Nakashima N. Japanese sentinel project and contribution of laboratory medicine. Rinsho Byori. 2013;61:501–10.

Achieving a Better Readability of Instructional Notes in the ICD-10-CM Tabular XML Files

Hari Nandigam MD, MSHI [1], Nima Behkami Ph.D[2]

[1]*Partners Healthcare System, Boston, MA.* [2]*Harvard Medical School, Center for Biomedical Informatics, Boston, MA, Boston, MA*

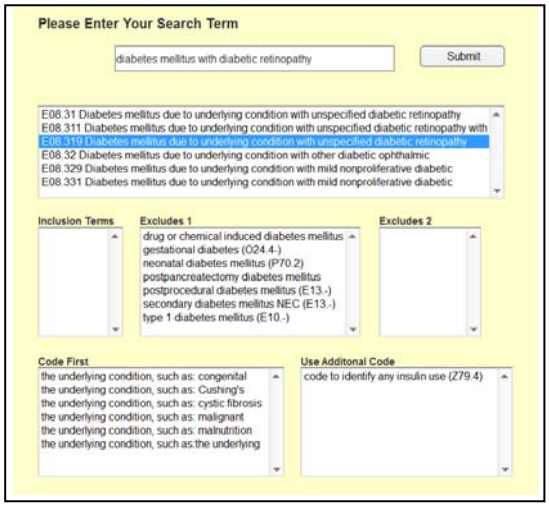
Introduction and Background: The Center for Medicare & Medicaid Services (CMS) published ICD-10-CM files in xml format in addition to pdf format. The advantage of having tabular and index files in xml format is that they can be uploaded to a database and easily queried. The tabular files published by CMS use colons (:) indicated in the text of the elements such as 'excludes1', 'excludes2', 'use additional code' and 'code first' to indicate that the condition in the first clause applies to the subsequent clauses. However the subsequent clauses are modeled at the same level as the parent clauses. For example the element -<excludes1> in 'G01' contains values such as 'meningitis (in):',' gonococcal (A54.81)', 'leptospirosis (A27.81)' etc. Ideally they are meant to be present as 'meningitis (in) gonococcal (A54.81)', 'meningitis (in) leptospirosis (A27.81)' etc. We came up with a model to explicitly indicate in this way using a transformation tool.

Methods: Using a transformation tool (Altova Mapforce) we loaded the initial version of 2015 ICD-10-CM tabular files and introduced logic to indicate that when the program sees a colon in the text of <excludes1>, <excludes2>, <code first> and <use additional code>, it explicitly indicate the condition before the colon in the second and subsequent clauses. This automation process helped us to handle the issue with <excludes1>, <excludes2>, <code first> and <use additional code>.

2015 ICD-10-CM tabular file	Modified ICD-10-CM tabular file
<diag> <name>G07</name> <desc>Intracranial and intraspinal abscess and granuloma in diseases cla <codeFirst> <note>underlying disease, such as:</note> <note>schistosomiasis granuloma of brain (B65.-)</note> </codeFirst> <excludes1> <note>abscess of brain:</note> <note>amebic (A06.6)</note> <note>chromomycotic (B43.1)</note> <note>gonococcal (A54.82)</note> <note>tuberculous (A17.81)</note> <note>tuberculoma of meninges (A17.1)</note> </excludes1> </diag>	<diag> <name>G07</name> <desc>Intracranial and intraspinal abscess and granuloma in diseases classified elsewhere</desc> <codeFirst> <note>underlying disease, such as: schistosomiasis granuloma of brain (B65.-)</note> </codeFirst> <excludes1> <note>abscess of brain: amebic (A06.6)</note> <note>abscess of brain: chromomycotic (B43.1)</note> <note>abscess of brain: gonococcal (A54.82)</note> <note>abscess of brain: tuberculous (A17.81)</note> <note>abscess of brain: tuberculoma of meninges (A17.1)</note> </excludes1> </diag>

Results: We modeled the tabular xml file and uploaded in the database. Upon submitting a term in the free text field, we get matching ICD-10-CM codes and their corresponding descriptions in the list box. On selecting each term in the listbox, their corresponding values for inclusion terms, Excludes1, Excludes2, Code First and Use additional Code show up in the additional list boxes below.

Discussion and Conclusion: we noticed at the time of this paper that 2015 version of ICD-10-CM is better than the older versions in explicitly indicating about the clauses. However we noticed the discrepancies with colons weren't handles in a consistent manner. Automation of this process will be very useful. A manual review is always recommended to verify if right changes are done by doing a comparison of the initial and final files. We took care not to deal with colons indicated within the elements such as 'visualImpairment', 'SevenChrNote' and 'Notes' in the tabular file. This transformation/tweaking of the tabular xml file helped in better expressivity, readability and understandability of the instructional notes



References:

1. ICD-10-CM Official Guidelines for Coding and Reporting 2015, Center for Medical Services, Editor. http://www.cms.gov/Medicare/Coding/ICD10/2015-ICD-10-CM-and-GEMs.html

The Can (thecan.apphb.com): a repository of decision support rules relating to laboratory tests

Scott D. Nelson, PharmD;[1,2] R. George Hauser, MD[3]
[1] Department of Veterans Affairs Medical Center, Salt Lake City, UT; [2] University of Utah, Salt Lake City, UT; [3] Laboratory Medicine, Yale-New Haven Hospital, New Haven, CT

Abstract
Clinical decision support (CDS) in health care has an important role in improving quality and appropriateness of patient care. We set out to develop a self-sustaining repository of best-evidence CDS rules for laboratory test ordering and interpretation through networking and collaboration. The first stage consists of the development of a rules repository (thecan.apphb.com), and populating it with laboratory monitoring rules. The second stage consists of developing a method for converting the natural language CDS rules into computable rules using a layered framework with and rule validation. The input of rules into the repository, "The Can", is accomplished through networking and collaborative efforts. For more information, and to participate, please visit: thecan.apphb.com

Background
Clinical decision support (CDS) in health care has an important role in improving quality and appropriateness of patient care. While much effort has been given to CDS in medication ordering, there is also a need for CDS in laboratory test ordering and interpretation. The challenge is that the practices of medicine and laboratory pathology are constantly changing, and a single organization may not have the resources and funding to create all of their own CDS rules. Additionally, while lab utilization research has been around for a while, finding a list of rules or recommendations is difficult. Therefore a method for creating, sharing, and updating CDS rules related to laboratory testing is needed.

Project goal
We set out to develop a self-sustaining repository of best-evidence CDS rules for laboratory test ordering and interpretation through networking and collaboration methods. The repository is named "The Can".

Methods
The first stage of the project is to create a repository for the laboratory test rules, and populate it with rules using natural language. We will recruit subject-matter experts (SMEs) with an interest in laboratory testing or lab utilization from outside our development committee for the input of laboratory test CDS rules. Market campaigning and recruitment of SMEs will be done through scientific meetings, professional panels, and peer-reviewed literature publications. These experts will then be asked to recruit others to contribute to and use the CDS repository. Similar successful attempts at this type of collaborative effort for placing expert knowledge into a community repository, due to usefulness of their content, are that of Stack Overflow and Wikipedia. SME input and use will be driven by repository content due to the difficulty of finding CDS lab rules from other sources. To encourage the generation of content in the repository, we have developed the website with priorities in access and usability. The website will help guide SMEs through a structured process for defining their CDS rules to facilitate later conversion to a computable format, using a layered framework. This common rule repository developed by SMEs will help empower SMEs (and their organizations) to use and trust the repository, increase collaboration between organizations, and keep 'The Can' flexible and adaptive to current user needs and the constantly evolving field of medicine. In order to increase evidence-based external validity and generalizability, the rules in the repository require at least one reference and peer review for validation. Ideally, the rules in 'The Can' would be computable and interoperable across systems; however, the current lack of national interoperability requires that institutions translate the rules into their specific systems. The second stage of the project consists of developing a method for making the CDS rules computable by mapping to terminology standards and through validation and curation process. While not currently computable, 'The Can' provides a repository for sharing CDS rules and promotes future interoperability.

Conclusion
To date, there have been over 100 validated rules added to the repository, with 9 collaborating organizations. While the project is new and currently a content repository in the first stage, we are piloting the implementation of some of the rules at a local institution. By networking and collaborating with other SMEs for CDS rule development, we will develop a scalable and up-to-date repository of laboratory test utilization and interpretation CDS rules. 'The Can' is a valuable resource for laboratory CDS rules. For more information, and to participate, please visit: thecan.apphb.com

RxClass - Navigating between Drug Classes and RxNorm Drugs

Thang Nguyen, M.S., Lee Peters, M.S., and Olivier Bodenreider, M.D., PhD
U.S. National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA
Contact information: RXNAVINFO@LIST.NIH.GOV

Motivation

Drug classes constitute important information about the drugs and are critical to important use cases, such as clinical decision support (e.g., for allergy checking). *RxNav*, our RxNorm browser, already displays the classes for RxNorm drugs, but its drug-centric perspective does not accommodate the exploration of drug classes. This is the reason why we developed a web-based companion browser, *RxClass*, which supports navigation between RxNorm drugs and drug classes from several sources, including ATC, MeSH, NDF-RT and Structured Product Labels from the Food and Drug Administration (FDA).

Drug Class Sources

The Anatomical Therapeutic Chemical drug classification (***ATC***) is a resource developed for pharmacoepidemiology purposes by the World Health Organization Collaborating Centre for Drug Statistics Methodology.

The Medical Subject Headings (***MeSH***), developed by the National Library of Medicine (NLM), provides a rich description of pharmacological actions for the purpose of indexing and retrieval of biomedical articles.

The National Drug File-Reference Terminology (***NDF-RT***), developed by the Department of Veterans Affairs, provides clinical information about drugs (therapeutic intent, mechanism of action, etc.) and integrates drug classes from FDA's Structured Product Labels.

Linking RxNorm Drugs to Drug Classes

Like *RxNav, RxClass* is supported by functions from an application programming interface (API), which can be used independently for integrating drug class information in programs. The API serves the latest information available from the drug information sources.

RxClass provides a graphical interface to explore the hierarchical class structures of each source and examine the corresponding RxNorm drug members for each class. Some features of *RxClass*:

- The user can navigate through the drug classes via the hierarchical menu, or use the search feature to identify a drug class or RxNorm drug.
- Drug class member results can be saved to a file in several different formats.
- Users of *RxNav* can link to *RxClass* to get class members.

A sample screenshot of *RxClass* is shown in Figure 1.

Conclusions

Providing drug class members has long been a missing piece of information in *RxNav*. With *RxClass*, we now provide a link between drug classes from various sources and RxNorm drugs.

RxClass is available from the main *RxNav* website (http://rxnav.nlm.nih.gov), along with additional information about *RxNav* and our APIs to various drug information sources.

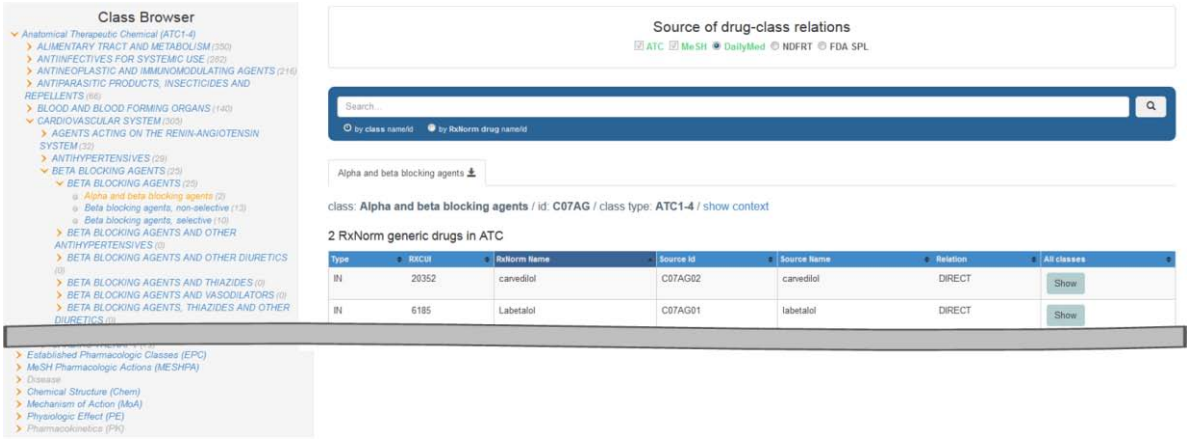


Figure 1. Sample screenshot of RxClass

Acknowledgments: This work was supported by the Intramural Research Program of the NIH, National Library of Medicine.

Patients and providers using secure messages on the patient portal for home telemonitoring: what are the informatics challenges and how can they be managed?

Frederick North, William Ward, Muhamad Elrashidi, Karen Ytterberg, Barbie Mundt, Eric Manley, SidnaTulledge-Scheitel

Introduction: Home telemonitoring is available for a number of patient measurements including blood pressure, blood glucose, weight, and oxygen saturation among others. Telemonitoring has generally sent data directly from a home monitor to an in home hub which then transmits data to a server. The patient portal has opened up another secure platform for telemonitoring data. Patients can self monitor blood pressure, glucose and other parameters and send the results to their provider in a secure portal message. We examined all secure messages to the Mayo Clinic patient portal from 2012 to determine use, quality and outcomes of messages for telemonitoring blood pressure.

Methods: We searched all 56,000 Mayo Clinic Rochester patient-generated secure messages from 2012 for content concerning blood pressure. We performed a detailed review on a random sample of 200 blood pressure messages. Measures for review included blood pressure reading counts, type of blood pressure data (individual readings, averages, and blood pressure range). We also examined whether the data provided was sufficient to treat blood pressure with a new medication. The message was evaluated for the blood pressure context (follow up blood pressure or specific elevated blood pressure concern). We captured the provider secure message response to the patient-generated blood pressure message and determined outcomes of prescriptions for antihypertensives, change in medication dose or if an appointment was suggested.

Results: Almost 1% of the 56,000 secure messages contained content about blood pressure. Of those messages with blood pressure content, the information was quite variable. Patients sent qualitative information (my blood pressure is too high) and qualitative information (blood pressure averages, ranges, and individual readings). Some sent lists of blood pressure readings. There was little specific information on dates the blood pressure was recorded, time of day, or conditions (during rest or activity). Few of the patient-generated messages had any evidence of medication reconciliation, pharmacy information, or other information usually needed to respond with a medication dose change or change in medication (new prescription). Provider responses also varied. There were responses to change medications and to start new blood pressure medications.

Conclusion: Patients and providers are using secure messages on the patient portal for blood pressure telemonitoring. Despite blood pressure content that is highly variable, providers respond with secure messages containing instructions for changes in medication and new antihypertensive medications. Our findings demostrate the informatics challenges of using secure messages on the patient portal for blood pressure monitoring. We suggest a specific template with required patient provided fields as an approach to managing patient-generated secure message home blood pressure readings. This could help standardize home blood pressure monitor readings for incorporation into the EMR. The template also requires patients to include other medication and pharmacy preferences so that providers have sufficient information to make medication changes.

Attitudes Towards Electronic Medical Records in Intensive Care

John C. O'Horo, MD, MPH, Pablo, Pickering, Vitaly Herasevich MD, PhD

Abstract: The degree to which electronic medical record (EMR) systems fit into intensive care unit (ICU) workflows is not well studied. We conducted a survey of critical care providers to determine the degree to which current systems fit their needs. Generally, EMRs were perceived as containing useful information, but not presenting it in a user friendly way that facilitates patient care in this setting.

Introduction: The fast pace of the intensive care unit (ICU) demands readily accessible and up to date data. However, most electronic medical records (EMRs) are designed for the outpatient or general care inpatient setting, and are not optimized for the ICU environment. As part of the implementation of a novel ICU EMR interface, the Ambient Warning and Response Evaluation (AWARE) system, we collected baseline data on attitudes towards the EMR in ICU providers.[1]

Methods: Critical care providers, including physicians, nurse practitioners, physician assistants, respiratory therapists and nurses, were surveyed at two tertiary care clinical campuses (Mayo Clinic Arizona and Florida). Providers were asked 10 questions about how accurate and user friendly the current EMR is on a 5 point Likert scale, with 5 being strongly agree, and 1 being strongly disagree. For summary analyses, both 4 and 5 were considered agreement, and 1 and 2 disagreement. Results were analyzed using JMP 9.0 statistical software (SAS, Cary, NC).

Results: We received responses from 233 providers; 63% of respondents were nurses, 24% respiratory therapists and the remainder physicians and midlevel providers. Most (89%) had worked in the ICU for more than a year, and the majority (62%) for more than 5 years. Ninety-four percent had worked with the existing medical record for at least a year. Overall, the majority of users thought the current EMR provided accurate information relevant to provider needs. However, fewer users were satisfied with the time required to use the EMR, with only 60% of users agreeing or strongly agreeing with the statement "I get the information I need in a timely manner using the EMR." Only 51% thought that the current EMR provides information in a useful format. More than 1/3 of providers thought he EMR made data gathering challenging. 55% felt that EMR use decreased the amount of time spent with patient/family. Prescribing providers (physicians, NP/PA) were more likely to be critical of the layout and efficiency of the EMR than non-prescribing providers.

Conclusion: Although the content of the current EMR provides relevant and accurate data, providers were not satisfied with the way in which data was presented. We will aim to improve these interface and efficiency issues with the novel AWARE interface.

References:

[1] Pickering BW, Herasevich V, Ahmed A and Gajic O. "Novel representation of Clinical Information in the ICU: Developing User Interfaces which Reduce Information Overload." *Applied Clinical Informatics*. 2010 1(2) p 116-131

Question	Prescribing providers	Nonprescribing providers	P value
The Electronic Medical Record (EMR) provides the precise information I need.	3.53 (1.10)	3.96 (0.81)	<0.01
The information content in the EMR meets my needs	3.66 (1.07)	3.99 (0.71)	0.10
The EMR provides me with sufficient information	3.75 (0.92)	4.02 (0.65)	0.11
I get information I need in a timely manner using the EMR	3.06 (1.22)	3.56 (1.03)	0.04
The EMR provides up-to-date information	3.69 (0.90)	3.85 (0.83)	0.34
I am satisfied with the accuracy of the EMR	3.59 (0.91)	3.92 (0.75)	0.06
I am satisfied with the way information is presented in the EMR, it is provided in a useful format	2.72 (1.25)	3.40 (1.12)	>0.01
The EMR makes the task of data gathering difficult/demanding	3.09 (1.17)	3.05 (1.13)	0.84
The EMR's display is designed to help increase patient safety	2.81 (1.23)	3.50 (0.97)	<0.01
The EMR's design facilitates more timely interventions	2.97 (1.18)	3.21 (1.02)	0.27
The time spent on data gathering through EMR reduces the time spent with the patient/family.	3.66 (1.15)	3.35 (1.13)	0.16
The ICU team discussion is encouraged by EMR's design	2.78 (0.91)	3.08 (0.86)	0.09

Table: Responses of 233 providers to individual survey questions. All questions were on a 5 point scale, where 1 was "strongly disagree" and 5 "strongly agree." Mean scores are presented with the standard deviation in parentheses. "Prescribing" providers were physicians, nurse practitioners and physician assistants. Non-prescribing providers were registered nurses and respiratory therapists. P values calculated using t- test.

Temporal patterns of alerts generated by a medication order-auditing program

Kevin O'Bryan MD[1], Charles H Andrus, MHA[1], S. Paul Hmiel MD, PhD[2], Phillip Asaro MD[2], Feliciano Yu, MD, MSPH[2,1] St. Louis Children's Hospital, St. Louis MO; [2]Washington University School of Medicine, St. Louis MO

Abstract

Pediatric drug dosing rules, embedded in computerized clinician order entry systems, may provide an additional mechanism to avoid adverse drug events. Alerts generated by a custom dose-range checking and auditing system were evaluated by medication, as well as by day of week (weekday vs. weekend) and shift (morning, afternoon, evening and early morning). Three medications represented nearly 65% of the total, with distinct increases in the number of alerts on weekends, and during the night and early morning shifts.

Introduction

Clinical decision support (CDS) engines provide a mechanism to provide relevant guidance to clinicians providing complex care. Alerts generated by these systems can provide insight into the complexities of clinical care, potentially identifying weaknesses of current practice, and opportunities for improving CDS. Medication dosing rules for infants and children are special example complex care, as the clinician needs to consider many factors in prescribing high-risk drugs. We previously described a robust dose range checking system that allows for complex medication dosing rules, while simultaneously producing few nuisance alerts. We now describe the temporal pattern of alerts generated by this system.

Methods

St. Louis Children's Hospital is a 275 bed tertiary care academic hospital, a member of the BJC Healthcare System, and affiliated with the Washington University School of Medicine. It includes a 75 bed neonatal intensive care unit, a 26 bed general Pediatric Intensive care Unit and 12 bed Cardiothoracic Intensive care unit, as well as general medical, surgical, oncology, and neuroscience units. Each unit is staffed by pediatric and surgical residents, supervised by subspecialty fellows and full time faculty, with a dedicated clinical pharmacist available for morning rounds, and throughout the day for consultation. A dose range checking and auditing system was developed as an adjunct clinical decision support system to provide more comprehensive information to clinicians, expanding on the vendor supplied dose range checking features of the EMR.
When the rule was applied to an order, it could result in any of 3 alert stop types: warning (user can proceed without a comment), soft stop (user must comment to proceed), hard stop (user cannot proceed). A real-time dashboard assesses alert effectiveness and user response.

Results

During the study period, a total of 2597 alerts were generated from thirty-three active medication-dosing rules. Alerts were most frequent for acetaminophen, representing 48.2% of total alerts, with other medications at <10%. Two high risk medications, morphine and clonidine, represented 9.1% and 6.4% of the alerts.
When analyzed by day of the week (weekend vs. weekday), a disproportionate number of alerts occurred on weekends (Saturday or Sunday), with 80.8% of acetaminophen alerts occurring on weekends, with similar results for morphine (80.1%) and clonidine (80.6%).

Time of day was grouped as morning (0600 to 1200 hr), afternoon (1200 t0 1800 hr), evening (1800 to 2400 hr) and early morning (0000 to 0600 hr), with the greatest number of alerts (1051) generated during the morning, when most medication orders are generated. The absolute number of alerts during the afternoon, evening and early morning, were similar, at 729, 561, and 411, respectively, in parallel with the decreasing number of overall orders during these time periods. The alert rate, expressed per total orders, was markedly higher during the night and early morning periods, however. Acetaminophen was again most common, representing 53.3% of night and 48.4% of early morning alerts, followed by morphine and clonidine.

Conclusion

Three medications, acetaminophen, morphine and clonidine, were responsible for nearly two thirds of the total alerts. A distinct temporal pattern of drug alerts was noted, with increased alerts on weekends, as well as during early morning and nights. This pattern may reflect the availability of the unit based clinical pharmacists in recommending and monitoring medication orders.

An exploratory factor analysis of socio-demographic and contextual factors associated with Dominican women concerned about HIV/AIDS

Michelle Odlum, BSN, MPH, EdD[1] and Suzanne Bakken, RN, PhD[1, 2]
Columbia University School of Nursing[1], Department of Biomedical Informatics[2], New York, NY;

Introduction

The feminization and ethnic diversification of the HIV epidemic, has resulted in a call for the development and evaluation of gender and culture specific HIV prevention strategies for high risk groups including Latinas[1]. The steadily changing demographic profile of the AIDS epidemic, challenges HIV prevention strategies to remain relevant and up-to-date. The long-term goal of this program of research is to develop a culturally-specific, literacy appropriate, smartphone-based, self-management intervention for Dominican women midlife and older. However, little is known about the HIV sexual-risk behaviors and the socio-demographic and contextual factors of Dominican women in all age groups. This preliminary research seeks use the Theory of Gender and Power (TGP) [2] theoretical framework to identify social/societal factors that may influence women's sexual risk behavior and targets for intervention.

Methods

The three constructs of the TGP will be used to guide the characterization of factors that potentially contribute to HIV sexual-risk behaviors. Items from a community-based survey, relevant to the TGP constructs of: 1) Affective influences/social norms; 2) Gender-specific norms and 3) Power and authority were considered for inclusion. Exploratory Factor Analysis (EFA) was conducted to identify latent constructs underlying items categorized under the three constructs[3]. Casewise deletion of missing data was conducted for all included items[4]. The final sample (N=205) was used in the EFA. Factors were extracted and orthogonally rotated using the varimax method. Solutions of two to five factors were examined for the TGP constructs. Factors were extracted based on the examination of eigenvalues, scree plots and latent construct interpretation.

Results

The original 28 items were reduced to 22 items. The EFA measurement model items were identified and two latent constructs were derived for each theoretical construct[3]. Affective influences/social norms: were characterized by Internet use and spending time with others. Gender-specific norms: were characterized by: food security, education and employment. Power and authority: was characterized by perceptions of perseverance and control (Table 1).

Table 1. Rotated Factor Loading for EFA with Varimax Rotation

Affective influences/social norms	**Factor**	
	Internet Use/ Social Circles	**eHealth seeking/Social Networking**
Item 1: Spend time with some don't live with you	-0.577	-
Item 2: Internet use	0.358	-
Item 3: Household members eHealth information seeking	-	0.556
Item 4: Social networking (Facebook, MySpace, or Twitter)	-	-0.312
Gender-specific norms	**Factor**	
	Food Security	**Education and Employment**
Item 5: Household not eat quality/variety of foods due to money	0.798	-
Item 6: Household not eat due to money	0.688	-
Item 7: Education	-	0.689
Item 8: Working for pay	-	-0.314
Power and authority	**Factor**	
	Perseverance	**Control**
Item 9: Doing things that other people thought could not be done	0.657	-
Item 10: Hard work has really helped me to get ahead in life	0.636	-
Item 11: Anything is going to be done right, I have to do it myself	0.633	-
Item 12: Important to do things in the way I want to do them	0.620	-
Item 13: Stand up for what I believe in, regardless of the consequences	0.594	-
Item 14: Stay with it until the job is completely done	0.539	-
Item 15: Don't let personal feelings get in the way of getting a job done	0.460	-
Item 16: Could make life pretty much what I wanted to	0.447	-
Item 17: Confident about ability to handle your personal problems	-	0.839
Item 18: Effectively coping with important changes in life	-	0.792
Item 19: Things were going your way	-	0.472
Item 20: Upset because of something that happened unexpectedly	-	-0.32
Item 21: Felt unable to control the important things	-	0.309

Only factors loading at ¦ = .3¦ are presented

Conclusion

The results of the EFA suggested a factorial structure for constructs of TGP. The latent constructs identified give insight into community-level factors to inform the content development for our smartphone-based, self-management intervention for HIV risk reduction.

Acknowledgement: This study is supported by WICER (R01 HS019853) . Dr. Odlum is supported by the Provost's Postdoctoral Scientist Fellowship.

References

1. Peragallo N, Gonzalez-Guarda RM, et al. The efficacy of an HIV risk reduction intervention for Hispanic women. AIDS and behavior. 2012;
2. Uhrig JD DK, et al. Behavioural precursors and HIV testing behaviour among African American women Health Ed 2010. 71:1-13.
3. Ruscio J and Roche B. Determining the number of factors to retain in an EFA. Psych Ass 24(2),282-292.
4. Yen P, Sousa KH, Bakken S. Examining construct and predictive validity of the Health-IT Usability Evaluation Scale JAMIA 2014 Feb 24

Clarifying requirements chronological visualization of medical data by investigating medical journal articles

Keisuke Ogawa, MS[1], Kazunori Matsumoto, MS, PhD[1], Masayuki Hashimoto, MS, PhD[1], Akiko Shibuya, RN, PHN, PhD[2], Yoshiaki Kondo , MD, PhD[2]
[1]KDDI R&D Labs, Saitama, Japan; [2]Nihon University School of Medicine, Tokyo, Japan

Abstract

To clarify the requirements for the chronological visualization of medical data, we performed a literature search of medical journal articles that presented data visualization methods (literature review). We found that the chronological visualization of medical data varied according to the type of research being conducted. In particular, the forms of visualization and time spans in animal experiments differed significantly from those in other types of research. By conducting this investigation, we were able to clarify the differences in the requirements for each type of research.

Introduction

An increasing amount of data is being stored in medical information systems. In hospital settings, the use of laboratory tests is increasing considerably[1]. In order to support the handling of such *medical big data*, medical data visualization systems have been proposed for a range of medical situations such as diagnosis, clinical trials, and medical research[2, 3, 4]. However, the scope of possible requirements for the representation of medical data is not yet clear[5], making it difficult to establish the appropriate representation capabilities for a chronologically based visualization system.

Methods and Results

We analyzed the potential requirements for medical data visualization systems by investigating the data visualization methods presented in medical journal articles (medical graphs). In this paper, we categorize the articles into three types of medical research[6] (cohort studies, animal experiments, and in-vitro experiments) and examine how medical data are displayed in chronological graphs and the associated time spans. We decided to focus our investigation on articles about chronic diseases, especially diabetes mellitus, because the treatment duration tends to be long in such cases. For this study, we selected 230 graphs from 44 articles in the top three diabetes-related journals (that have "diab" in the title and high impact factors) published in January 2012. The investigation results are shown in Tables 1 and 2.

Table 1: Data visualization methods(graph count)

AIC

	Flow chart	Polyline	Vertical bar	Horizontal bar	Table	Scatter plot	Total
Animal exp.	11	69	16	5	13	1	115
†1 Cohort study	3	24	1	0	22	0	50
†2 In-vitro exp.	2	16	12	0	0	0	30
†3 Misc(6 kinds)	9	26	0	0	0	0	35
Total	25	135	29	5	35	1	230

Table 2. Average time span (graph count)*

F-test

	~1 hour	~1 day	~1 week	~1 month	~1 year	~1 century	Variance of time spans(normalized)
Animal exp.	10	57	4	4	22	0	0.08
‡1 Cohort study	3	0	0	1	0	43	0.03
‡2 In-vitro exp.	1	7	14	1	0	0	0.04

*Only those graphs having explicit time spans are counted in Table2

From Tables 1 and 2, it can be seen that there are significant variations in the ways these graphs are presented and the time span in *animal experiments* varies far more than in the other two types of research. The differences are statistically significant. We performed a cross-table analysis using the Akaike Information Criteria[7] and performed an F-test on the time-span variances; AIC differences between dependent and independent models are -98.9(† 1 in Table1), -74.5(† 2 in Table1), and -69.6(† 3 in Table1). P-values are 0.001 (‡ 1 in Table2) and 0.02(‡ 2 in Table2). We concluded that the requirements for chronological visualization of medical data vary according to the type of research. In particular, it was found that representation varies depending on the type of research and that the time spans adopted in *animal experiment* are significantly different from those in cohort and in vitro studies. Satisfying such diverse needs will require systems that allow the user to present and examine data using a wide range of time spans.

Conclusion

This study clarified that the requirements for the chronological visualization of medical data vary according to the type of research. In particular, chronological representations differ depending on the research category and the time spans in animal experiments are significantly different from those in other types of medical research.

References

1. M.Mindemark et.al., Longitudinal trends in laboratory test utilization at a large tertiary care university hospital in Sweden., Uppsala Journal of Medical Sciences. 2011; 116:34-38.
2. Alex A.T.Bui et al., TimeLine: Visualizing Integrated Patient Records, IEEE TITB. 2007; 11:462-473.
3. Y.Kondo, A new proposal of an active object-oriented project management system for electronic health records, AMIA Annual Symposium Proc. 2008.
4. K.Ogawa et al., Mobile TimeLine EMR system, 5th BIOSTEC. 2012; HEALTHINF:23-29.
5. A.Shibuya et al., Usability of "Problem-oriented Contiguous Timeline View (PCTLV)" using a mobile device for clinical knowledge management, 32nd JCMI. 2012: 360-361.
6. Stephen B. Hulley et al., Designing Clinical Research, 3rd Edition Lippincott Williams & Wilkins. 2004.
7. Yoshiyuki Sakamoto et al., Johoryo Tokeigaku, Kyoritsu Shuppan Co., Ltd. Tokyo. 1983:92-106

Perceptions of Health Care Quality in an Emergency Department During a Planned Electronic Health Record Downtime

Author Block *Amit M. Mehta, Nnaemeka G. Okafor*. University of Texas - Health Science Center at Houston, Houston, TX

Abstract:
Study Objectives:
To evaluate the impact of an Electronic Health Record (EHR) system downtime on health care quality in an emergency department (ED). Health care quality is defined by The Institute of Medicine using six aims: safe, effective, timely, efficient, patient centered and equitable. As more hospitals implement EHR systems, there is a need to understand how those six aims are affected when the EHR systems are unavailable.
Methods:
The project was conducted in an academic ED with 70,000 annual visits and a commercial enterprise-wide EHR system. The ED Informatics Division created a Downtime Summary questionnaire aimed at understanding the clinician's perception of the health care quality issues during the planned 12-hour EHR downtime. The standard EHR downtime protocol included patient care documentation, blank ordering forms, and a whiteboard in each ED care area for tracking patients, diagnostic testing and clinician assignment status. For this project, all ED technicians, nurses, and clinicians were instructed to document any perceived safety (harm to the patient), timeliness (delays in care), effectiveness (missed evidence-based interventions) or inefficient processes attributable to the downtime on the questionnaire attached to every patient chart. The questionnaire forms were then collected from each patient chart.
Results:
A total of 80 patients were evaluated in the ED during the EHR downtime, only 69 (86%) questionnaires were retrieved with any information. 52% of the patients evaluated were discharged, 4% were hospitalized, 39% were still in process, and the remaining 3% were transferred, eloped or left against medical advice. There were no reported issues related to patient safety. The only reported quality issue related to effectiveness was the lack of utilization of approved order sets. Quality issues related to timeliness centered on delays in provider notification of imaging and laboratory results often leading to multiple calls by ED providers to determine the status of these tests. The vast majority of quality issues were regarding inefficient processes. The reported inefficient processes included: difficulty locating newly roomed patients and their respective charts, difficulty finding the appropriate forms to place orders or complete patient care dispositions, problems accessing radiology images or reports and inefficient patient, diagnostic testing and clinician assignment status tracking using the boards. The boards frequently contained incomplete information because they could not be updated in real-time and had a limited physical size. Providers did not review the boards regularly because of their stationary location and limited information.
Conclusion:
The impaired ability to track patient status and diagnostic data, which is typically provided by the EHR track board, resulted in inefficient processes due to the decreased situational awareness of the providers. Development of methods to remotely track these elements during a downtime is an important aspect of maintaining health care quality in an ED.

Decomposition of Quality Requirements to Evaluate Electronic Health Records Systems

Marília E. S. Oliveira, Ms[1, 3], Magdala A. Novaes, PhD[2, 3], Alexandre M. L. Vasconcelos, PhD,[1]

[1]Center of Computer Science of Federal University of Pernambuco, Recife, PE, Brazil;
[2]Internal Medicine Department, Federal University of Pernambuco, Recife, PE, Brazil;
[3]Telehealth Center, Clinics Hospital, Federal University of Pernambuco, Recife, PE, Brazil;

Abstract

Quality evaluation of Electronic Health Record Systems (EHR-S) is an important point of discussion in relation to implantation and acceptance of this kind of system. This work presents the application of a requirements decompose method to evaluation of EHR-S in relation to requirements that is scope of Brazilian Certification for Electronic Health Record Systems – SBIS-CFM Certification. A case study has shown, like initial results, this method may contribute to EHR-S evaluation in relation to SBIS-CFM Certification.

Introduction

Technical standards and certifications have been created in order to establish the minimum requirements expected for EHR-S. On the other hand, EHR-S suppliers should to introduce in their process development an evaluation step for check their products systems according to these requirements. [1] However in some case, these requirements are not expressly verifiable and may bring different interpretations to the reader. This characteristic may become more difficult conduce the evaluation process.

This work presents the initial results of a case study that has evaluates an EHR-S in relation to requirements of SBIS-CFM Certification [2] using a method for decompose quality requirements in minor quality attributes. [3] The aims were analyze if this strategy would help EHR-S suppliers teams to clarify the understanding about requirements and audit test scripts when evaluate their software products, before submitted its on official certification process.

Methods and Results

This work has used official test scripts of Brazilian Certification for Electronic Health Record Systems, SBIS-CFM Certification, and has produced a new evaluation test suit adding a set of help questions to each test script. Test scripts have been decomposed in a set of minor checkpoints that include examples and test data options. Objective of the checkpoints was become more easy identify if the respective requirement was met or not.

Case Study was executed at Telehealth Center of Federal University of Pernambuco and its objective was verify if proposed evaluation method may help Telehealth center team to conduce a internal evaluation in its EHR-S in relation to SBIS-CFM Certification. For each test script, at first moment, participants had to try to execute only official script and after that, they could use respective help questions. During study, it was possible note that participants had to use the help questions to solve their doubts in 39% of total test scripts.

Conclusion

Decompose test scripts in help questions may help the developer team conduce previous evaluations in their EHR-S in relation to SBIS-CFM Certification requirements.

References

1. Smith B, Austin A, Brown M, King JT, Lankford J, Meneely A, and Williams L. Challenges for protecting the privacy of health information: required certification can leave common vulnerabilities undetected. In SPIMACS '10, 2010 ,pp. 1–1
2. SBIS-CFM - Certification Manual for Electronic Health Records Systems (EHR-S), Public Consult, 2011, www.sbis.org.br, accessed at december 2012.
3. ISO/IEC 25040, Software engineering - Software product Quality Requirements and Evaluation (SQuaRE) – Evaluation reference model and guide, ISO/IEC JTC1/SC7/WG6, 2011

Employment of Penalized Models to Predict Cognitive Domain Performance Using Proteomics

Shauna M. Overgaard, MHI, S.Charles Schulz MD, Gyorgy Simon, PhD
University of Minnesota, Minneapolis, MN

Abstract

Detection and quantification of protein levels may lead to the differentiation of illness-type and the development of diagnostic tests to earlier identify vulnerable patients. Our objectives are to i) employ ridge regression analysis and adaptive lasso in the identification of biomarkers associated with cognitive deficit, and to ii) test the ability of the identified biomarkers to differentiate illness severity groups. Distinct biomarkers are associated with validated deficits shown to differentiate illness groups.

Introduction

Challenges in psychiatric medicine have motivated endophenotype identification and integration of dimensional approaches to diagnosis. Detection and quantification of protein levels may lead to the differentiation of illness-type and development of diagnostic tests to identify vulnerable patients prior to frank symptom manifestation. Schizophrenia is marked by a decrease in cognitive performance[1]. The MATRICS[2] Consensus Cognitive Battery (MCCB) is employed to test seven cognitive domains demonstrated to be effected by schizophrenia: Speed of processing, Attention/Vigilance, Working memory, Verbal learning, Visual learning, reasoning and problem solving, and social cognition. The primary objectives of the analysis are to i) identify proteomic (analyte) predictors that are associated with deficits in the seven cognitive domains and to ii) quantify their ability to distinguish between illness severity groups (control, prodrome, schizophrenia).

Materials and Methods

To achieve our first aim of identifying analyte predictors for each standardized MCCB domain score, we employed adaptive lasso regression. To evaluate whether the resultant analytes can better differentiate between the three disease severity groups than analytes that are not necessarily associated with a cognitive domain deficit, we built a baseline ridge[3] regression model that predicted disease severity based on all analytes. We then built a "differential" regression model to predict the disease severity group based on the analytes identified in our first aim. This was achieved by including the prediction of the baseline ridge model as offset. Analytes that are significant in the second (differential) model can distinguish between the severity groups better than the analytes that are not necessarily associated with a cognitive domain.

Results

Cognitive domain scores known to predict schizophrenia yield predictive value for prodrome classification. Handedly, the levels of analytes that are predictive of cognitive domain are significantly different between clinical groups. The resulting predictions of the ridge regression analysis using MCCB domain scores yielded a receiver operating characteristic area under the curve (ROC-AUC) of .80 for classification of schizophrenia, .79 for Prodrome and .50 for controls. The resulting predictions for the ridge regression using protein analyte predictions of diagnostic class yielded a receiver operating characteristic area under the curve (ROC-AUC) of .78 for classification of schizophrenia, .76 for Prodrome and .64 for controls. Using adaptive lasso, protein phenotypes that appear to support a deficit in working memory and verbal memory are identified. Distinct biomarkers are associated with validated cognitive domains shown to differentiate illness groups. The data contribute to a foundation for a clinical model integrating biomarkers and behavioral indicators associated with early identification and intervention of schizophrenia.

References

1. Seidman, L., Giuliano, A., Meyer, E., Addington, J., Cadenhead, K., Cannon, T., McGlashan, T., & Perkins, D. (2010). Neuropsychology of the prodrome to psychosis in the NAPLS consortium. *Archives of General Psychiatry*, *67*(6), 578-588.
2. Nuechterlein, K.H., Green, M.F., (2006). MATRICS Consensus Battery Manual [MCCB]. MATRICS Assessment Inc., Los Angeles.
3. Hoerl, A. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 1970. 12(1), 55-57

DIGITILIZATION OF THE MULTIDISCIPLINARY MEDICAL INTENSIVE CARE UNIT CHECKLIST

Latifat A. Oyekola, MD [1,2], John T. Finnell, MD , MS[1,2]
[1]Indiana University School of Medicine, Indianapolis, IN; [2]Regenstrief Institute Indianapolis, IN

Abstract With the wide adoption of the electronic health record systems, it is imperative to incorporate existing paper based checklists in medical practice as part of the electronic health record. It is important that checklist is self-developed, precise and operationally suited.

Wishard-Eskenazi Health Hospital in Indianapolis, Indiana is in the process of automating the MICU checklist.

Introduction and Background Checklists can be effective tool to improve care processes and lead to better outcome. Checklists are common in some medical fields like surgery, intensive care unit and emergency medicine. The implementation of multidisciplinary daily rounding checklist in the intensive care units has been shown to reduce mean ventilation-associated pneumonia, reduce mean catheter-related blood stream infection and in the process improve the quality of care while also reducing cost.

Method The MICU team indicated their interest in automating the multidisciplinary paper-based checklist which they call the 'Orange Sheet". The Clinical Informatics team analyzed the workflow of the MICU team after which a meeting with the stakeholders (MICU team) was convened to clarify terms used in the existing paper form and define the functional requirements of the software.

Results Some of the requirements desired were the ability of the software to interact with the users, reminder and prompting ability, messaging features, pre-populate antibiotics, hyperlink to evidence –based practices, leveraging this in multiple spaces (for example order interface) and it's use as part of nursing care plan.

Conclusion The first meeting between the developers of the electronic medical record system (the New Gopher) raised questions about some of the functional requirements specified which could not be immediately addressed as the developers brainstorm on the design for the project. This project is ongoing and is expected to follow the sequence outlined in the diagram.

Differences in Occurrence and Recorded Times of Care Delivery Events as Documented in Electronic Health Records

Mustafa Ozkaynak, PhD[1], Oliwier Dziadkowiec, PhD[1], Sara Deakyne, MPH[2], Tiffany Callahan, BA[1], Eric Tham, MD, MS[1,2]
[1]University of Colorado, Aurora, CO; [2]Children's Hospital Colorado, Aurora, CO

Abstract

Electronic Health Record (EHR) systems can serve as rich data sources to examine clinical workflow. We extracted event logs for every encounter that took place in an academic hospital emergency department and its satellite urgent care sites in 2013. We identified delays between the occurrence and the recorded times of care delivery activities. Workflow researchers should consider these delays to ensure event sequences are captured accurately when using EHR systems as a data source.

Introduction

Electronic Health Record (EHR) systems automatically capture massive numbers of clinical and non-clinical data related to patient care. Although EHRs were designed and are primarily used for patient-care purposes, secondary use of these data can potentially provide with workflow-characterization insights[1]. However, meticulous use of such data source is required for a valid data analysis and interpretation. The purpose of this study is to examine the differences between time of activities and their documentation time of these activities as recorded in the EHRs. Understanding these time differences will inform the reliable and systematic use of EHRs for workflow analysis.

Methods

We extracted event logs for each encounter that took place in the emergency department (ED) of an academic hospital and its satellite urgent care sites in 2013. Our analysis included a total of 2,815,568 events from 133,586 visits by 88,936 patients. Data extraction included the following data elements: activity type, activity occurrence and activity recording times. Activity occurrence is the time that is displayed in the EHR as the time the event occurred, which may be changed manually by a staff member for some events. Activity recording time is the timestamp in the system when the information was recorded.

Results

We identified 966,821 events (34.3% of all recorded events) that were recorded after they occurred. The most frequent five events that recorded after their occurrences are given in Table 1.

Table 1. The five most frequent events recorded after their occurrences (Sorted by number of delayed recordings).

Event Type (number of delayed recordings)	Average Delay (minutes)	Median Delay (minutes)	% of Events with delay
Charting Complete (225,651)	0.51	0.50	81.6%
Patient departed (134,388)	34.58	9.77	99.8%
Patient arrived in ED (134,377)	2.59	1.97	99.8%
Patient discharged (123,530)	33.90	11.43	99.4%
Nurse/tech Assigned (103,696)	0.50	0.50	59.3%

Discussion

Results indicated delays in recording majority of the five frequent ED activities in the EHR by the computers and clinicians. Delayed recording can be due to fast pace of work in EDs, in which treatment outweighs documentation at the time it occurs. Researchers should consider these delays to ensure that event sequences are captured accurately when using EHR systems as a data source. Supporting EHR data with radio-frequency identification technology, which captures real-time data on clinicians' activities, can enhance the quality of the data for workflow research.

References

1. Vankipuram M, Kahol K, Cohen T, Patel VL. Toward automated workflow analysis and visualization in clinical environments. J Biomed Inform. 2011 Jun;44(3):432–40.

Automating Extraction and Calculation of Daily Dose and Duration for Medications in EHRs

Jennifer A. Pacheco[1], William K. Thompson, PhD[1], Kathryn L. Jackson, MS[1], Abel N. Kho MD, MS[1]

[1]Feinberg School of Medicine, Northwestern University, Chicago, IL

Abstract

As part of the eMERGE (electronic Medical Records and Genomics) project, we are engaged in studies that require the extraction of medication data from EHRs (Electronic Health Records). In order to calculate daily dose and duration of medications, we found it necessary to supplement the existing NLP (Natural Language Processing) tool we are using with additional techniques. The result is an open-source workflow that can be readily re-targeted to multiple medications and is 90% accurate.

Introduction

Increasingly in clinical research, and especially in pharmacogenetic studies[1], it is important to extract and calculate numerical dose and duration for medication usage from EHRs. Current NLP applications such as MedEx produce results which are often not numeric nor measured in consistent units. At least one study has extracted such data by extending and supplementing tools such as MedEx, but only for 1 specific medication[1].

Methods and Materials

We created an executable workflow in KNIME (Konstanz Information Miner, knime.org) for calculating average daily dose and duration for medications, which extracts the numerical data needed for the calculations in 3 major steps: 1) combining the results of the MedEx NLP tool on prescription order signature line text with discrete medication data from our EHR where available, both of which are mostly non-numeric; 2) using pattern matching (RegEx), Java, and other data transformation tasks in KNIME, to convert dose and duration into numeric fields in standardized units; and 3) applying logical rules, for duration in particular, prioritizing fields to determine an accurate daily dose and duration.

We tested our algorithm on statins and metformin, from which a random representative subset of 20 patients' 100 medications records were selected for manual review in order to compute accuracy statistics. The review was done by reviewing medication records, discharge summaries, and encounter notes, in both inpatient and outpatient EHR systems (Cerner for inpatient and Epic for outpatient), by JAP, who is an experienced biomedical informaticist.

Results

Ninety percent of the medication records from the KNIME workflow were accurate when compared to the medication records found during manual review. Many if not most of the errors were due to the NLP algorithm not finding a given data point needed to calculate daily dose or duration, i.e. missing data such as 22% of dose frequency missing. We are making our KNIME workflow publicly and freely available at http://bit.ly/1nowjqt.

Discussion and Conclusion

This work builds on existing work but expands to include multiple medications and is tested on 2 different commercially available EHR systems. Future work includes expanding to more types of medications, i.e. more than just orally administered medications measured in mg, and using algorithms to fill in missing data (with most frequently prescribed dose frequency for a given medication for example).

Acknowledgement. This work has been supported by funding for the eMERGE Network which was initiated and funded by NHGRI in part through the following grant: U01HG006388 (Northwestern University).

References

1. Xu H, Jiang M, Oetjens M, Bowton EA, Ramirez AH, Jeff JM, Basford MA, Pulley JM, Cowan JD, Wang X, Ritchie MD, Masys DR, Roden DM, Crawford DC, Denny JC. Facilitating pharmacogenetic studies using electronic health records and natural-language processing: a case study of warfarin. J Am Med Inform Assoc. 2011;18:387-91

Problem-Solving Methods for Public Health Informatics Practice and Training: Insights from Technical Assistance Projects

Sridhar R. Papagari Sangareddy, MS[1], Laura Franzke, MPH, PhD[1], Herman Tolentino, MD[1]

[1] Centers for Disease Control and Prevention, Atlanta, GA

Abstract

A core component of the Centers for Disease Control and Prevention's Public Health Informatics Fellowship Program (PHIFP) is the Info-Aid — a short-term technical assistance project that PHIFP fellows lead to solve public health informatics (PHI) problems for public health agencies. Info-Aids are used in a range of PHI problems and require different problem-solving methods (PSMs). A taxonomy for classifying PSMs for PHI is presented that is based on a problem-solving framework and systems concepts.

Description

PHI is the systematic application of the knowledge of systems that capture, manage, analyze, and use information for improving population health. PHI practice involves problem solving, which requires the systematic application of PSMs. PHI PSMs are not well-documented in literature. In Info-Aids, PHIFP fellows are engaged in real-world PHI problem solving for various PHI problems. To solve them, fellows adapt and adopt PSMs from referent disciplines (e.g., computer science, information science, or the social sciences). Using lessons learned from Info-Aids, we analyzed and then synthesized these methods into a taxonomy (Table 1) that can be the basis of reasoning steps and knowledge types needed to solve problems.[1] This taxonomy can serve as a resource for fellows to classify and organize PSMs they need to apply in a given Info-Aid; the taxonomy also can be useful for other PHI training programs. The taxonomy can be further refined and validated to identify PHI problems in practice.

Table 1. A Preview of Taxonomy of PSMs Based on Information System Structure, Function, and Process

Problem-solving method		SWOT[b] analysis	Data modeling	Record linkage
Problem space related to public health information systems[a]	**Structure**	Organization	Data	Information
	Function	Assess	Capture	Manage
	Process	Planning	Design	Development
Problem-solving goal		Strategic planning	Database design	Integration of data from different sources

[a] *Structure* defines dimensions of an information system, as represented by Richard Heeks' Onion Ring model (Source: Heeks R, Bathnagar S. Understanding success and failure in information age reform. In: Heeks R, ed. Reinventing Government in the Information Age: International Practice in IT-Enabled Public Sector Reform. Abingdon, Oxon: Routledge; 1999: 49–74.); *function* defines the steps in generating value from an information system (Source: Taylor RS. Value-added processes in the information life cycle. Journal of the American Society for Information Science 1982:33:341–6); and *process* explicitly defines the phases of an information system implementation.

[b] SWOT = strengths, weaknesses, opportunities, and threats.

References

1. Fensel D, Motta E, Decker S, Zdrahal Z. Using ontologies for defining tasks, problem-solving methods, and their mappings. In: Plaza E, Benjamins VR, eds. Knowledge Acquisition, Modeling, and Management. Berlin, Germany: Springer-Verlag. Lecture Notes in Computer Science 1997;1319:113–28.

Network Analysis of Common Eligibility Criteria in Cancer Clinical Trials

Chin S. Park[1], MBA, RN; Jacqueline A. Merrill[1, 2], PhD, MPH, RN; Chunhua Weng[2], PhD
[1]School of Nursing, [2]Department of Biomedical Informatics
Columbia University, New York, NY

Abstract

To understand the patterns in eligibility criteria across different cancer types, we performed a network analysis of frequent eligibility criteria for 124 cancer types available on ClinicalTrials.gov. Free text eligibility criteria were extracted using natural language processing. The results indicate that substantial eligibility criteria were shared across various cancer clinical trials. The findings have implications for knowledge reuse in clinical trial designs.

Introduction

Eligibility criteria are specifications written by clinical researchers regarding who would qualify for a clinical study. We hypothesize that a study of eligibility criteria patterns in the cancer clinical trial research community can help us identify systematic knowledge for subject selection for cancer research. This study aims to identify which patient characteristics are frequently shared among cancer clinical trials, within and across cancer types.

Methods

Cancer clinical trial summaries for 124 cancer types archived between years 1999-2013 were extracted from ClinicalTrials.gov. An unsupervised text mining method was utilized to identify frequent eligibility criteria from all these studies [1]. Each eligibility criterion had information about its prevalence, e.g., 50% indicates that a criterion appeared in 50% of all cancer trials of a selected cancer type. The frequency information was converted to a corresponding rank, as shown in Table 1. Network analysis and visualizations were accomplished using ORA-NetScenes. The nodes indicate cancer types (green) or frequent eligibility criteria (red), which were eligibility concepts such as "anemia" or "concurrent malignant neoplasms". The links indicate the associations between cancer types and eligibility concepts.

Table 1. Common Eligibility Criteria (CEC) frequency of appearance converted to corresponding rank.

CEC Frequency	Rank
90% - 100%	10
80% - 89%	9
70% - 79%	8
60% - 69%	7
50% - 59%	6
40% - 49%	5
30% - 39%	4
20% - 29%	3
10% - 19%	2
< 10%	1

Figure 1A. Network of 124 cancer types linked to 11,195 CECs.

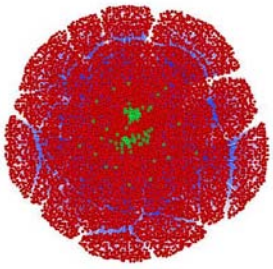
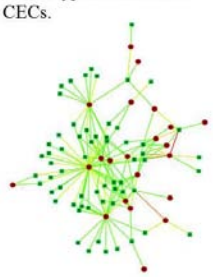


Figure 1B. Network of 62 cancer types linked to 22 CECs.



Results

We identified 11,195 unique frequent eligibility criteria, and the resulting network presented 81,145 links out of 1,388,180 possible links, indicating a network density of 5.8% (Figure 1A). When the criteria were filtered by those only ranked 6 or higher, only 22 were linked to 62 cancer types (Figure 1B). Degree centrality measures were calculated for the complete network of 124 cancer types and 11,195 criteria. The cancer node with the highest in-degree centrality (number of links from other criteria) was "leukemia erythroblastic acute," and the criteria node with the highest out-degree centrality (number of links to other criteria) was "therapeutic radiology procedure." The top 25 nodes for both in-degree and out-degree centrality were identified and listed in decreasing order by rank.

Discussion

The centrality measurements indicated that criteria such as "therapeutic radiology procedure" and "disease characteristic" are nodes with high out degree centrality and therefore linked to more cancer types than the other criteria. This concurs with published cancer research literature and demonstrates face validity of the results. Also, this network analysis found only 22 of 11,195 unique criteria were linked to half of 124 cancer types when looking for criteria that appear in 50% or more of the research trials. The patterns revealed by the network analysis suggest that certain criteria are common links across different types of cancers. Continued work in identifying central criteria of eligibility criteria may help indicate the types of patient information that are important to collect in electronic health record (EHR) systems in order to facilitate electronic eligibility screening for cancer clinical trials.

References

[1] Miotto R, Weng C, Unsupervised Mining of Frequent Tags for Efficient Clinical Eligibility Text Indexing, J Biomed Inform, 2013 Dec; 46(6):1145-51

Clinical Data Modeling EHR Data for Understanding Factors Related to Hospital-acquired Catheter-Associated Urinary Tract Infection (CAUTI)

Jung In Park, BS, RN[1]; Steven G. Johnson, MS[2]; Matthew D. Byrne, PhD, RN[3]; Beverly Christie, DNP, RN[4]; Lisiane Pruinelli, MS, RN; Suzan Sherman, PhD, RN[4]; Bonnie L. Westra, PhD, RN, FAAN, FACMI[1]

[1]University of Minnesota, School of Nursing; [2]University of Minnesota, Institute for Health Informatics; [3]St. Catherine's University; [4]Fairview Health System

Problem

The purpose of this study was to create a clinical data model that includes normalized and standardized data from an academic health center clinical data repository (CDR) as a preliminary step for knowledge discovery research planned in a subsequent study.

Introduction and Background

Catheter-Associated Urinary Tract Infections (CAUTIs) are the most common type of infection in hospital, accounting for 40% of all nosocomial infections per year. The Centers for Medicare and Medicaid Services (CMS) announced hospital acquired Urinary Tract Infections (UTIs) as one of eight conditions for which hospitals will not receive additional reimbursement. Studies identifying risk factors and evidence-based guidelines for CAUTIs exist, but the occurrence rates are still high. There is a clear need for further understanding of either additional factors and/or subgroups of patients to understand predictors associated with hospital-acquired CAUTI to reduce the occurrence rates. With the implementation of electronic health records (EHRs), it becomes possible to conduct research using large data sets to gain new insights into which patients benefit the most from different approaches to prevention of CAUTI.

Methods and Data Source

The University of Minnesota developed a CDR which includes more than two million patient and multiple data types (demographics, encounters, diagnoses, procedures, medications, and lab values) for ambulatory and hospital data. Additionally it includes flowsheet data which contain documentation of essential data for assessment and prevention of CAUTI. A data set from the CDR was evaluated and compared with the literature to create a data model to discover patient characteristics and interventions by subgroups of patients that can provide new insights to reduce infections.

Results/ Discussion

A clinical data model from EHR data for CAUTI will be presented. It was created from patients' health records in the CDR for patients hospitalized between 10/1/10 and 12/31/2013. Data normalization and standardization will be discussed. In a future study, data mining techniques will be employed to discover patterns of patient characteristics and provider interventions within subgroups of patients to gain new insights for reducing the rates of hospital-acquired CAUTI.

Acknowledgment

"This was supported by Grant Number 1UL1RR033183 from the National Center for Research Resources (NCRR) of the National Institutes of Health (NIH) to the University of Minnesota Clinical and Translational Science Institute (CTSI). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the CTSI or the NIH. The University of Minnesota CTSI is part of a national Clinical and Translational Science Award (CTSA) consortium created to accelerate laboratory discoveries into treatments for patients."

Automatic Engine for Mapping Mycobacteriology Reports to SNOMED-CT

Olga V. Patterson, PhD[1,2], Scott D. Nelson, PharmD[1,2], Makoto M. Jones, MD[1,2], Kimberly Findley, RN[3], Kevin L. Winthrop, MD, MPH[4], Kevin P. Fennelly, MD, MPH[3,5], Scott L. DuVall, PhD[1,2]

[1] VA Salt Lake City Health Care System; [2]University of Utah, Salt Lake City, UT; [3]Office of Public Health, Veterans Health Administration, Gainesville, FL; [4]Oregon Health and Science University, Portland, OR; [5]University of Florida, Gainesville, FL

Background

The genus *Mycobacterium* contains a number of pathogens of public health significance that cause infections ranging from tuberculosis and leprosy to debilitating infections in individuals with structural lung disease. Since administrative codes underreport mycobacterial infections, coded microbiology data are necessary to gain a complete picture of the burden of disease. Detection of mycobacteria is performed in the laboratory using a combination of culture and molecular methods. The generation of coded data from microbiology reports in large population, such as in the US Department of Veterans Affairs (VA), requires the development of an innovative strategy for data extraction. Previous research efforts involved contacting individual laboratories, collecting data directly from sources, and aggregating results across multiple laboratories.[1] This approach is time intensive and, without incorporating other information from the medical record, only allows analysis of trends and distribution of positive tests. Generating coded microbiology is challenging because of insufficient interoperability of laboratory systems and clinical patient record, especially in cases when an outside lab performs mycobacterial cultures. Typically, the test results are passed to the requesting organization as text reports. We developed a natural language processing (NLP) system to identify mentions of mycobacteria in lab reports, map the mentioned organisms to SNOMED-CT codes, and extract test outcomes.

Methods

A corpus of 528,083 laboratory reports from 128 VA stations from 2008 to 2012 was split into 3 parts – half of the notes for training, a quarter for development testing, and a quarter for final validation. Four clinical annotators manually reviewed 11,200 lab reports randomly selected from the training subset to identify mentions of mycobacteria and relevant test results. A dictionary was developed using these annotations that was manually mapped to appropriate SNOMED-CT codes. Text indicating positive, negative, and unknown test results was transformed into co-occurrence patterns. An NLP system was built using a set of libraries that facilitate rapid creation and deployment of high throughput, parallel NLP systems.[2,3] Term identification, mapping to SNOMED-CT codes, and test outcomes were implemented using regular expressions for each of the targeted organisms. Rules were developed to determine the final test result in cases with multiple mentions of the same organism in the same document.

Results

System performance was evaluated using 5,400 documents selected from the validation subset. Extracted test results from each documents processed by the NLP system were reviewed by two clinical annotators. Document-level classification accuracy was 0.95 for positive test results and 0.942 for negative test results.

Conclusion

Automatic detection of positive laboratory test results using NLP with mapping to a controlled terminology can detect mentions of mycobacteria, and may help with real-time surveillance of Veterans with mycobacterial infections.

Acknowledgements

This work was supported using resources and facilities at the VA Salt Lake City Health Care System with funding support from the VA Center for Occupational Health and Infection Control and the VA Informatics and Computing Infrastructure (VINCI), VA HSR HIR 08-204.

References

1. Cassidy PM, Hedberg K, Saulson A, McNelly E, Winthrop KL. Nontuberculous mycobacterial disease prevalence and risk factors: a changing epidemiology. *Clin Infect Dis*. 2009;49(12):e124–9.

2. DuVall SL, Patterson OP, Ginter T, Cornia R, Nebeker JR. FLAP: Flying Towards Real-Time Decision Support. In: *NIH workshop on Natural Language Processing: State of the Art, Future Directions and Applications for Enhancing Clinical Decision-Making*. Bethesda, MD; 2012.

3. Ferrucci D, Lally A. UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Nat Lang Eng*. 2004;10(3-4):327–348.

Kidney transplantation web portal for the citizens in Brazil

Alissa Peres Penteado, MSc Fellow[a], Frederico Molina Cohrs, MSc, Rafael Fábio Maciel, MD MSc[b], Bartira de Aguiar Roza, PhD Associate Professor[a], Cristina Lúcia Feijó Ortolani, PhD Full Professor[c,a], Ivan Torres Pisa, PhD Associate Professor[a]

[a] *Universidade Federal de São Paulo UNIFESP, Brazil* [b] *Instituto Social de Assistência a Saúde, Caruaru, Brazil* [c] *Universidade Paulista UNIP, São Paulo, Brazil*

Abstract

The kidney transplantation process in Brazil is based on laws, decrees, ordinances, and resolutions, and it's unique for all the country. It's a complex process for non-specialized healthcare professionals, non-specialized healthcare managers, and citizens. This paper presents our effort to model this organizational process using business process modeling notation considering it a simple language and to model a web portal to support this process, legislation repository, and news about transplants in Brazil.

Introduction

The kidney transplantation process in Brazil is defined through laws, decrees, ordinances, and resolutions (1), but there is no defined theoretical map describing the entire process. Business process modeling notation (BPMN) (2) has been used to represent actions in organizations, aiming to increase efficiency, quality, and maturity (3). Taking it into consideration we decided to create a diagram that can represent the flow of this entire process. In order to make this process public available for the Brazilian citizens we will design a web portal including graph representation, a legislation repository, and news about transplant in an easy and dynamic way. This study took into consideration just the kidney transplantation, because Brazil has good global indices in this type of transplant in absolute numbers (4). But when it comes per million population (pmp) rate, Brazil is only the 33th in the world rankings. This is the reason why we decided to analyze this process and try to describe why this process is too complex, its barriers and modification possibilities.

Methods

For the creation of the kidney transplantation process the approach taken was an exploratory observational study with analysis on documents and processes (5). We identified and gathered official digital documents (laws, decrees, and ordinances) relating to the kidney transplantation process in Brazil (26 regional states plus Distrito Federal) and we analyzed them using process analysis methodology (5). Then a kidney transplantation process in Brazil based on BPMN was created using Bizagi version 2.4.0.8 (bizagi.com). We conduct 2 different evaluation tasks to ensure that the process was properly represented. The first evaluation was using Delphi methodology with 4 specialists in kidney transplantation field. The second evaluation was a survey application with 27 specialists (one from each state from Brazil plus Distrito Federal) in the kidney transplantation field as well. In order to make the process created available for the Brazilians citizens, a web portal is under construction. We have been using PHP language and MySQL database considering web and mobile access. In this the non-specialized healthcare professionals, non-specialized healthcare managers, and citizens from Brazil will find in an easy and dynamic way the process in a graph representation, a legislation repository, information about healthcare provider, and news about transplant.

Results

The main result is the kidney transplantation process in Brazil represented in BPMN available at telemedicina6.unifesp.br/projeto/kidneytransplant/amia2014. We analyzed 18 digital documents that resulted in a process with 40 activities and events, 6 organizations involved and 6 different stages. In consequence a web portal that makes this work public available for the Brazilians citizens is under construction. The web portal will be available at transplante.info in November 2014.

Conclusion

A representation of the kidney transplantation process in Brazil was created using BPMN, which can be read and interpreted by the professionals involved in this process. In add, with the future availability of this process in a web portal, Brazilian citizens must understand it easily. We acknowledge CAPES for scholarship financial support.

References

1. Medina-Pestana JO, Galante NZ, Tedesco-Silva Jr. H, Harada KM, Garcia VD, Abbud-Filho M, et al. Kidney transplantation in Brazil and its geographic disparity. Jornal Brasileiro de Nefrologia. Dec 2011;33(4):472–84.
2. Business Process Model and Notation (BPMN) version 2.0, Object Management Group, Jan. 2011.
3. Sakr S, Awad A. A framework for querying graph-based business process models. Proceedings of the 19th International Conference on World Wide Web [Internet]. New York, NY, USA: ACM; 2010 p. 1297–300.
4. Global Observatory on Donation and Transplantation, GODT. Organ Donation and Transplantation Activities, 2012. Available at <http://www.transplant-observatory.org/Pages/Data-Reports.aspx>. Accessed at July 20, 2014.
5. Valle R, Oliveira SB. Análise e modelagem de processos de negócio: foco na notação BPMN (Business Process Modeling Notation). São Paulo: Atlas, 2009.

Toward a Conceptual Understanding of Social Network Analysis in Public Health Derived from Published Literature

Catherine Pepper, MLIS, MPH, Juliana J. Brixey, PhD, MPH, MSN, RN
School of Biomedical Informatics, UTHealth, Houston, TX

Abstract

Social network analysis (SNA) is an empirical research method used in public health to evaluate organizational effectiveness in collaboration and knowledge flow [1]. Content analysis of articles from a systematic literature review resulted in a concept map that visualized intersecting components of SNA and the public health system. SNA provides a theoretical and scientific basis for describing public health network structures, and the concept map suggests a linkage between "SNA" and "public health information systems."

Introduction

Social network analysis (SNA) is an empirical research method that can graphically model "the mapping and measuring of relationships and flows between people, groups, organizations, computers, or other information/ knowledge processing entities."[1] SNA has been used in public health to describe organizational collaboration and information flows[1], but a conceptual basis to inform these studies has yet to be articulated. Whereas previous systematic reviews of the literature have examined SNA in disciplines such as the social sciences, business, and healthcare[2], this conceptual analysis identifies SNA studies in the context of public health organizations and maps linkages between network measures and components of the public health system. The aim is to derive a conceptual understanding of social network analysis and public health information systems.

Methods

A systematic literature review was conducted to retrieve published research on the application of social network analysis in public health. Content analysis of retrieved articles was performed: first, to identify the conceptual components of SNA and of the public health system by extracting keywords and indexing terms; then, to categorize articles according to study design, network measures used, and data collection and analysis methods. Concept mapping is an accepted method to link concepts in the literature[3]. A concept map was constructed by extracting and relating the main ideas from individual articles, with derived linkages between SNA and the public health system.

Results

The literature searches yielded a study set of 34 articles, with an additional two articles that provided background. Content analysis of these articles identified a variety of study designs, including qualitative, cohort, mixed-methods, experimental designs, and descriptive analyses. Network measures included centrality, density, and socio-spatial. Data analyses included network analyses and sociograms of various variables, including information sharing and collaboration. These elements were depicted in a concept map that provided a visualization of intersecting components of SNA and the public health system.

Discussion

Both SNA and the public health system are displayed in a concept map as complex, multi-layered domains. Two primary targets of SNA investigations emerged from this review: Information diffusion/distribution and coordination/performance of public health systems as measured by organizational collaborations. Results show that SNA provides a theoretical and scientific basis for describing the complexity of the public health system, and the concept map suggests a linkage between "social network analysis" and "public health information systems."

Conclusion

The concepts of social network analysis and the public health system intersect at key points. SNA has been used to describe and analyze the structure and components of the public health system. Disease surveillance systems, a sub-set of the public health system, may be viewed as social networks, thereby candidates for SNA. Further research will focus on defining optimal social network structure for performance of public health surveillance systems.

References

1. Luke DA, Harris JK. Network analysis in public health: History, methods, and applications. Annu. Rev. Public Health. 2007; 28:69–93.
2. Chambers D, Wilson P, Thompson C, Harden M. Social network analysis in healthcare settings: a systematic scoping review. PLOS ONE. 2012; 7(8):e41911.
3. Heinrich, K. T. Mind-mapping: A successful technique for organizing a literature review. Nurse Author & Editor. 2001;11 (2): 4, 7-8.

Categorizing RxNorm Concepts by Treatment Intent

Pascal B Pfiffner[1], Joshua C Mandel[1], and Kenneth D Mandl[1,2]

[1]*Boston Childrens Hospital Informatics Program at Harvard-MIT Health Sciences and Technology, Boston, MA*
[2]*Center for Biomedical Informatics, Harvard Medical School, Boston, MA*

Introduction & Background

For research and user facing applications alike, categorizing drug products by treatment intent is a desirable goal. National Library of Medicine's (NLM) RxNorm terminology does not yet explicitly categorize drugs on the RxNorm concept id (RXCUI) level. However the National Drug File - Reference Terminology (NDF-RT), produced by Veterans Affairs (VA), is distributed alongside RxNorm and provides additional drug information. This presents two interesting approaches to categorizing RxNorm concepts:

Therapeutic Intent & Contraindication provides treatment intent via a `may_treat` relationship for a range of drugs and ingredients. This allows to arrive at drug-level therapeutic intents by looking at a medication's ingredients.

VA Drug Classification classifies drugs, primarily of the "clinical drug"-type concepts (`TTY=[S]CD`) and allows to infer therapeutic intents by looking at a drug's class. Different RxNorm concept types (e.g. branded drug forms, `TTY=SBD`) are not explicitly linked to a drug class, however their class can be deduced by walking RxNorm relationships.

Previous work has determined that NDF-RT covers about 46% of "semantic clinical drug" (`SCD`) type RxNorm concepts (Pathak and Chute, 2010). We were interested whether combining our two approaches would complement each other and cover a higher percentage of drug-type RxNorm concepts, not limited to `SCD`. Having access to treatment intents for e.g. branded drug concepts is useful for patient-facing apps where a patient might know the drug by brand name.

Materials & Methods

RxNorm 1/6/2014 Full Update Release was downloaded from nlm.nih.gov on January 20, 2014. Data was imported into an SQLite database and our lookup algorithms were applied using Python scripts, available from github.com/chb/py-umls, as follows:

RxNorm concepts of `SCD*`, `SBD*`, `*IN`, `BN` and `*PCK` types were retrieved from the database. Each concept's ingredients (`TTY=IN`) were determined by recursively looking up relationships in the `RXNREL` table. Related treatment intents of these ingredients could then be looked up in the `RXNCONSO` table and assigned to the original RXCUI.

A table was created to store known mappings between concepts and VA drug class, as defined in the `RXNCONSO` table. Heuristically defined, immediate relationships of these known concepts—e.g. `tradename_of` for `SBD` to arrive at an `SCD` with known drug class—were then looked up and the same drug class was assigned to these concepts and stored. This process was repeated 5 times, each time looking at relationships for newly mapped concepts.

Results

Of the 203,175 RxNorm concepts, 72% (146,213) were assigned one or more treatment intents, 71% (143,263) were assigned one or more VA drug classes and 80% (163,339) were assigned at least one of the two. While we have yet to qualitatively analyze these assignments, exploratory analyses suggest that results from both approaches are well aligned and complementary. An (shortened) example concept of `SBDF` type is shown in figure 1.

```
{
  rxcui: "364073",
  tty: "SBDF",
  label: "Codeine / Guaifenesin ... [Bron-Tuss]",
  ingredients: ["2670", "5032"],
  va_classes: [
    "[RE301] OPIOID-CONTAINING ANTITUSSIVES/EXPECTORANTS"
  ],
  treatmentIntents: [
    "Bronchitis",
    "Cough",
    ...
  ]
}
```

Figure 1: Example JSON document for a branded drug form (RXCUI 364073, TTY `SBDF`) with drug class and appropriate treatment intents. The "label" and "treatmentIntents" properties were shortened to fit one page.

Conclusions

Depending on use-case, having quick access to either treatment intents or drug class—and in many cases both—might prove useful for a range of medication related computational tasks, not least in view of patient-centered outcomes research with patients managing their medications.

Meaningful Use & EHR Scope

Valerie J. H. Powell, MS, PhD, RT(R)[1], Amit Acharya, BDS, MS, PhD[2], Robert H. Posteraro, MD, MBI, FACR[3], Thankam Paul Thyvalikakath, DMD, MDS, PhD[4,5]
[1]Robert Morris University, Moon Township, PA; [2]Marshfield Clinic Research Foundation, Marshfield, WI, [3]Health Sciences Center, Texas Tech University, Lubbock, TX, [4]School of Dentistry, Indiana University (IUPUI), Indianapolis, IN, & [5]Regenstreif Institute, Inc., Indianapolis, IN

Abstract
Assuring that the scope of the electronic health record (EHR) includes the entire human body (the stomatognathic system along with the "rest of the body") would contribute to advancing the goals of Meaningful Use (MU).

Introduction
MU is defined as using certified EHR technology in order to improve quality, safety, efficiency, & reduce health disparities, engage patients & family, improve care coordination, & population & public health, as well as to maintain privacy & security of patient health information. The scope of the EHR is often bounded by the traditional separation of the scope of health care into medical & dental categories & therefore does not necessarily include the dental portion of the patient's record. However, microbes & disease processes are under no obligation to respect any division of health care delivery/research & patient records into medical & dental silos. If we examine MU, we find among the core measures, CMS69v1, NQF 0421, Preventive Care & Screening: Body Mass Index (BMI) Screening & Follow-Up. Evidence reveals relationships on BMI & obesity with two oral chronic conditions, caries & periodontal disease.[1,2,3] Another core measure is CMS68v1, NQF 0419, Documentation of Current Medications in the Medical Record. Because, as detailed in Powell et al., eds.,[4] a given patient can receive concurrent medical & dental care, including surgery & prescriptions from the dental provider, documentation of current medication requires listing medications that are in the patient's dental record to be complete. Yet another core measure is: CMS90v1, Functional status assessment for complex chronic conditions. As long as the separation of scopes of medical & dental care & their respective patient records systems persist, neither the common systemic nor the oral health chronic conditions can be adequately addressed through separated streams of care. Extensive clinical literature, for example, documents the interrelationships between periodontal disease & diabetes mellitus. Examining Löe[5] reveals how well-established the understanding of this connection is. To respond to the report of Chaffee et al.[6] to improve care safety & quality would require availability of the maternal dental record during the prenatal period in order to prevent caries & early childhood caries (ECC) in the newborn. To improve quality & safety for patients, in view of the implications of the research of Rubinstein et al.[7], revealing a connection between a common oral microbe & colorectal cancer, requires integrated medical & dental care & patient records. A decision as to how to prioritize categories of patients who care would benefit in quality & safety from an EHR scope that explicitly includes oral health data can be made utilizing the 30 plus contact points discussed in Powell et al. eds[4].

Conclusion
Evidence indicates that MU objectives & core measures would be served by assuring that the scope of the EHR includes the stomatognathic system & associated structures, as well as the remaining systems of the body. Therefore we propose that the scope of the EHR for purposes of MU be established as the entire human body.

References

1. Al-Zahrani MS (2003). Obesity & Periodontal Disease in Young, Middle-Aged, & Older Adults. J Perio 74(5):610-615, doi 10.1902/jop.2003.74.5.610
2. Morita I, Okamoto Y, Yoshii S, Nakagaki H, Mizuno K, Sheiham A, Sabbah W (2011). Five-year incidence of periodontal disease is related to body mass index. J Dent Res. 90(2):199-202. doi: 10.1177/0022034510382548.
3. Subramaniam P, Singh D (2011). Association of age specific body mass index, dental caries & socioeconomic status of children & adolescents. J Clin Pediatr Dent 36(2):175-9.
4. Powell V, Din, F, Acharya A, Torres-Urquidy, MH, eds. (2012). Integration of Medical & Dental Care & Patient Data. Springer-UK.
5. Löe H (1993). Periodontal disease: the sixth complication of diabetes. Diabetes Care 16(1): 329-333.
6. Chaffee BW, Gansky SA, Weintraub JA, Featherstone JDB, Ramos-Gomez FJ (2014). Maternal Oral Bacterial Levels Predict Early Childhood Caries Development. J Dent Res 93:238-244. doi: 10.1177/0022034517713
7. Rubinstein MR, Wang X, Liu W, HaoY, Cai G, Han YW (2013). *Fusobacterium nucleatum* Promotes Colorectal Carcinogenesis by Modulating E-Cadherin/β-Catenin Signaling via its FadA Adhesin. Cell Host & Microbe 14, 2: 195-20

Temporomandibular Disorder Treatment Adherence Improved By A Mobile SMS-Based Intervention

Cristiana S. Prado, MSc Fellow[1], Frederico M. Cohrs, PhD Fellow[1], Cristina L. F. Ortolani, PhD, Full Professor1,[2], Evandro E. S. Ruiz, PHD, Associate Professor[3], Ivan T. Pisa, PhD, Associate Professor[1]

[1]Universidade Federal de São Paulo – UNIFESP, Brazil, [2]Universidade Paulista – UNIP, Brazil, [3]Universidade de São Paulo – USP, Brazil.

Abstract

The use of mobile phone short message service (SMS) leaveraging follow-up therapy has been studied in various areas of health care. We present a research aimed to assess the effects this messaging media as intervention technique on temporomandibular disorder (TMD) treatment adherence during by a randomized controlled trial.

Introduction

The TMD refers to a set of conditions that occur in the orofacial region that compromise the comfort and normal functioning of the hard and soft tissues of the masticatory system[1]. Literature presents positive results in researches using SMS reminders via cellphone to improve patients' adherence to treatments of different kinds of diseases[2].

Methods

The development of web system used in this study, DTM Alert System, is based on an existing model[3]. A randomized controlled trial was conducted at the TMD and orofacial pain outpatient clinic at Universidade Federal de São Paulo (UNIFESP). The population consisted of 38 patients, selected according to inclusion and exclusion criteria and allocated randomly into two groups. The control group (n=19) received standard treatment and the intervention group (n=19) received standard treatment plus SMS reminders. Each patient was followed up for three months and adherence was measured by assessing the patient`s clinical evolution through objective variables (vertical extent of mouth opening, right, left maximum bite force) and visual analog pain scale (VAS) subjective variable. Non-attendance related to follow-up visits and satisfaction questionnaires regarding the delivery and outpatient care were also assessed. This study was approved by Brazilian Ethics Committee SISNEP (# 2135/11).

Results

The DTM Alert System was tested and approved as a project platform for inpatient SMS communication. The tests included message delivery likelihood, system usability and messages content analysis. A descriptive data analysis was carried out. The mouth opening average values, maximum bite force and VAS in the intervention group had a better average when compared to the control group (12% vs. 6%; 11% vs. -12%; 78% vs. 46% improvement respectively). The percentage of control group non-attendance was 22%, while the percentage of the intervention group non-attendance was 11%. The responses to the satisfaction survey were positive.

Conclusion

The system was considered useful and feasible for use in the outpatient clinic where the research was carried out. The results related to the clinic evolution variables and non-attendance suggest that the support offered by SMS reminders had a positive impact in the TMD symptoms and adherence to treatment.

Acknowledgment: CAPES/DS – #33009015 financial support, KATU - Intelligent Systems for Health, José Marcio Duarte for the web system development and Antônio Sérgio Guimarães for the clinical support.

References

1. Dworkin S F, LeReshe L, De Rouen T, Von Korff M. Research diagnostic criteria for temporomandibular disorders: review, criteria, examinations and specifications, critique. J Craniomandib Disord. 1992; 6: 301-355.
2. Déglise C, Suggs LS, Odermatt P. SMS for disease control in developing countries: a systematic review of a mobile health applications. J Telemed Telecare. 2012 Jul;18(5): 273-81.
3. da Costa TM, Barbosa BJ, Gomes e Costa DA, Sigulem D, de Fátima Marin H, Filho AC, Pisa IT. Results of a randomized controlled trial to assess the effects of a mobile SMS-based intervention on treatment adherence in HIV/AIDS-infected Brazilian women and impressions and satisfaction with respect to incoming messages. Int J Med Inform. 2012 Apr;81(4):257-69.

Data Mining Methodologies to Discover Best practices for Diabetic Patients with Health Disparities

Lisiane Pruinelli, MSN, RN[1]; Sanjoy Dey, PhD-C[2]; György J. Simon, PhD[3]; Pranjul Yadav[2]; Andrew Hangsleben[2]; Katherine Hauwiller[2]; Vipin Kumar, PhD[2]; Connie W. Delaney, PhD, RN, FAAN, FACMI[1]; Michael Steinbach, PhD[2]; Bonnie L. Westra, PhD, RN, FAAN, FACMI[1]

[1]School of Nursing, University of Minnesota; [2]Department of Computer Science & Engineering, University of Minnesota; [3]Institute for Health Informatics, University of Minnesota;

Abstract

The use of data mining techniques for analyzing data sets from electronic health records (EHR), and evaluating and expanding evidence based practice (EBP) guidelines requires new approaches. Mining patterns of care for sub-populations over a period of time enables the discovery of best practices for diabetic patients with health disparities and strengthens the content and appication of EBP for clinical decision support.

Introduction

The use of multidisciplinary scientific EBP guidelines during hospitalization can assist diabetic patients to regain their health and thus reduce rehospitalization. Our research reuses EHR data to address new research questions that explore patterns of patients characteristics and resources, EBP interventions, and health impovement. To acomplish these goals, we develop approaches for feature extraction from an EHR to characterize the patient's state at hospital admission and track their progression over time. This analysis will identify the factors (interventions or other aspects of the patient's health status or social determinants) that contribute to successful and unsuccessful outcomes for the patients with diabetes. The interventions both medical and nursing guidelines as well as discovering data derived evidence (DDE). This study aims to describe a methodology for assessing the value of the EBP guidelines and finding DDEs for subgroups of patients, particularly for those with health disparities.

Methods

After IRB approval, data were extracted from a clinical data repository (CDR) at the UMN. The CDR contains more than 2 million patients from 8 hospitals and 40 clinicals using the same EHR. Our de-identified dataset included patients hospitalized between 1/1/09 and 12/31/13 with a primary or secondary diagnosis of diabetes and their follow up for at least 2 years. Data included : social and demographics, diagnoses, vitals, labs, procedures, medications and flowsheets.

One part of this research involves analysis of complications of diabetes: cardiovascular and peripheral vascular diseases, nephropathy, retinopathy, neuropathy, and ultimately death. Different sets of comorbid conditions have different risks for outcomes. Survival association techniques using the Cox-proportional hazards model with Martingale residuals and Association rule mining can be used to detect interactions between co-morbid conditions and risk factors.

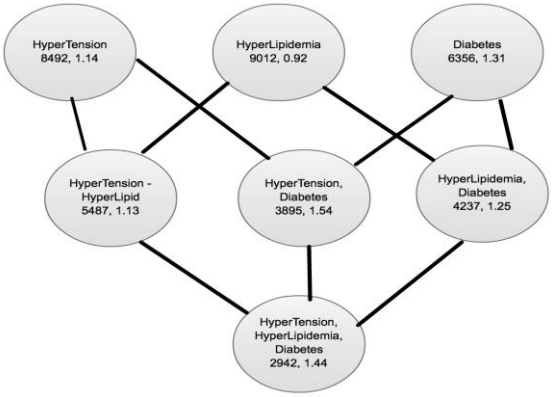


Figure 1: Diagram with sample size and relative risk of mortality for each condition.

Results/Conclusion

In the figure, we present the sample size and relative risk of mortality for diabetes related disease progression. We can observe how the relative risk increases significantly when the patient has hypertension and diabetes, as compared to when the patient is diagnosed with hypertension and hyperlipidemia.

Acknowledgment

This study is supported by NSF grant NSF IIS-1344135 and by Grant Number 1UL1RR033183 from the National Center for Research Resources (NCRR) of the National Institutes of Health (NIH) to the University of Minnesota Clinical and Translational Science Institute (CTSI).

Using EHR Data to Automate Colorectal Cancer Screening at Community Health Centers: Opportunities and Barriers

Jon Puro, MPA:HA[3]; Gloria D. Coronado, PhD[1]; Amanda Petrik, MS[1]; Josue Aguirre[2]; Tanya Kapka, MD, MPH[1,2]; Beverly Green, MD, MPH[4]; Thuy Le[3], Tim Burdick, MD[3]

[1] Kaiser Permanente Center for Health Research
[2] Virginia Garcia Memorial Health Center
[3] OCHIN
[4] Group Health Research Institute

Abstract / Background. *The Strategies and Opportunities to Stop Colorectal Cancer (STOP CRC) study is a collaboration between a health informatics system, clinic networks, and research institutions. The study will assess the reach and effectiveness of an intervention program using electronic health record (EHR) clinical decision support (CDS) tools to improve rates of colorectal-cancer screening in federally qualified health centers (FQHCs). Few studies, and no large studies, aimed at raising CRC screening rates have utilized an EHR-embedded system. Our study strives to automate CRC screening to as great a degree as possible, thus improving clinical efficiencies and patient outcomes. However access to comprehensive clinical data varies in completeness and quality in an "open" health care network, which led to implementation challenges.*

Methods: Selecting patients for automated CRC screening intervention depends on access to data regarding previous such screening done on patients to determine if they are due for another screening: data pertaining to inclusion criteria (patient age, PCP assignment, provider type, viable mailing address, and visit in past 12 months), and data pertaining to exclusion criteria (e.g., known conditions and diagnoses such as colorectal disease, and previous screening). We examined multiple sources of data within the EHR as well as outside sources to find patients eligible for automated screening. Chart audits were performed to validate inclusion and exclusion processes.

Results: We created processes for identifying patients due for CRC screening, but with caveats and compromises on ideal data availability. With multiple lab interfaces processing fecal occult blood tests, multiple outside sources of colonoscopy results, and systemic barriers to indigent patients receiving recommended care came challenges that had to be surmounted or taken into account. Our findings from the initial validation work led to restrictions on the exclusion criteria to cast a wider net over patients in need of screening. This led to instituting new clinical workflows and data collection tools to improve future patient selection methods and, as a result, improve screening rates.

Discussion: In an ideal world, all patients would receive necessary CRC screening, and all such data would be available to clinicians and researchers. This is not the case. We will discuss (a) barriers to adequate patient care, (b) problems with data availability and quality, and (c) methods for working around and improving these issues.

Learning Objective: Attendees will be able to formulate an approach to the design and implementation of a population-based intervention requiring integration of lab, procedure, and other clinical data.

Funding source: National Center for Complementary & Alternative Medicine of the National Institutes of Health under Award Number UH2AT007782

Examining the Potential for CPOE System Design and Functionality to Contribute to Medication Errors

Arbor J.L. Quist[1], Alexandra Robertson[1], Thu-Trang Thach, MPH[1], Lynn Volk, MHS[2], Adam Wright, PhD[1,2], Shobha Phansalkar, RPh, PhD[1,3], Sarah Slight, PhD[1,4], David W. Bates, MD[1,5], Gordon D. Schiff, MD[1,5].

[1]Brigham and Women's Hospital, Boston, MA; [2]Partners HealthCare, Wellesley, MA; [3]Medi-Span, Indianapolis, IN; [4]School of Medicine, Pharmacy and Health, The University of Durham, UK; [5]Harvard Medical School, Boston, MA.

Abstract

To understand how the design of computerized prescriber order entry (CPOE) systems can facilitate medication errors, we collected functionality data from 10 different CPOE systems and examined software team support reports. We found that inconsistencies in CPOE functionality, confusing design, and lack of standardization contribute to medication errors and user frustration.

Background

The Institute of Medicine has estimated that more than 1.5 million people are harmed by medication errors each year; thus, reducing medication errors is essential in improving patient safety[1].While CPOE systems reduce medication errors by using effective clinical decision support (CDS) and creating legible prescriptions, they can also introduce new types of errors[2]. Inflexible ordering screens, inability of information systems to communicate, and prescribers relying on often inaccurate CPOE dose displays are examples of ways that CPOE systems can facilitate medication errors[3]. The Brigham and Women's Hospital (BWH) CPOE Medication Safety (CPOEMS) Project was funded by the FDA to better understand the functionality, features, and interfaces of 10 CPOE systems and how they may contribute to errors.

Methods

We studied 10 CPOE systems that varied in providers, vendors, and care settings across 6 different healthcare facilities. We collected CPOE functionality data through both usability testing and detailed interviews with 35 CPOE system users and IT specialists at each site. We created test patients in each system and developed a CPOE Medication Safety Assessment Tool to assist in collecting standardized information on the systems, including details on drug name presentation, alerts, and system history. These interviews and system evaluations were then recorded, transcribed, and organized to highlight the similarities and differences between functionality in the systems. Finally, we reviewed integrated software team reports to understand how CPOE issues lead to medication errors.

Results

We found a large number of inconsistencies within and across systems, including confusing CDS and system specific issues that could cause medication errors. For example, one site's inpatient and outpatient systems fire different alerts for the same drug, causing confusion in medication profiles. A review of integrated software team reports in one outpatient system revealed that alerts fire only when medications are prescribed anew, and not when re-ordered. We also observed variation in ways that medications are discontinued. One system allows, but does not require, a discontinuation reason while another requires the user to choose one of 10 reasons for discontinuation.

Conclusion

We developed a better understanding of CPOE functionality by examining error reports, usability assessments, and vulnerability testing. This analysis helps us to conclude that CPOE system functionality can be improved to prevent frustrations and confusion errors through more standardized and less confusing interfaces and design features.

References

1. National Research Council. Preventing Medication Errors: Quality Chasm Series. Washington, DC: National Academy Press, 2007.
2. Boehne J, Seger A., Amato M., Whitney D, Schiff G. Analysis of USP MEDMARX Medication Error Data: CPOE as Contributing Factor. Presented July 13, 2011 NCCMERP Meeting.
3. Koppel,R., Metlay, J.P., Cohen, A., Abaluck, B., Localio, A.R., Kimmel, S.E., Strom, B.L. Role of Computerized Physician Order Entry Systems in Facilitating Mediation Errors. JAMA. 2005;293:1197-1203.

With Whom Will I Share?
A Quantitative Data Analysis of the Special Project of National Significance Survey for Persons Living with HIV/AIDS.

S. Raquel Ramos, MSN, RN, MBA, Peter Gordon, MD, Suzanne Bakken, RN, PhD
Columbia University Medical Center, New York, NY

Abstract

Health information exchange (HIE) is utilized nationally. However in New York State (NYS), HIE participation is under appreciated and potentially more so from those living with chronic illness, such as HIV. The purpose of this study was to quantitatively explore factors associated with persons living with HIV (PLWH) willingness to share all of their personal medical data electronically with specific types of healthcare personnel.

Introduction

Patient tracking, linkage and retention in care have been empowered through health information technology (HIT) innovations, such as HIE. HIE allows clinicians to access pertinent up-to-date health information that could potentially save lives, yet it may be underutilized in NYS by PLWH. Moreover, it has the potential to strengthen patient self-agency[1], patient care and outcomes in the midst of HIV stigma, fear and discrimination. As a chronic illness with over 1 million persons infected, HIV has been associated with co-morbid conditions that affect the cardiovascular, hepatic and endocrine systems. Therefore it is imperative that continuity of care be facilitated through the HIE. Thus, we sought to use quantitative methods to explore what factors were associated with PLWH willingness to share all medical data electronically with specific types of healthcare personnel.

Methods

A survey from the Special Projects of National Significance (SPNS) was completed by PLWH (N=291) participating in a Medicaid Special Needs Plan for PLWH in New York City. Independent variables selected for analysis were: age, gender, income, education, being US born, ethnicity and sexual orientation. The dependent variables were sharing all personal medical data electronically with specific healthcare clinicians including specialists. Descriptive statistics and non-parametric techniques were performed to select variables for the logistic regression using SPSS v21[2]. A binary logistic regression was performed and the Bonferroni correction[3] was applied for the multiple dependent variables; p – value = 0.0125 was used as the alpha for significance.

Results

No independent variables significantly influenced a PLWH willingness to share all of their personal medical data electronically with their PCP or other clinicians. However, being US born significantly influenced PLWH willingness to share all personal medical data electronically with other health providers (OR=3.083, CI=1.633-5.810) and with non-HIV specialists (OR= 2.504, CI = 1.32-4.569). Persons that are US born are three times more likely to share all of their personal medical data electronically with other health providers (ER/Hosp personnel) and non-HIV specialists (cardiologists/OBGYN) than those who are not US born.

Conclusions

With the expansion of HIE in healthcare, further insights are warranted to better understand what specific cultural factors influence HIV- positive non-US born person's willingness to share all of their medical data electronically.

Acknowledgment: Survey funded by Health Resources and Services Administration (H97HA08483).
Ms. Ramos is supported by T32NR007969 (Bakken, PI)

References

1. Teixeira, PA,, Gordon, P, Camhi, E, Bakken, S. HIV patients' willingness to share personal health information electronically. Patient education and counseling. 2011; 84:e9-e12.
2. Pallant J. *SPSS survival manual: A step by step guide to data analysis using SPSS.* McGraw-Hill International. 2010.
3. Munro B H. *Statistical methods for health care research.* Lippincott Williams & Wilkins. (Vol. 1) 2005.

Evaluation of Existing Phenotype Authoring Tools for Clinical Research

Luke V. Rasmussen[1]; Jie Xu, MS[1], Ruijue Liu[2], Qian Zhu, PhD[3], Jennifer A. Pacheco[1], Jyotishman Pathak, PhD[3], William K. Thompson, PhD[4], Joshua C. Denny, MD, MS[5], Huan Mo, MD, MS[5], Richard C. Kiefer[3], Peter Speltz[5], Enid Montague, PhD[1]
[1]Northwestern University, Chicago, IL; [2]National University of Singapore, Singapore; [3]Mayo Clinic, Rochester, MN; [4]NorthShore University HealthSystem, Evanston, IL; [5]Vanderbilt University, Nashville, TN

Abstract
In the course of developing phenotype algorithm definitions, graphical tools and query editors are often used to assist with abstracting away the complexities of the underlying data. Many tools have been created and reported in the biomedical literature to assist with definition building, but no work to date has performed a comparison of the features available in these tools. This work conducts a preliminary analysis of the features available in ten publicly reported tools for phenotype algorithm authoring.

Introduction. Phenotyping is the process of systematic collection and analysis of phenotypic data [1]. The implementation of electronic health records (EHRs) on the national scale has a potential to tremendously scale the phenotyping process for clinical research [2]. However, one challenge is the lack of a validated tools to author and apply standardized phenotype algorithms across different sites and EHR systems accurately and efficiently [3]. The aim of this study was to identify and evaluate platforms that may be used to author phenotype algorithms.

Methods. The research team identified a number of tools that are able to construct phenotype definitions of varying complexity, through literature review and expert inquiry. These platforms were evaluated for the following functionality deemed by the study team to be relevant to phenotype construction and/or execution: capability of using temporal operators (specific date, time interval, time overlap, and time sequence) and Boolean operators (basic and nested AND/OR and NOT), search functions (search by codes, keywords, and other advanced search functions), user interface (dashboard display of results, web-based application, drag-drop function, and chart visualization), and other functions (information returned from query and if the platform is open source). Evaluation was conducted on published materials supporting the tool, including scientific publications, system documentation, and execution of public demos.

Results. Ten platforms that can be used for authoring phenotype algorithms were identified and reviewed; the Biomedical Translational Research Information System (BTRIS), Duke Enterprise Data Unified Content Explorer (DEDUCE), eMERGE Network Record Counter (eRC), Eureka!, HealthFlow, Informatics for Integrating Biology and the Bedside (i2b2), Measure Authoring Tool (MAT), Stanford Translational Research Integrated Database Environment (STRIDE), Visual Aggregator and Explorer (VISAGE), and Vanderbilt University's Synthetic Derivative (VUSD). With respect to certain features, six of the ten tools lacked the ability to relate items occurring at the same time. Only four of the ten made their source code readily available, and three of the ten were created for a specific institution or network. Additional details of the analysis will be presented.

Discussion. While there is a great deal of overlap between functionality in the reviewed systems, there is also great variance, which may be attributed to the different scenarios for which the tools were designed. For example, tools such as the MAT are used purely for logic representation and not execution, and the eRC and i2b2 serve as platforms for feasibility queries. The lack of temporal operators overall, while often not needed for simple phenotypes, is often used to create more robust phenotype definitions. Future work will evaluate the graphical paradigms used in these tools to implement features, which may be used to inform enhancements and the design of future platforms.

Acknowledgement. This work has been supported in part by funding from the National Institutes of Health (R01-GM105688).

References

1. Freimer N, Sabatti C. The Human Phenome Project. Nature Genetics. 2003;**34**(1):15-21.
2. Hripcsak G, Albers DJ. Next-Generation Phenotyping of Electronic Health Records. Journal of the American Medical Informatics Association. 2013;**20**(1):117-21.
3. Pathak J, Kho AN, Denny JC. Electronic Health Records-Driven Phenotyping: Challenges, Recent Advances, and Perspectives. Journal of the American Medical Informatics Association. 2013;**20**(e2):e206-e11.

Using EHR Timestamps for Analyzing Ophthalmology Clinic Workflows

Sarah Read-Brown[1*], Michelle R. Hribar, PhD[2*], Leah Reznick, MD[1], Thomas Yackel, MD, MPH, MS[2], Michael F. Chiang, MD, MA[1,2] (*Co-first Authors)

[1]Ophthalmology and [2]Medical Informatics & Clinical Epidemiology; OHSU, Portland, OR

Introduction

Patient flows in clinical ophthalmology are complex: patients see multiple providers, undergo dilation, and may undergo office procedures. Due to the complexity of and multistep nature of this process, patient wait times can accumulate, affecting patient satisfaction[1] and future care[2]. Being able to predict variability and patient exam times has potential to greatly improve scheduling strategies[3]. We propose using timestamps from historical exam data from an electronic health record (EHR) for this prediction, and to validate them with observed time-motion data.

Methods

This study took place at Oregon Health & Science University, focusing on a single ophthalmology provider's clinic (LR). Clinic exams consisted of the following steps: 1) Patients are initially examined by an ophthalmic technician, orthoptist and/or resident, 2) their eyes may or may not be dilated, and 3) they are examined by a physician. The authors (SRB, MRH) observed a clinic for 3 days, using time-motion methods to record the times that patients spent during each part of their exam using a combination of paper and electronic forms. We then compared these observed times to timestamps recorded in the EHR (Epic; Verona, WI) during exams. These included time points such as scheduled appointment time, patient check-in time, check-out time, dilation time, and audit log data. For each provider, we used the timestamps of the first and last audit log entries made during the exam time recorded on workstations in exam rooms. Differences between the observed vs. EHR timestamps were calculated.

Results

Table 1 shows the individual exam times by each provider, as well as total exam times. Overall, 67% of exam times from EHR timestamps fell within 3 minutes of observed times. Among the 14/82 (17%) of exams in which discrepancies between these times were >5 minutes, review of time-motion observation notes determined that providers did not use the EHR at the start of the exam and/or stop using the EHR at the end of the exam.

Table1: Comparison between EHR and Observed Exam Times

Time Difference EHR Timestamp vs. Observed Data					
	< 1 Min	1-3 Min	3-5 Min	> 5 Min	N
Initial Exam	9 (32%)	13 (46%)	5 (18%)	1 (4%)	28
Resident Exam	1 (50%)	-	-	1 (50%)	2
Physician Exam	9 (33%)	12 (44%)	1 (4%)	5 (19%)	27
Total Exam	7 (28%)	4 (16%)	7 (28%)	7 (28%)	25
Total	26 (32%)	29 (35%)	13 (16%)	14 (17%)	82

Conclusion

EHR timestamp data can be used to estimate exam times as long as providers follow a consistent workflow that captures start and end times of exams. This creates potential for using these data for strategies to evaluate and improve clinical efficiency. More work is needed to expand and validate this EHR time stamps method to other providers and clinics.

References

1 McMullen, M., & Netland, P. A. (2013). Wait time as a driver of overall patient satisfaction in an ophthalmology clinic. *Clinical Ophthalmology, 7*, 1655-1660.

2 Lee, B.W.a , Murakami, Y.a , Duncan, M.T.a , Kao, A.A.b , Huang, J.-Y.b , Lin, S.b , Singh, K.a (2013). Patient-related and system-related barriers to glaucoma follow-up in a county hospital population. *Investigative Ophthalmology and Visual Science*, 54 (10), pp. 6542-6548.

3 Salzarulo, P. A., Bretthauer, K. M., Côté, M. J. and Schultz, K. L. (2011), The Impact of Variability and Patient Information on Health Care System Performance. *Production and Operations Management*, 20: 848–859.

Effect of Pre-annotation on Annotation Time

Andrew Redd, PhD[1, 2], Youngjun Kim, MS[2, 3], Stéphane M. Meystre, MD, PhD[3, 4], Julia Heavirland[4], Allison Weaver[5], Jenifer Williams[2] , Jennifer Garvin, PhD, MBA[2, 4],

[1]Division of Epidemiology, Department of Internal Medicine, University of Utah; [2]VA Health Care System, Salt Lake City, Utah; [3]School of Computing, [4]Department of Biomedical Informatics, University of Utah; [5]Centers for Medicare & Medicaid Services

Abstract

In the Department of Veterans Affairs (VA) project Automated Data Acquisition for Heart Failure (ADAHF) we had annotators annotate documents with and without pre-annotation to assess the impact of pre-annotation of documents on the time required to perform annotation. We found that pre-annotation was statistically significant and saved time on annotation.

Introduction

Human capital is the largest expense in almost all research. In Informatics, a significant task involved in training information extraction algorithms is the time needed for human review and annotation of documents [1]. In the Department of Veterans Affairs(VA) project Automated Data Acquisition for Heart Failure (ADAHF), annotators reviewed documents for concepts related to VA heart failure treatment performance metrics. The annotated documents were used to train algorithms for automated information extraction. We examined the effect of pre-annotation of documents with significant keywords on the time needed to perform annotation.

Methods

The corpus, consisting of 150 documents, was separated into 7 batches, which were then randomly split into three sets. For each set one annotator received the documents preannotated with concepts; the other two annotators received the documents in a raw state with no pre-annotations. Preannotations were created through a combination of machine learning tools [2], which detected ejection fraction related concepts, and a dictionary lookup, which detected medication related concepts. The annotators used Protégé [3] with the Knowtator plugin [4] to perform the annotation task. Time for each annotator to complete the task was measured and recorded in a spreadsheet.

Statistical Methods

The statistical model used fit the log transformed time with the effects of annotator, pre-annotation status, and their interaction, with a random blocking effect for the document to adjust for document complexity. The model was checked for and satisfied standard validity checks for linear models. Similar papers have reported on the benefit of preannotation before [1] [5] [6], however we additionally modeled the effect of preannotation with the effect of individual annotators as well as the possible interaction between annotators and preannotation.

Results

Similar to Lingren et al. [6], preannotation reduced the time required by 13%. For a typical document in our corpus, pre-annotation saved 8.7 seconds from a total 66.4 seconds in the annotation task. The log transformation in the analysis indicates that the larger number of concepts and the longer the document (the annotation task) the more pre-annotation will reduce time; for example, if a document would otherwise take 2 minutes to annotate, pre-annotation would take a full 15.8 seconds off the total time of the annotation task. We found both the annotator and the pre-annotation effects to be statistically significant, and the interaction to not be significant. In comparison the effect from the annotator is larger than the effect from the pre-annotation.

Discussion

We found that all annotators benefit similarly from pre-annotation. However, the pre-annotation does not outweigh the effect of individual annotators, indicating that pre-annotation does not substitute for well-trained and efficient annotators.

Acknowledgements

The research reported here was supported by the Department of Veterans Affairs, Veterans Health Administration, Office of Research and Development, Health Services Research and Development Service IBE 09-069. Dr. Redd is a Statistician at the VA Salt Lake City Health Care System IDEAS Center. The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the United States government.

References

1. Ringger E, Carmen M, Haertel R, Seppi K, Lonsdale , McClanahan P, et al. Assessing the Costs of Machine-Assisted Corpus Annotation Through a User Study. In Proceedings of the Sixth Language Resources and Evaluation Conference; 2008; Marrakech, Morocco. p. 3318-3324.
2. Garvin JH, et al. Automated extraction of ejection fraction for quality measurement using regular expressions in Unstructured Information Management Architecture (UIMA) for heart failure. JAMIA. 2012; 19(5).
3. Stanford Center for Biomedical Informatics Research. Protege. [Online].http://protege.stanford.edu/ accessed Feb 2014
4. Ogren P. Knowtator Website. [Online].http://knowtator.sourceforge.net/ accessed 1 Jun 2011
5. Ganchev K, Mandel M, Pereira F, Carroll S, White P. Semi-automated named entity annotation. In Proceedings of the linguistic annotation workshop. p. 53-56.
6. Lingren T, et al. Evaluating the impact of pre-annotation on annotation speed and potential bias: natural language processing gold standard development for clinical named entity recognition in clinical trial announcements. JAMIA. 2014; 21(3): p. 406-413.

Visualizing Technology Types for Gait Speed Detection in Older Adults

Blaine Reeder, PhD[1], Kamin Whitehouse, PhD, MS[2]
[1]University of Colorado, Aurora, CO [2]University of Virginia, Charlottesville, VA

Abstract

Gait speed is a predictor of adverse events in older adults. We conducted an integrative review of gait speed technologies used in studies that enrolled older adult participants to identify distinct detection approaches and the potential for these technologies to support gait speed monitoring in everyday life. This presentation provides a visual overview of gait speed technology types and future directions for research in this area.

Introduction

Changes in gait speed are an important indicator of functional change in older adults. Slow gait speed is associated with falls, mobility limitations, hospitalization and mortality[1]. Older adults prefer to live independently in their own homes or "age in place". Technologies that monitor gait speed in everyday life have the potential to support this desire through alerts in functional change. Recent years have seen a number of different approaches to monitor gait speed at home but reported evidence is fragmented because it is indexed in discipline-specific literature repositories. To address this gap, we conducted an integrative review of gait speed technologies. Here we present a visualization of gait speed technology types identified during our review.

Methods

Using Whittemore and Knafl's methodology[2], we conducted an integrative review of technologies to detect gait speed used in studies that enrolled older adult participants. PubMed and IEEE Xplore databases were searched with keyword synonym groups for "gait speed", "sensor" and "older adult". Five hundred thirty-nine total articles were identified for review. Technologies were classified according to type and visualized in a graphic to facilitate explanation and understanding of their function.

Results

Sixteen articles that described ten different technology types published between 2002-2014 were included. Technology types were split into categories of *installed* and *body-worn* sensors (Figure 1). Installed sensors provide comprehensive home-based data but video may present privacy challenges. Worn sensors provide data outside the home but present adherence issues. In just over a decade, gait speed technology research has advanced beyond the use of customized research hardware prototypes to the point where use of consumer available technologies is common.

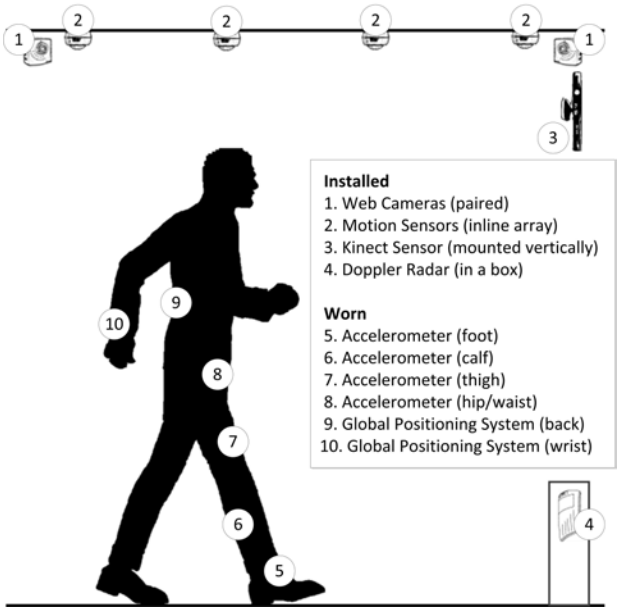


Figure 1. Types of installed and body-worn gait sensors

Conclusion

Gait speed technology research still requires deep technology expertise and substantial integration efforts to achieve research aims, especially with regard to data analysis and presentation. Future research should include development of visualization software and exploration of ways to incorporate gait information into personal health records, health care provider workflows, and older adults' everyday lives to support behavioral interventions for independent living. In addition, older adults' perceptions of different technology types should be compared within the same studies.

References

1. Montero-Odasso M, Schapira M, Soriano ER, et al. Gait velocity as a single predictor of adverse events in healthy seniors aged 75 years and older. *J Gerontol A Biol Sci Med Sci.* 2005;60(10):1304-1309.
2. Whittemore R, Knafl K. The integrative review: updated methodology. *J Adv Nurs.* Dec 2005;52(5):546-553.

Incorporating an Electronic Ventilator-Associated Event (VAE) Tool within a Hospital's Internal Electronic Surveillance System

Ervina Resetar, MIM, PMP[1,3], Kathleen M. McMullen, MPH, CIC[2], Anthony J. Russo MPH[2], Joshua A. Doherty, BS[3], Kathleen A. Gase, MPH, CIC[3], Keith F. Woeltje, MD, PhD[1,3]

[1]Washington University School of Medicine, Saint Louis, MO
[2]Barnes Jewish Hospital, Saint Louis, MO
[3]BJC HealthCare, Saint Louis, MO

Abstract
Mechanical ventilation provides an important, life-saving therapy for severely ill patients, but ventilated patients are at an increased risk for complications, poor outcomes, and death during hospitalization.[1] Using electronic health record (EHR) systems along with microbiology, vital signs, and antibiotic data, BJC HealthCare implemented an electronic Ventilator-Associated Event (VAE) tool and incorporated it within an existing, intranet based, electronic surveillance system.[2] Incorporating the electronic VAE tool resulted in a reduction of Infection Prevention (IP) labor efforts involved in identifying VAE cases and provided a foundation for electronic reporting of VAE cases to the National Healthcare Safety Network (NHSN).

Background
Healthcare-associated infections (HAI) are an important cause of preventable harm in hospitalized patients. Ventilator-associated pneumonia (VAP) is among the most common HAI, but accurate surveillance for VAP has been historically difficult because of the lack of objective definitions. The Centers for Disease Control and Prevention (CDC) National Healthcare Safety Network (NHSN) Working Group developed a new and more objective approach to surveillance that focuses on Ventilator-Associated Events (VAE). These include Ventilator-Associated Conditions (VAC), Infection-related Ventilator-Associated Complications (IVAC), possible VAP and probable VAP.[1] Using the previous VAP definitions, Infection Prevention (IP) staff at BJC HealthCare (BJC) spent approximately 10 hour per intensive care unit (ICU), equating to about 60 IP hours per month at just one facility.

Methods
In response to this surveillance burden, BJC developed an electronic VAE tool and incorporated it within an existing, intranet-based, electronic surveillance system called Surveillance Assistant (SA).[2-3] The clinical decision support system (CDS) consists of registration, lab, vital signs, pharmacy, microbiology and selected nursing assessment data (including ventilator-control settings data), allowing BJC to implement a VAE tool for electronic surveillance for the respiratory status (VAC) and the infection and inflammation components (IVAC) of VAE. Amount of IP staff time spent completing surveillance for VAE was quantified pre- and post-successful implementation of the electronic algorithm.

Results
Prior to the successful launch of the electronic algorithm, IP staff continued to commit about 10 hours per unit per month (60 hours/months) manually applying the NHSN VAE definitions. As a result of this implementation project, the electronic VAE tool automatically identifies new VAC and IVAC cases on a daily basis and loads them into SA. The SA interface provides IP staff with patient level detail data needed for accurate determination of possible and probable VAP; IP staff makes the appropriate selections and saves that information using the same web interface. The complete automation of the VAE definitions decreased IP surveillance time to about 30 minutes per unit per month (3 hours/month), saving roughly 57 hours of IP time monthly in one facility.

Conclusion
Implementation of an electronic VAE tool and integrating it within existing electronic surveillance tools resulted in reduction of labor efforts involved in manual application of VAE algorithm and provided a solid foundation for electronic reporting of VAE cases to NHSN.

References
1. Centers for Disease Control and Prevention National Healthcare Safety Network. Definitions for Ventilator-associated Events. http://www.cdc.gov/nhsn/acute-care-hospital/vae/index.html (accessed February 24, 2014).
2. Resetar E, McMullen KM, McCormick S, Woeltje KF. Development of Electronic Surveillance for Ventilator- Associated Events (VAE) in Adults. AMIA Annu Symp Proc. 2013.
3. Huang Y, Noirot LA, Heard KM, Reichley RM, Dunagan WC, Bailey TC. Migrating toward a next-generation clinical decision support application: the BJC HealthCare experience. AMIA Annu Symp Proc. 2007;344–348.

Using participatory design to optimize capture of information needed for public health reporting processes

Debra Revere, MLIS, MA[1], Jennifer Williams, MPH[2], Rebecca Hills, PhD, MSPH[1,3], Shaun J. Grannis, MD, MS[2,4], Brian E. Dixon, PhD, MPA[2,5,6]

[1]School of Public Health, University of Washington (UW), Seattle, WA; [2]Center for Biomedical Informatics, Regenstrief Institute, Indianapolis, IN; [3]School of Nursing, UW, Seattle, WA; [4]Indiana University (IU) School of Medicine, Indianapolis, IN; [5]IU School of Informatics and Computing, Indianapolis, IN; [6]Dept of Veterans Affairs, VHA, HSR&D Service, Indianapolis, IN

Objective. To engage public health end-users in the participatory design of Communicable Disease Report (CDR) forms that are pre-populated with HIE-derived clinical data (patient demographics, laboratory results, and treatment) and transmitted via the HIE.

Background. Prior research demonstrates a need to improve public health (PH) agencies' ability to more completely and accurately capture routinely collected CD case information from clinical providers. Traditionally these processes involve manually completed paper forms by providers and/or clinic staff. In settings that utilize electronic health record systems, one proposed pathway for improvement is deploying CDR forms pre-populated with electronic data derived from clinical care settings.[1] Participatory design (PD) principles advocate including real users and stakeholders when designing an information system to ensure high ecological validity of the product, incorporate relevance and context to the design, reduce misconceptions designers can make due to insufficient domain expertise, and ultimately reduce barriers to adoption of the system.[2] Prior to deploying a pre-populated CDR form, we engaged PH users and stakeholders as co-designers.

Methods. Over a 9-month period, we engaged stakeholders in a multi-phase PD process as follows:

1. Initial Focus Groups: Focus groups (5-6 participants/group) were held with PH health stakeholders representing communicable and infectious disease investigation in two PH agencies to review the list of existing CDR form data elements and identify data elements that PH routinely request from clinical reporters and are desirable for inclusion on a CDR form, focused on seven high-priority conditions previously identified by PH: Chlamydia; Gonorrhea; Hepatitis B, acute; Hepatitis C, chronic; Histoplasmosis; Salmonella; and Syphilis.
2. IT Input: Project staff held meetings (n=4) with IT staff to categorize and document the technical feasibility of extracting both existing and additional desirable data elements.
3. Survey: Practitioners and investigators from the two PH agencies prioritized data elements feasible for extraction with regards to data needed to close a CDR case.
4. CDR Form Mock-ups: Based on survey data, IT staff developed a modified, pre-populated CDR form mock-up for each high-priority condition.
5. Closing Focus Groups: The original PH stakeholders reconvened to review and achieve consensus on the collated list of final prioritized data elements and provide feedback on the CDR form mock-ups.

Results. By involving PH users in a PD process, we made significant improvements to the layout and functionality of the forms. Focus groups with PH stakeholders identified a number of state-mandated fields that are not highly used or desirable for disease investigation and their elimination allowed engineers to focus on re-purposing form space to capture higher priority data elements. Establishing appropriate user expectations is a critical factor with respect to adoption and receiving early input through PD methods provided new insights into PH workflow and allowed the team to quickly triage user requests while managing PH user expectations within the realm of engineering possibilities. A final CDR form for each condition was designed to meet the needs of PH stakeholders and the technical capabilities of the HIE to automatically pre-populate the forms with available, prioritized data elements. Deployment of the modified, automatically pre-populated CDR forms is in process in pilot clinic settings and evaluation of these improvements on CDR data quality and reporting rates is ongoing.[3]

Conclusions. Innovative IT strategies must be aligned with the requirements and expectations of their users. Engaging PH as a co-designer not only ensured the new CDR forms will meet real-life needs, but also will support development of a product that will improve CDR reporting.

References

1. Grannis SJ, Stevens KC, Merriwether R. Leveraging Health Information Exchange to support public health situational awareness: the Indiana experience. *Online J Public Health Inform* 2010;2:2.
2. Schuler D, Namioka A. Participatory design: principles and practices. Hillsdale NJ: L Erlbaum Associates; 1993.
3. Dixon BE, Grannis SJ, Revere D. Measuring the impact of a health information exchange intervention on provider-based notifiable disease reporting using mixed methods: a study protocol. *BMC Med Inform Decis Mak* 2013;13:121.

Acknowledgments
This project was supported by grant number R01HS020909 (PI: Dixon) from the Agency for Healthcare Research and Quality (AHRQ) and is a collaboration between Indiana University, Regenstrief Institute and the University of Washington. The content is solely the responsibility of the authors and does not necessarily represent the official views of AHRQ.

Patient Perspectives of Mobile Phones' Effects on Healthcare Quality and Medical Data Security and Privacy: A Nationwide Survey

JE Richardson, PhD, MLIS; M Silver, MS, JS Ancker, MPH, PhD
Center for Healthcare Informatics and Policy, Weill Cornell Medical College, NY, NY

Abstract: *Given the growing interest in utilizing mobile phones including cell phones and smartphones for health management, we sought to gauge consumer perceptions of these technologies as related to security, privacy, and healthcare quality. A national random-digit-dial telephone survey was conducted in 2013. Subjects were more likely to endorse the belief that medical data on mobile phones would worsen privacy and security than they were to express similar concerns with electronic health records (EHRs) or health information exchange (HIE). Fewer than half of respondents endorsed a statement that using mobile phones to share personal health data with physicians would improve healthcare quality, which was much lower than the percentage who agreed that physicians' use of EHRs or HIE would improve quality. Consumer security and privacy concerns may need to be addressed before the application of mobile phones for healthcare quality improvement can be fully realized.*

Introduction

Healthcare researchers perceive mobile phones as having the potential to revolutionize the ways in which consumers collect, monitor, and share their medical data with providers.[1] Yet consumer concerns around data privacy and security may limit use[2] and thus negate any potential healthcare quality improvements that could result. We therefore wanted to gauge consumer perceptions of mobile devices on medical data security and privacy, so to inform policies that can target mobile devices as ways to improve healthcare quality.

Methods

In 2013, we conducted a national random-digit-dial telephone survey of subjects with access to landline telephones and cell phones in the United States. Subjects were asked whether healthcare quality and the privacy and security of medical information would improve or worsen with the use of specific technologies: electronic health records (EHRs), health information exchange (HIE) among physicians, and cell phones/smartphones used to store medical data or send it to a physician's EHR.

Results

A total of 1000 respondents were included (a 64% response rate), producing a 3.1% margin of error. Most respondents endorsed the belief that privacy and security would worsen if mobile phones were used to store personal medical data (74%) or send this information to a physician's EHR (69%). By contrast, only 41% expressed similar privacy and security concerns about EHRs or HIE as used by physicians. Fewer than half of respondents endorsed a statement that using mobile phones to share data with physicians would improve healthcare quality (48%), which was much lower than the percentage who agreed that EHRs or HIE would improve quality (61% and 74%). Respondents who believed that EHRs would improve healthcare quality were nearly 3 times as likely to believe that using a mobile device to share data with an EHR would also improve healthcare quality (OR 2.94). Younger individuals were slightly more likely to believe that mobile phones would improve healthcare, but age was not associated with privacy concerns.

Discussion

Despite the growing popularity among consumers for using mobile phones and health-related mobile applications, consumers have concerns about the privacy and security of medical information, and many are skeptical that using mobile phones to share information with physicians will improve healthcare quality. Consumer security and privacy concerns may need to be addressed before the application of mobile phones for healthcare quality improvement and patient engagement can be fully realized.

References

1. Mosa ASM, Yoo I, Sheets L. A systematic review of healthcare applications for smartphones. BMC Med Inform Decis Mak. 2012;12:67.
2. Prasad A, Sorber J, Stablein T, Anthony D, Kotz D. Exposing privacy concerns in mHealth. San Francisco, CA; 2011. Available from: https://www.usenix.org/legacy/events/healthsec11/tech/

The UCDHS Tethered Meta Registry: A Tool for Patient Health and Organizational Quality Improvement

Albert Riedl, MS; Larry Errecart; Sharon Myers, PhD, MPH; Colleen Gordon, MS; Kent Anderson, MS
UC Davis Health System, Sacramento, CA

Abstract

The UC Davis Health System has demonstrated successes in the secondary use of Electronic Health Record (EHR) digital content: EHR content has been summarized to create clinical outcome databases and registries, and where appropriate, clinical data has been de-identified and loaded to Cohort Discovery (i2b2) and other research support software datasets. Challenges to further leveraging clinical data re-use for population health and decision making include data fragmentation or duplication between systems, lack of data entry standards, and the amount of contextual knowledge required to understand the data. Further complicating data re-use is the ongoing evolution of EHR implementation and slow adoption of national healthcare data standards, thus limiting industry knowledge transfer to methodological sharing. UCDHS developed the Tethered Meta Registry (TMR) to overcome fragmented and erroneous data, encourage common institutional data standards and definitions, highlight the importance of data contexts, and integrate with industry standards.

Introduction

The UC Davis Health System captures data across a myriad of clinical and supporting IT systems. Although paperless, the majority of these systems are not designed for data reporting or analysis, nor does any single system contain a complete picture of institutional operations or patient care at the population level. Furthermore, patient data sets or registries that were created to meet departmental or organizational measurement objectives were done so independently and managed in isolation. Although UCDHS has demonstrated success in the re-use of electronic health records, the challenges to advancing this initiative further are data fragmentation, duplication between systems, incomplete data, data errors, and lack of comprehensive data entry standards.

Strategy for Change

The paperless infrastructure at UCDHS is leveraged to create a Tethered Meta Registry (TMR), bringing together various silos of clinical content including EHR, billing and coding, radiology, and Laboratory Information Systems. The TMR is a repository of curated data based on the Epic EHR and augmented by high value raw and derived data from other source systems. These curated data are maintained via a 'tethered,' near real-time connection. 'Meta Registry' refers to a series of tested and validated data sets, known as registries. These registries emerged from the institutional mission to improve patient care, optimize operations, and centralize and standardize external reporting. Each registry is created in partnership with key subject matter experts who understand the data within the tethered systems. Each registry also expands the breadth of the TMR, adding data elements for future use and allowing for cross-registry interoperability.

Results

Currently, the TMR manages about 300 defragmented, de-duplicated, well understood data concepts, aligning over 1 billion data points, and supporting registries which vary in purpose from clinical quality improvement to patient and disease tracking. The TMR has spurred quality improvement projects reducing mortality in critical care settings (specifically, reducing sepsis mortality by 25%), freed resources previously spent on data entry/documentation, helped direct resource allocation across the primary care network, informed the improvement of clinical documentation workflow, and supported numerous grant applications leading to incoming extramural funding. Additionally, it has catalyzed institutional involvement encouraging participation and developed lasting strategic partnerships in UCDHS's effort to re-use clinical data.

Conclusion

The TMR is an effective and efficient way to produce registries to identify, track, and analyze patient populations as well as the care they receive. By providing this information to the clinical teams in a meaningful way, care can be improved. The TMR development process has led to increased institutional engagement in data management, furthering a shared understanding and shared ownership of registry data sets.

Error Propagation in EHRs via Copy/Paste: An Analysis of Relative Dates

Kirk Roberts, PhD, Amos Cahan, MD, Dina Demner-Fushman, MD, PhD
U.S. National Library of Medicine, Bethesda, MD

Abstract

We present a method for identifying errors in EHRs caused by copying and pasting text containing relative dates. Our method utilizes a sequence alignment method for recognizing instances of copy/paste, and regular expressions for relative dates. We furthermore present an analysis of our method on the MIMIC-II dataset.

Introduction

Electronic health records (EHR) are a powerful enabling technology, but concerns exist about the dangers of copying text from one clinical record to another.[1,2] While strategies exist for mitigating copy/paste in text mining approaches, little work has been done to evaluate the types of errors introduced to the EHR through copy/paste. This work studies errors due to the copying of relative dates. When a note contains a relative date (e.g., "*3 days ago*", "*past 4 weeks*", "*43 y/o*"), performing temporal reasoning requires this date be grounded, typically to the date the report was written. However, if the text is copied to a later report, the ability to ground this date is lost. We propose a method called DupLink, which links pasted text back to its original source (Figure 1). We then utilize regular expressions to identify relative dates in pasted text, enabling such cases to be flagged or even automatically corrected.

Methods

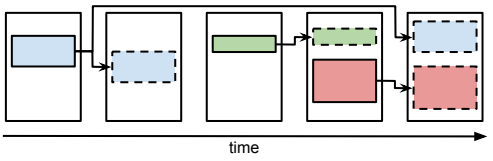


Figure 1: DupLink

DupLink recognizes copy/paste using Smith-Waterman local sequence alignment, a dynamic programming algorithm for detecting similar regions from two sequences. For each patient in MIMIC-II[3], reports are chronologically ordered, and Smith-Waterman is run pairwise to identify duplicate spans. A text span can act as the source (copy) to any number of targets (paste); a target can only have one source, creating the linking structure in Figure 1. DupLink iteratively finds the best local alignment on each report pair until no more high-scoring alignments are found. For each instance of copy/paste, regular expressions identify six classes of relative dates, best illustrated by example: (P1) *fourteen days ago*, (P2) *last year*, (P3) *on Saturday*, (P4) *yesterday*, (P5) *this July*, (P6) *45 years old*. To evaluate the severity of copying errors, we used five classes: (C1) *updated* in the pasted text, (C2) *valid*, as little time has passed, (C3) *resolvable* by a nearby absolute date, (C4) inconsistent, but *minor*, and (C5) inconsistent, and *serious*. Some analysis can be performed automatically (e.g., when reports are from the same day), otherwise human analysis is necessary. An experienced internist (AC) annotated a sample of each of the six pattern classes.

	matches	annotations	%C1	%C2	%C3	%C4	%C5
P1	156	142 (100)	-	53	26	20	2
P2	50	50 (7)	-	86	2	8	4
P3	47	47 (18)	11	85	4	-	-
P4	1278	731 (100)	6	62	17	12	1
P5	790	432 (100)	8	71	9	9	-
P6	388	199 (100)	-	100	-	-	-

Table 1: Results. Annotations in parenthesis are the manual annotations. Percentages are projections from manual and automatic annotations.

Results

See Table 1. Given the size of MIMIC, there were few instances of copy/paste. In other institutional cultures, however, copy/paste is more prevalent. Most copy/paste instances in MIMIC occur on the same day (C2). The resolvable class is the next most common, suggesting there are many relative dates that should be resolved to a local anchor instead of the report date. While few serious inconsistencies were found, they do exist along with many minor inconsistencies. This suggests the importance of integrating a copy/paste detection system like DupLink into the EHR. Further information as well as a software implementation of DupLink can be obtained by contacting the authors.

Acknowledgement This work was supported by the intramural research program at NLM/NIH.

References

1. Robert E. Hirschtick. Copy-and-Paste. *Journal of the American Medical Association*, 295(20):2335–2336.
2. Jesse O. Wrenn, Daniel M. Stein, Suzanne Bakken, and Peter D. Stetson. Quantifying clinical narrative redundancy in an electronic health record. *Journal of the American Medical Informatics Association*, 17:49–53, 2010.
3. DJ Scott, J Lee, I Silva, S Park, GB Moody, LA Celi, and RG Mark. Accessing the public MIMIC-II intensive care relational database for clinical research. *BMC Med Inform Decis Mak*, 13(9), 2013.

An International Evaluation of User Perceptions of Drug-Drug and Drug-Allergy Interaction Alerts

Alexandra Robertson[1], Pamela Neri, MS[2], Elisabeth Burdick MS[1], Sarah P. Slight MPharm, PhD, PGDip [1, 3], David W. Bates MD[1, 2, 4], Shobha Phansalkar RPh, PhD [1,4,5]

[1]Division of General Internal Medicine, Brigham and Women's Hospital, Boston, MA; [2]Partners HealthCare, Wellesley, MA; [3]School of Medicine, Pharmacy and Health, University of Durham, UK; [4]Harvard Medical School, Boston, MA; [5]Wolters Kluwer Health, Indianapolis, IN.

Background

Electronic medical records (EMRs) with integrated CDS have the potential to improve medication safety. Increasing adoption, refining the delivery and content of existing CDS, and improving user-centered design are key to meeting core measures for Meaningful Use. However, only limited evidence supports a direct link between CDS and measurable improvements in patient safety. Poor user acceptance of alerts, poor alert design, and lack of contextual specificity have been cited as causes for negative perceptions about the utility of CDS, alert fatigue, and override rates which are estimated to be as high as 49-96%.

Methods

To understand the relationship between alert-acceptance and user perceptions of the number and content of alerts, we surveyed physician perceptions of DDI and DAI alerts via a validated survey tool developed by Zheng, et al.[1] We surveyed users of 4 home-grown and 2 vendor systems across 7 multi-national institutions. 1,423 internal medicine physicians were invited to participate. To assess the impact of provider perceptions of the volume of alerts they receive, providers were asked to quantify the DDI/DAI alerts they: (a) received in an average week, (b) read thoroughly, (c) found clinically relevant, (d) felt changed their prescribing behavior, and (e) overrode.

We performed descriptive statistics of responses to establish correlations between perceptions of alert frequency, alert relevancy, and alert override. We ran means in 3 groups based on the number of alerts reported per week; 1-10 alerts/week (group 1), 11-50 (group 2), and greater than 50 (group 3).

Results

Of the 1,423 physicians invited, 342 consented to participate for a response rate of 24%. Overall, participants estimated receiving a greater number of DDI (22.9) than DAI (13.8) alerts per week, but were more likely to override DAI than DDI alerts, with reported override rates of 83.22% and 78.5%, respectively. For both DAI and DDI alerts across all 3 groups, we found that as the number of perceived alerts increases, the percentage of providers who report reading, finding these alerts relevant, or changing prescribing behaviors, based on the information provided, decreases, while the number of alerts overridden increases.

Physicians who reported receiving greater than 50 DDI/DAI alerts per week also reported reading only 25.4/30.5%, finding only 25.5/7.8% clinically relevant, identifying only 15.2/16.5% as changing their prescribing behavior, and reported overriding 90.4/96.9%, respectively. Comparatively, those who reported receiving between 1-10 DDI/DAI alerts per week reported: reading 69.6/82.7% of those alerts, finding 40.7/55.1% clinically relevant, identified 34.7/48.5% as changing their prescribing behavior, and overrode 64.2/68.5%, respectively.

Conclusions

This is the first study to establish an empirical correlation between physician perceptions of alerts and alert acceptance. Physicians who believe they receive a high number of alerts are less likely to read them, find them clinically relevant, allow them to affect their prescribing behavior, and more likely to override them. Future research should focus on how perceptions of alert volume can be improved. Decreasing the volume of interruptive alerts may foster more positive physician attitudes towards CDS alerting in EMRs.

References

1. Zheng K, Fear K, Chaffee BW, Zimmerman CR, et. al. Development and validation of a survey instrument for assessing prescribers' perception of computerized drug-drug interaction alerts. J Am Med Inform Assoc 2011; **18**:i51-i61.

This study was funded by grant # U19HS021094 from the Agency for Healthcare Research and Quality (AHRQ).

The importance of mental models in the design of integrated PHRs

Inês Rodolfo, BA[1]; Liliana Laranjo, MD, MPH[2]; Nuno Correia, PhD[1]; Carlos Duarte, PhD[3]
[1]Faculty of Science and Technology, UNL, Lisbon, Portugal; [2]Portuguese School of Public Health, UNL, Lisbon, Portugal; [3]Faculty of Science, UL, Lisbon, Portugal

Abstract
In order to address the challenges of designing a national integrated personal health record (PHR), we conducted remote software-based individual and closed card sorting sessions to detail mental models for the creation of a meaningful information architecture. We were able to compare how people from different user groups (senior people, younger people, healthcare professionals and non-professionals) think about PHR-related content

Environment relevance and problem
Several government programs around the world are implementing healthcare system reforms with the goal of creating sustainable networks that combine interoperable health information collected both from healthcare providers and patients. In this context, integrated PHRs have emerged. The main difference from stand-alone and tethered PHRs is that this model has the ability to incorporate patient self-reported information with data from multiple health-related sources [1], providing holistic views of each patient's health. Our research is focused on a design proposal for our National Patient Portal in the scope of the European Patients Smart Open Services project. The platform combines this new PHR model counting over 750, 000 registered users who can access their unique summary care record and use both healthcare services and self-tracking tools to manage their health. Past research revealed the need to detail mental models of lay individuals and providers in order to achieve comprehensive system information architectures that reflect the patient's roles and needs [2]. This need inspired the goal of this study: to create content and navigation structures that respond to the involved user profiles: providers who need to access patient's data and the citizens who use the platform (including the difference between younger and older adults).

Methods
For the portal redesign we followed a user experience design approach to gain more understanding of users, combining functional analysis with other emotional and social aspects of people's relations with technology. To detail mental models we applied card-sorting, an evaluative straightforward method that helped us understand and compare how people think, label, and organize content in a way that makes sense to them [3]. We applied closed card sorting because of the complexity of health content. Through this method we evaluated the first two navigation levels of the system information architecture (IA), developed from a previous extensive literature review and functionality gathering process. We had a total of 73 participants representing two user groups as a result of two different participation environments (a senior university and through social networking): a senior user group with whom moderated sessions were held in a senior university with 11 participants (average age: 64); and a broader audience group (average age: 33) with whom remote sessions were conducted during 4 days with 62 participants.

Results and conclusions
The results influenced significantly the final structure of the IA. Both first and second navigation levels suffered changes in the given category labels and content grouping. The first level groups that worked best for both user groups were "Personal information", "My Health" and "Health Education". The groups that presented more conflicts and doubts for both user groups were "Healthcare Management" and "Healthcare System". As expected, healthcare professionals were the ones who showed more agreement with the initial IA, as they are more aware of how the healthcare services work and more comfortable with clinical content, in comparison with lay people. We did not find significant differences between the mental models of senior people and other demographic groups, except regarding 'Privacy Control' content. This is probably due to the fact that senior people are less used to this type of content when compared to younger people, who tend to have more software experience. We also considered participants' suggestions and comments in the IA restructure: e.g., to divide "My Health" group into two other sub-levels ("Health History" and "Basic Information") as this group was considered to be too overwhelming.

References

1. Detmer, D.E., Bloomrosen, M., Raymond, B. and Tang, P. 2008. Integrated personal health records: Transformative tools for consumer-centric care. *BMC Medical Informatics and Decision Making*. 8,1(2008), 45.
2. Kaelber, D.C.D., Jha, A.K.A., Johnston, D.D., Middleton, B.B. and Bates, D.W.D. 2008. A research agenda for personal health records (PHRs). *JAMIA*. 15, 6 (Aug. 2008), 729–736.
3. Spencer, D. 2011. *Card Sorting*. O'Reilly Media, Inc.

Automatic coding of Free-Text Medication Data recorded by Research Coordinators

Laritza M. Rodriguez, MD, PhD, Vojtech Huser, MD, PhD, Olivier Bodenreider, MD, PhD, James J. Cimino, MD, PhD

Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health Clinical Center, National Institutes of Health, Bethesda, MD

Abstract

Medication reconciliation and clinician-friendly recording of medication is an ongoing informatics challenge. Users often prefer unrestricted free-text entry. We sought to see if simple Natural Language Processing (NLP) methods can be used to convert real-life medication records into coded concepts without additional interface changes to clinicians. We used data from the National Institutes of Health Clinical Center (NIH CC). Outpatient medications are recorded by research nurses using a semi-structured form that does not alert for incorrect data entry. We present data on accuracy of the NLP pipeline and suggest improvements to RxNorm tools.

Objective: To utilize simplified automatic NLP tools for drug entry harmonization.

Materials and Methods: We used National Library of Medicine's RxNorm API through the RxMix tool (http://rxnav.nlm.nih.gov/, accessed on 2014-01-30) to convert free-text medication entries into RxNorm concepts, namely RxNorm ingredients**.** Medication free text entries were first preprocessed with regular expressions to extract the drug names, drug strength, and drug forms from the original string that also included dosing and frequency information. The resulting simplified string was processed using the function *getApproxMatch* to find matching concepts in RxNorm (RxCUIs). The RxCUI was then processed using the *getAllRelatedInfo* function to obtain the ingredient and additional information. We used the concept match score and rank provided by the *getApproxMatch* function to further consider only well-matched concepts. For example, from the original input string: "*Metronidazole 250 mg , 2 tablet , PO , Every 12 hours.*", we first extract the drug entity "*metronidazole 250 mg*", which maps to the RxNorm concept *"Metronidazole 250 MG"* (316300) with a score of 100. Its ingredient is *"Metronidazole"* (6922). The remainder of the input string drug dose and frequency can be used to estimate drug doses or as part of a free text drug entry system automatic fill option.

Results: The original 43,303 medication strings were preprocessed into distinct 9,466 drug input strings. A total of 5,680 (60%) strings mapped to either RxNorm Ingredient (IN) concept or Clinical Drug (SCD) or both. Online appendix at http://dx.doi.org/10.6084/m9.figshare.960060 shows complete numerical results together with the impact on the string mapping counts when score thresholds are used to classify strings into two scenarios: (A) likely correction to a single target concept (score 80-100; single misspelled character: *"atenlol"* to *"atenolol"*) or (B) nurse may be asked to pick the best term from several matching contexts (score 50-79; multiple misspellings). A total of 502 concepts had either no matches or a score of less than 50 most of which are not ingredients or meaningless strings eg: *"vitamin juice", "zyrec"*.

Evaluation: The evaluation of the automatic mapping was done by one of the authors (clinician LMR). The unmatched set of strings was analyzed to detect those that should have found a match and did not. Two mapping errors were found: ***"codeine"*** and ***"chlorpheniramine"*** mapped to two unrelated ingredient names with the highest score of 100. (These errors were reported to and corrected by the RxMix team). Overall, the method had a 0.99 precision, 0.97 recall and 0.98 F1-measure. The manual evaluation was performed by only one clinician.

Conclusions: The results demonstrate that simple NLP methods that use the existing free RxNorm API can be used to convert free text medications records into RxNorm concepts. Coded concepts offer researchers better data queries that can utilize drug class hierarchies and outperform the existing free-text search. The quality of the mapping using the automated string matching is excellent provided the total number of input strings with only two mapping errors (employing also the above score filters). Although the approximate matching function of RxMix was primarily developed to support matching of clinical drug names, not ingredients alone, we found it useful in the context of this experiment where our input data had various degrees of specification.

References

1. Zhou L, Plasek JM, Mahoney LM, Karipineni N, Chang F, Yan X, et al. Using Medical Text Extraction, Reasoning and Mapping System (MTERMS) to process medication information in outpatient clinical notes. AMIA Annu Symp Proc AMIA Symp AMIA Symp. 2011;2011:1639–48.

Implementation of a mobile communication application to enable secure electronic access and exchange of patient health information.

Taisha Y. Roman, MD, MPH, Allen Hsiao, MD, FAAP
Yale School of Medicine, New Haven, CT

Introduction

Alphanumeric pagers are the preferred mode of communication in many institutions because they allow providers to communicate pertinent patient information quickly over a secure hospital network.[1] However, these pagers do not satisfy current security recommendations for mobile devices used to transmit personal health information (PHI). The current recommendations call for the use of data encryption, user authentication, and remote disabling features.[2] Studies have also shown that the use of pagers can lead to medical errors if relayed lab results are erroneously reported.[3] In an attempt to ensure accurate and secure exchange of personal health information, we have piloted a mobile communication application that enables efficient communication between physicians and allows mobile access to important patient information such as lab results.

Methods

A secure mobile communication application (Mobile Heartbeat,™ Woburn, MA) was implemented in the pediatric and adult emergency rooms. The application was placed on preprogrammed mobile phones with passcode protection, remote disabling capability, and two-way encryption at rest and in transit. Shared devices were placed in a common area in each department. Providers were assigned a device at the start of their shift and returned it when the shift was completed. Anonymous pre and post implementation surveys were used to evaluate the application.

Results

A total of 60 providers completed pre and post implementation surveys. The implementation of Mobile Heartbeat decreased the use of unsecure modes of communication. Pager use during shifts decreased from 53% to 38% and the use of personal cellphones decreased from 22% to 10%. The application also improved communication, making it easier to contact providers within the department (24% found it easy prior to implementation vs 75% post implementation) and outside of the department (16% found it easy prior to implementation vs 45% post implementation). The device was well received by providers; 77% of users believed that mobile heartbeat improved workflow and 77% believed that it improved patient safety.

Discussion

Although alphanumeric paging is an efficient means of communication between physicians, studies have shown that it can pose a risk to patient privacy and patient safety.[3] Current mobile technology enables secure, efficient exchange of personal health information without compromising patient care. Our institution implemented a secure mobile communication application in the pediatric and adult emergency departments in order to reduce unsecure exchange of PHI and improve communication among providers. This application decreased the use of unsecure methods of communication such as personal cell phones and alphanumeric pagers. There was also a perceived improvement in communication, workflow, and patient safety. The results of our pilot show that secure mobile communication technology is a viable alternative to the alphanumeric pagers currently being used in most healthcare institutions.

References

1. Wong BM, Quan S, Shadowitz S, Etchells E. Implementation and evaluation of an alpha-numeric paging system on a resident inpatient teaching service. J Hosp Med. 2009 Oct;4(8):E34-40
2. http://www.healthit.gov/providers-professionals/how-can-you-protect-and-secure-health-information-when-using-mobile-device
3. Espino S, Cox D, Kaplan B. Alphanumeric paging: a potential source of problems in patient care and communication. J Surg Educ. 2011 Nov-Dec;68(6):447-51

Lessons Learned Bringing Public Health into the Primary Care Clinic through an EHR-based Application

Caryn Roth, MPH[1], Randi E. Foraker, PhD[2], Marcelo A. Lopetegui, MD, MS[1], Philip R.O Payne, PhD[1]

[1]Department of Biomedical Informatics, College of Medicine [2]Division of Epidemiology, College of Public Health, The Ohio State University Columbus, OH

Abstract

We developed an application embedded in the EHR to incorporate prevention guidelines into primary care clinics. The dynamic visualization facilitates patient-provider communication and motivates behavior change to improve health. We discuss the development and usability challenges involved in deploying such a tool, as well as strategies to overcome these issues.

Introduction

Chronic diseases disproportionally burden the healthcare system, and in order to tackle them successfully, we must shift our focus to disease prevention, rather than treatment. We aimed to develop and validate a personalized and interactive health visualization module in order to enhance communication between healthcare providers and patients to promote cardiovascular health (CVH) and prevent or manage cardiovascular disease.

Methods

By leveraging data visualizations and relying on tight integration with our commercial electronic health record (EHR), we designed SPHERE (**S**troke **P**revention in **H**ealthcare Delivery **E**nvi**R**onm**E**nts), a data-driven risk assessment instrument that enables seamless point-of-care interactive risk profiling for primary care physicians and their patients. We employed a composite CVH score, modified from the AHA's *Life's Simple 7*™ prevention campaign, and designed an interactive visualization tool to engage patients and improve patient-clinician communication. We piloted SPHERE among academic medical center primary care providers serving an urban, low socio-economic status patient population and catalogued strategies and solutions to successfully deploy such a tool.

Results

The final tool met all the specified requirements: we created an EHR-integrated application that leverages existing clinical data to provide a real-time representation of the patient's CVH status, and promotes discussion between healthcare providers and patients. Moreover, the interactivity feature provides immediate feedback about the impact of individual behavior changes on CVH.

Discussion

Automating health profiling and displaying visual cues to patients and physicians within the EHR may promote behavior change and improve cardiovascular health across our entire patient population. The informatics and public health lessons we uncovered while evaluating and refining SPHERE are critical to the success of such interventions. In addition to modifying the tool to be clinically relevant for our physicians, we uncovered issues relating to data capture in the EHR, technical differences across our academic medical center, and usability and workflow obstacles. While challenging, harmonizing clinical practice, public health, and information systems for such purposes can truly impact healthcare delivery and patient outcomes. Our tool is one example of how we can harness the power of the EHR to bring evidence-based public health interventions into primary care clinics, reduce risk for chronic disease and conserve valuable healthcare resources.

The Use of Multiple Emergency Department Reports per Visit for Improving the Accuracy of Influenza Case Detection

Victor M. Ruiz, BS[1], Ye Ye, MS[1], Amie J. Draper, BS[1], Fuchiang (Rich) Tsui, PhD[1]
[1]Real-time Outbreak and Disease Surveillance Laboratory, Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA

Abstract*: This study assessed the diagnostic accuracy of probabilistic Bayesian models that detect influenza from one or multiple emergency department (ED) reports per visit. We extracted clinical findings from ED reports using two natural language processing (NLP) tools. We measured and compared the area under the ROC (AUROC) curve of existing models applied to single and multiple ED reports per visit respectively, and found that using multiple reports per visit increased the detection performance significantly (p<0.01).*

Introduction: ED reports document the care received by patients in the ED, and contain clinically relevant information including medical history, medications, clinical findings, and ED course. Recent studies demonstrated that ED reports improve influenza detection performance over chief complaints.[1] Those studies used one ED report per visit. However, in ED practice, multiple reports exist for one visit due to multiple clinicians (e.g., residents, attending physicians) seeing one patient and/or one clinician dictating/writing multiple (supplemental) reports. To the best of our knowledge, no previous research studied the impact of using single vs multiple ED reports per visit on the performance of influenza detection. We hypothesize that the use of multiple ED reports per visit improves the accuracy of influenza detection through the collection of more clinical findings across multiple reports over time.

Methods: We used the Topaz[2] and MedLEE[3] NLP tools to extract 31 clinical findings from free-text ED reports, retrieved from electronic health records of 11 hospitals in Allegheny County, Pennsylvania. Those findings were selected by board-certified physicians as relevant for the diagnosis of influenza, and served as input to existing Bayesian network (BN) models for influenza detection.[4] These models share a static, expert-defined structure and were trained using one report per visit. They differ in the NLP tool used to extract clinical findings from training data.

When incorporating data from multiple ED reports per visit, it is possible to find contradicting mentions of clinical findings. We used two rules to resolve conflicts and evaluated the performance of the BN models for each rule. First, we assigned a 'True' value to findings with at least one positive mention. Alternatively, we used the value of the most recent mention of each finding.

To assess the contribution of using multiple ED reports per visit for the detection of influenza, we compared the AUROC of the BN models using the first available ED report for each visit to the AUROC of the same models using multiple reports. The testing dataset included 131 polymerase chain reaction (PCR)-confirmed influenza cases and 1179 randomly selected controls, from January 1, 2011 to June 30, 2013.

Results: BN models achieved the highest accuracy using the most recent finding values as conflict resolution rule, independently of the selection of NLP for finding extraction. With MedLEE, the AUROC increased to 0.90 (95% CI 0.87-0.93) from 0.86 (95% CI 0.83-0.90) when using multiple vs single ED reports per visit ($p < 0.01$). With Topaz, The AUROC increased to 0.91(95% CI 89-0.93) from 0.88 (95% CI 0.85-0.91), with $p < 0.01$.

Conclusion: The contribution of using multiple ED reports per visit to the diagnostic accuracy of influenza detection models was statistically significant. This indicates that updating the influenza probability estimation with the latest data can improve the accuracy of case detection systems (CDSs). Future work will assess if diagnostic accuracy can be further improved by using additional findings.

References

1. Elkin PL, Froehling DA, Wahner-Roedler DL, Brown SH, Bailey KR. Comparison of natural language processing biosurveillance methods for identifying influenza from encounter notes. Ann Inter Med. 2012;156(1 Pt 1):11-8.
2. Harkema H, Dowling JN, Thornblade T, Chapman WW. ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports. J Biomed Inform. 2009;42(5):839-51.
3. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc : JAMIA. 2004;11(5):392-402.
4. Ye Y, Tsui FR, Wagner M, Espino JU, Li Q. Influenza detection from emergency department reports using natural language processing and Bayesian network classifiers. J Am Med Inform Assoc : JAMIA. 2014.

Authors:

- **Amer Saati, MD, MS** Clinical Informatics Specialist at North Shore LIJ Health System, Lake Success, NY 11042
- **Michael Oppenheim, MD** Chief Medical Informatics Officer at North Shore LIJ Health System, Lake Success, NY 11042
- **Eytan Behiri, MD** Director of CIS Evidence Based Medicine and Order Sets at North Shore LIJ Health System, Lake Success, NY 11042

Background:

Adoption of Computerized provider order entry (CPOE) with embedded clinical decision support (CDS) has been a promising vehicle to improve quality and consistency of care, better cognitive workload, reduce errors, increase adherence to evidence-based knowledge guidelines. Clinical decision support (CDS) includes a group of computerized functions such as reminders, alerts that provide knowledge to clinicians to enhance health care. A mainstay of CPOE systems are order sets, a grouping of orders, related to a specific diagnosis or symptom, used by providers to create orders that are clinically related.

In our study, we chose to look at a unique CDS scenario that occurred in our Electronic Medical Record (EMR) system that allowed us to compare 2 types of CDS alerts ("hard" stop and "smart" stop alerts) in the same clinical workflow situation. As we went live with the EMR system, we embedded a Venous Thromboembolism (VTE) prophylaxis order set as a subset (VTE Subset) of ALL inpatient order sets in CPOE. In order to make sure a VTE order was selected, the system had a hard stop that required providers to order from the VTE Subset. Physicians could not proceed in their orders until they addressed VTE prophylaxis, even if documented previously, i.e. a provider might encounter several such hard stops, even if the patient had an active order for VTE Prophylaxis. At a later stage we changed the alert to a "smart" stop. The VTE Subset was still linked to all inpatient order sets, but no longer required a mandatory stop. It was based on a rule engine algorithm that only alerts providers of missing VTE prophylaxis on exiting an order set or on entry to the CPOE module. We hypothesized that the smart stop subset would influence prescribers to use order sets embedded in CPOE in greater numbers and in addition to gain better adherence to VTE prophylaxis guidelines as opposed to the hard stop subset.

Methods:

Study design and Locations: This is a retrospective, quasi experiment one group pre post study conducted on all inpatient admissions in two major tertiary hospitals in North Shore LIJ Health System: Long Island Jewish Hospital in Lake Success, New York and North Shore University Hospital in Manhasset, New York. Long Island Jewish Hospital: 63,000 admission With T1 (Hard stop VTE subset) from 04/11/2011 to 02/28/2012 and T2 (smart stop VTE subset) from 04/11/2012 to 02/28/2013; North Shore University Hospital:33,000 admissions With T1 (Hard stop VTE subset) from 10/15/2011 to 02/28/2012 and T2 (smart stop VTE subset) from 10/15/2012 to 02/28/2013.

Variable: The main outcome is to determine the difference in number of times VTE prophylaxis was utilized among all hospitalized patients in two tertiary hospitals using a smart stop VTE in comparison to a hard stop VTE subset. The secondary outcome was to measure the difference in number of times VTE prophylaxis was utilized using VTE diagnostic Order set or subset vs. individual orders from main order catalog among the same population in order to evaluate CPOE efficiency. Compliance criteria were met if at least one VTE prophylaxis item was addressed from the VTE list.

Data Collection and Statistical Analysis: Queries in inpatient database (Sunrise Clinical Manager) were performed and Data were exported to a spreadsheet and imported to SPSS for analysis. Chi square test was conducted to compare the differences to measure primary and secondary outcomes. Other variables were analyzed such as age, gender, medical unit, time from admission and time before discharge in addition to distinguishing type of VTE prophylaxis (mechanical vs. medication) and source of VTE utilization.

Results:

A total number of 63,000, 33,000 patients were assessed retrospectively in Long Island Jewish Hospital and North Shore University Hospital in two different time frames as mentioned in methodology. After implementing the smart stop VTE subset, over all adherence to VTE improved in LIJ and NSUH from 85.2 % to 94.9 % (odds ratio [OR], 3.17; 95% CI, 2.993-3.365; P<.001) and from 84.7 % to 96.6 % (odds ratio [OR], 5.17; 95% CI, 4.700-5.684; P<.001) respectively. Using Chi Square in both facilities, we proved that there is an association between implementing a smart VTE subset and increasing number of VTE orders for hospitalized patients prior to discharge. .X^2 (2, N = 62608) = 1781.556, p <0.05 in LIJ and X^2(2, N = 33148) = 1480.638, p <0.05 in NSUH. Furthermore, Patients in the Post group has statistically significant higher probability for VTE utilization via main order set as opposed to using individual VTE orders with Odds ratio >1.

Conclusion:

Using smart stop VTE subset is associated with increase compliance to VTE guidelines in comparison to using hard stop VTE subset. Also, that increase in VTE orders was associated with ordering from Order set catalog as opposed to individual catalog items and that association was statistically significant.

Feasibility of the SBAR Discharge Summary Format

Farrant H. Sakaguchi, MD, MS[1,2], Leslie A. Lenert, MD, MS[3]
[1]University of Utah, Salt Lake City, UT; [2]Intermountain Healthcare, Murray, UT; [3]Medical University of South Carolina, Charleston, SC

Abstract

To improve the clinical handover at hospital discharge, we have proposed reformatting the discharge summary using the SBAR mnemonic. We tested the feasibility of our template with five residents. They found the template relatively easy to use and agreed that the template helped focus on the clinical handoff.

Introduction

The ability of the discharge summary (DCS) to promote continuity in care may be limited by its structure of summarizing previous care without focusing on transmitting information needed for future care. Clinicians are frustrated by more extensive documentation for non-clinical purposes. The increasing prevalence and capabilities of electronic health records (EHRs) may facilitate the completeness and standardization of documentation but at the risk of losing the pithiest communication that clinicians often value; there are concerns that the expansion of electronic documentation may obfuscate clinical communication.

The Situation-Background-Assessment-Recommendation (SBAR) structure for handoff's is common in healthcare and has facilitated the sharing of clinical contexts, clarified roles and expectations, and increased critical thinking and analysis. We proposed reformatting the DCS using SBAR to shift the paradigm away from being a historical "Captain's Log" towards being a handoff document.[1]

Methods

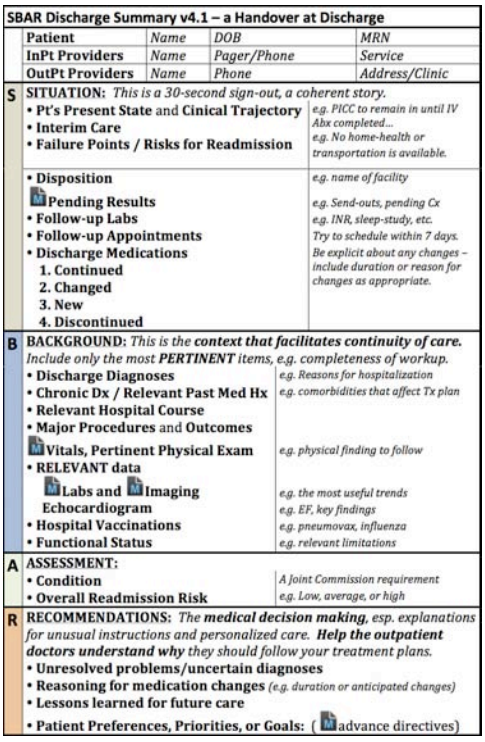
We created an SBAR-DCS summary template (Figure 1) through an iterative process with a panel of five practicing physicians. The M icons refer to information specifically available in the EHR. Internal medicine residents were recruited to evaluate the feasibility of using a paper, pocket-sized SBAR-DCS template while dictating. After each dictation, they completed a brief questionnaire.

Results

Five residents created a total of 17 DCS using the SBAR-DCS template. They found the template relatively easy to use even though the template took 20 minutes to use instead of the estimated usual 10 minutes. However, in all discharges (17/17), they agreed that the template helped focus on the handoff. The template frequently reminded them of things otherwise likely forgotten (16/17 discharges), and emphasized the importance of lessons-learned for the follow-up provider (15/17 discharges).

Discussion

Outpatient physicians felt that several of the DCS based on the SBAR-DCS template were more useful than those they typically received. In one instance, an emergency room physician changed plans for readmitting a patient based on the insights available in the SBAR-DCS. Next steps will quantify the clarity of the handover from the recipient's perspective and the feasibility of an embedded, semi-automated SBAR-DCS template.



SBAR Discharge Summary v4.1 – a Handover at Discharge				
	Patient	Name	DOB	MRN
	InPt Providers	Name	Pager/Phone	Service
	OutPt Providers	Name	Phone	Address/Clinic
S	SITUATION: *This is a 30-second sign-out, a coherent story.*			
	• Pt's Present State and Cinical Trajectory • Interim Care • Failure Points / Risks for Readmission	*e.g. PICC to remain in until IV Abx completed...* *e.g. No home-health or transportation is available.*		
	• Disposition	*e.g. name of facility*		
	M Pending Results	*e.g. Send-outs, pending Cx*		
	• Follow-up Labs	*e.g. INR, sleep-study, etc.*		
	• Follow-up Appointments	*Try to schedule within 7 days.*		
	• Discharge Medications 1. Continued 2. Changed 3. New 4. Discontinued	*Be explicit about any changes – include duration or reason for changes as appropriate.*		
B	BACKGROUND: *This is the* ***context that facilitates continuity of care.*** *Include only the most* ***PERTINENT*** *items, e.g. completeness of workup.*			
	• Discharge Diagnoses	*e.g. Reasons for hospitalization*		
	• Chronic Dx / Relevant Past Med Hx	*e.g. comorbidities that affect Tx plan*		
	• Relevant Hospital Course			
	• Major Procedures and Outcomes			
	M Vitals, Pertinent Physical Exam	*e.g. physical finding to follow*		
	• RELEVANT data			
	M Labs and M Imaging	*e.g. the most useful trends*		
	Echocardiogram	*e.g. EF, key findings*		
	• Hospital Vaccinations	*e.g. pneumovax, influenza*		
	• Functional Status	*e.g. relevant limitations*		
A	ASSESSMENT:			
	• Condition	*A Joint Commission requirement*		
	• Overall Readmission Risk	*e.g. Low, average, or high*		
R	RECOMMENDATIONS: *The* ***medical decision making****, esp. explanations for unusual instructions and personalized care.* ***Help the outpatient doctors understand why*** *they should follow your treatment plans.*			
	• Unresolved problems/uncertain diagnoses			
	• Reasoning for medication changes *(e.g. duration or anticipated changes)*			
	• Lessons learned for future care			
	• Patient Preferences, Priorities, or Goals: (M advance directives)			

Figure 1: SBAR-DCS Template

References

1. Lenert, L. A., Sakaguchi, F. H. & Weir, C. R. Rethinking the discharge summary: a focus on handoff communication. *Acad Med* **89,** 393–398 (2014).

Acknowledgement: Funding provided by National Library of Medicine Training Grant T15LM007124

The Establishment of Health Informatics Laboratory for Specialized Wireless Remote Monitoring Training and R&D

Hasan Sapci, MD [1], Aylin Sapci, MD
1. Adelphi University, Garden City, NY

The American Board of Medical Specialties approved clinical informatics as a board certified subspecialty in 2011.[1] Academic institutions have been launching new medical informatics programs and adding new courses in their curriculum over the last decade and according to the Commission on Accreditation for Health Informatics and Information Management Education (CAHIIM), there are more than 100 accredited associated, baccalaureate and master level health/medical informatics programs in the United States.[2] American Telemedicine Association states that there are about 200 telemedicine networks in the United States and one million patients have been using remote monitoring devices. Considering the fact that this number reflects just one medical specialty which is cardiology, the importance of structured educational programs that provide hands-on remote patient monitoring and telemedicine skills is apparent.[3]

For a long time students, clinicians and information technology professionals have been trying to follow-up changing trends by themselves, either by forming special interest groups or traveling to other states in order to learn these specific skills from their colleagues' experience. Because health/medical informatics is a quite new field, professional organizations and institutions have been re-defining the core content recently. Competencies required to design and implementation of telemedicine systems, and to monitor, transfer and analyze data are still not listed in most health informatics programs' curricula.[4]

Traditional informatics training methods has various limitations and most students acquire these skills and knowledge from books or lectures only. They do not have access to data-driven healthcare applications, medical systems and devices. Practical hands-on skills are crucial to design, evaluate and implement state-of-the art telemedicine and remote monitoring systems.

We designed a new curriculum and established a health informatics laboratory with U.S. Department of Health's support in order to provide access to specific remote monitoring software and hardware such as medical scopes, exam cameras and vital sign monitors. The authors designed a replicable training model for academic institutions, received IRB approval and conducted pre and post training surveys on 60 graduate level health informatics students. This quantitative research compares traditional health informatics education with the proposed new problem-based learning model and focuses on different sophisticated modules that complement each other.

This presentation will describe a new methodology for the establishment of a health informatics laboratory, share the results of this research and propose a state-of-the art model for remote patient monitoring, and telemedicine training curriculum.

REFERENCES

1. American Board of Medical Specialties News Release. Retrieved from http://www.abms.org/News_and_Events/news_archive/release_Announcing_TwoNewSubspecialties_10312011.aspx on 08.01.2014.

2. CAHIIM Accredited Programs Directory. Retrieved from http://www.cahiim.org/accredpgms.asp on 08.01.2014

3. American Telemedicine Association, Telemedicine Frequently Asked Questions. Retrieved from http://www.americantelemed.org/about-telemedicine/faqs#.U9vlAmN6fxI on 08.01.2014

4. Gardner, R. M., Overhage, J. M., Steen E. B., Munger, M. S., Holmes, J. H., Williamson, J. J., & Detmer, D. E. (2009). Core Content for the Subspecialty of Clinical Informatics. *Journal of the American Medical Informatics Association*, 16, 153-157.

The role of HMG-CoA reductase inhibitors in the management of sepsis in a cohort study of adult intensive care unit patients

Raymond Francis Sarmiento, Paul Fontelo
Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Bethesda, MD

Abstract
Recently, HMG-CoA reductase inhibitors, better known as statins, have been suggested as an element in the management of sepsis patients in the ICU because of its immunomodulating properties. We found in this study of the MIMIC-II database that statins are associated with reduced mortality in septic shock ICU patients but not in sepsis and severe sepsis.

Introduction
In the last 40 years, the number of sepsis cases has increased to 350% in the United States with 750,000 severe sepsis cases occurring annually.[1] Adjunct therapies have been found to modulate the marked inflammatory response found in sepsis patients.[2] Observational studies suggest that statins possess anti-inflammatory and antioxidant properties that may reduce the body's inflammatory response.[3] Previous statin use has also been suggested as protective against sepsis and stopping statins during severe sepsis might be deleterious for prior users. However, a multicenter randomized trial of atorvastatin therapy on severe sepsis patients in the ICU found that only prior, not *de novo*, statin use was associated with improved survival. Our objective in this study was to examine the relation between statin use and mortality among sepsis patients admitted into the ICU using the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC-II) database.

Methods
To extract our dataset, we performed SQL queries on the MIMIC-II version 2.6 database, which contains 32,425 unique patients. We excluded 7,919 non-adult patients from the analyses and extracted cohort groups of sepsis, severe sepsis, and septic shock patients based on ICD-9 codes on the remaining 24,506 adult patients. We removed duplicate records, giving a final tally of 2,203 patients (477 sepsis only, 739 severe sepsis, and 987 septic shock patients). We analyzed the data by considering demographic as well as clinical covariates based on clinical knowledge and existing literature on sepsis and critically ill patients in the ICU.

Results
We developed models for analysis and ran the following covariates in our final logistic regression model: male gender, ethnicity, comorbidities (hypertension, diabetes, alcoholism, chronic kidney disease, COPD, and cancer), ICU interventions (dialysis, insulin, transfusion, total parenteral nutrition), and statin use. We found that even after adjusting for multiple potential risk factors, statin use seemed to have a beneficial effect on the survival of septic shock patients, but not on sepsis and severe sepsis. We found that septic shock patients had a lower mortality risk at 30 days and one year after ICU admission if given statins compared to sepsis only and severe sepsis patients.

Conclusion
This finding could provide more evidence on the role of statins in modulating the body's inflammatory response in extreme states of stress as in septic shock, although a similar response was not seen among sepsis only and severe sepsis patients. Further studies to investigate the effect of statins on acutely ill patients such as those in the ICU are warranted.

References

1. Martin GS, Mannino DM, Eaton S, Moss M. The epidemiology of sepsis in the United States from 1979 through 2000. N Engl J Med 2003; 348:1546-155.
2. Hotchkiss RS, Karl IE. The pathophysiology and treatment of sepsis. N Engl J Med 2003; 348:138-150.
3. Mekontso-Dessap A, Brun-Buisson C. Statins: the next step in adjuvant therapy for sepsis? Intensive Care Med 2006;6:242-248.

Customized Reference Ranges for Laboratory Values Decrease False Positive Alerts in Intensive Care Unit Patients

Christopher N. Schmickl, MD, MPH, Oguz Kilickaya MD, Ahmed Adil, MBBS, MSc, Juan Pulido, MD, James Onigkeit, MD, Kianoush Kashani, MD, Ognjen Gajic, MD MSc, Vitaly Herasevich, MD, PhD, Brian Pickering, MD
Multidisciplinary Epidemiology and Translational Research in Intensive Care (METRIC), Mayo Clinic, Rochester, MN

Abstract: Objective was to compare the accuracy of customized laboratory reference ranges (based on expert surveys) against standard reference ranges. Blinded to abnormal-value indicators, critical-care experts reviewed the laboratory results of 37 critically ill patients in the context of patients' problem list, judging about which of the values they would like to be alerted (gold standard). Customized reference ranges resulted in significantly less false positive results, thus possibly decreasing alert fatigue in health-care providers.

Background: Traditional electronic medical record (EMR) interfaces mark laboratory tests as abnormal based on standard reference ranges derived from healthy, middle-aged adults. This yields many false positive alerts with subsequent alert-fatigue when applied to complex populations like critically ill patients. Novel EMR interfaces using adjusted reference ranges customized for specific patient populations may ameliorate this problem. Our objective was to compare the accuracy of abnormal laboratory value indicators in a novel[1] vs traditional[2] EMR interface.

Methods: Laboratory data from intensive care unit (ICU) patients consecutively admitted during a two-day period were recorded. For each patient, available laboratory results and the problem list were sent to two mutually blinded board-certified critical care physicians, who marked the values about which they would like to be alerted. All disagreements were resolved by a third physician. Based on this gold standard, we calculated and compared the sensitivity, specificity, positive and negative predictive values (PPV, NPV) of customized vs traditional abnormal value indicators.

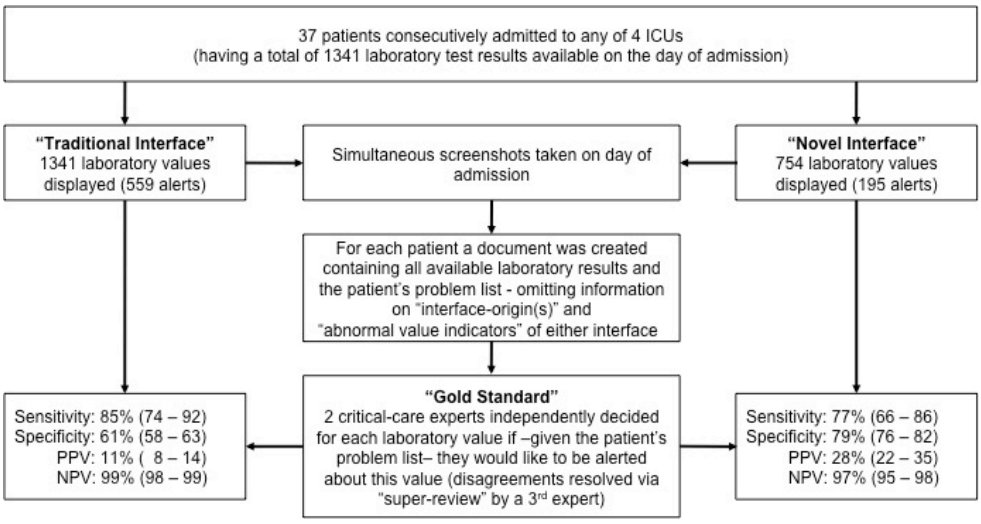


Figure 1. Studyflow. Results are displayed as estimate (95%-Confidence Interval).

Results: Thirty seven patients with a total of 1341 laboratory results were included (Figure 1).

Experts' agreement was fair (kappa=0.39). Compared to the traditional EMR, custom abnormal laboratory value indicators had similar sensitivity (77% vs 85%,P=0.22) and NPV (97.1% vs 98.6%,P=0.06) but higher specificity (79% vs 61%,P<0.001) and PPV (28% vs 11%,P<0.001).

Conclusions: Reference ranges for laboratory values customized for an ICU population decrease false positive alerts. Disagreement among clinicians about which laboratory values should be indicated as abnormal limits the development of customized reference ranges.

References

[1] Pickering BW, Herasevich V, Ahmed A, Gajic O. Novel Representation of Clinical Information in the ICU: Developing User Interfaces which Reduce Information Overload. Appl Clin Inform. 2010;1(2):116-31

[2] Carpenter PC. The electronic medical record: perspective from Mayo Clinic. Int J Biomed Comput. 1994 Jan;34(1-4):159-7

Reproducibility of Health Care Datasets

Daniel H Schneider[1], MS, Nathan D Sisterson[1], Luke V Rasmussen[1]
[1]Northwestern University, Chicago

Abstract
Healthcare data is being used in a wide array of statistical analyses and interpretations. Discrepancies in data points can lead an investigator to differing conclusions. We posit that in examining data collection methods by rerunning exact copies of executable data searches we will derive varying results which in some cases may be significant.

Introduction
With the growth of electronic health record data (EHR) available for institutional research as well as operations, reporting from these data has become routine. Since health care data is derived from many transactional systems this data is rarely static and has its own organic properties of modifications and alterations that occur over time. As a result, re-executing the same data searches over time will, intuitively, produce varying results. In addition, improvements made to data warehouses and integration efforts across systems can result in improved results.[1] The purpose of this study was to gain more understanding of both the magnitude and characteristics of such changes in results.

Methods
Ten previously defined datasets across multiple medical specialties were investigated to identify potential variation found within EHR data. In order to look back at the initial collection of these datasets we utilized an internal applications developed at the Northwestern Medicine Enterprise Data Warehouse (NMEDW.) This application (MyCohorts) allows a user to define and store patient identifiers to use for tracking subjects in a particular study. This allows us to reference a static snapshot of the dataset as existed when first created. These ten cohort definitions were then rerun to measure any difference in number of subjects meeting the eligibility criteria.

Results
The table shows that eight of ten queries did not produce stable results over time. In addition, the direction of change was not uniform, with both increases and decreases in patient counts over time. Three contributing factors were associated with a changing number of subjects identified. Poor query design, application updates, and source system data changes all contributed to the changing number of subjects. In one case, the placing of a date filter on an encounter date while failing to do so for a diagnosis date added subjects diagnosed after the query was initially run (cohort 1.) Deprecation of columns within a database as well as the removal of entire databases caused two of our queries to break completely (cohorts 2,9.) Examples of system data changing over time were also noted (cohorts 3,4,5,6,7,8,10.) Diagnoses were recoded, encounter types changed, and in some cases ages had been adjusted. All of these factors led to the changing number of subjects identified. The poster will examine the causes, and potential solutions for, these changes in more detail.

#	Cohort	Original Pt. Cnt.	New Pt. Cnt.	Difference
1	Outpatients with Meniere's Disease	341	495	154
2	Patients with Keloid Scar	386	0	-386
3	2009 Stem Cell or Bone Marrow Transplant Patients	231	231	0
4	All NMH Inpatients	121,056	104,958	-16,098
5	Patients with Aortic Stenosis and an Echo in 2010	390	389	-1
6	Patients with an HIT antibody test	1,310	1,291	-19
7	Spine Surgery Patients 2000-2010	2,598	2,598	0
8	NMH Patients with Hip or Knee Replacement	5,553	5,508	-45
9	Low Bone Mineral Density	247	0	-247
10	Sleep Study Patients	238	246	8

References

Strong DM, Lee YW, Wang RY. Data quality in context. Comm of the ACM 1997;40(5):103-110.

A Literature Review of Electronic Health Record Redesign for Optimization

Alissa A. Schultz, RN, MA, Albert M. Lai, PhD, Po-Yin Yen, RN, PhD
Department of Biomedical Informatics, The Ohio State University, Columbus, OH

Abstract

Meaningful Use has been a major driver behind most Electronic Health Record (EHR) implementations with mixed reviews of adoption and usability across disciplines, stakeholders and roles. Moving beyond initial installations, we wanted to learn how EHRs are being redesigned to optimize utilization for various performance outcomes given the Affordable Care Act and the healthcare reform. Our literature review revealed relatively few studies on the redesign of EHR systems that look at measurable outcomes of such initiatives.

Background

In 2009, the government offered monetary incentives for providers and institutions to implement Electronic Health Records (EHRs). Penalties and fines are being imposed when meaningful use requirements are not met. Most healthcare settings use EHRs to some extent, but perhaps not optimally. Much research exists on the usability and evaluation of EHRs, but a lack of standards within Health Information Technology make redesign difficult to study given the variability and dynamic nature of healthcare. Future focus will be on EHR optimization through redesign efforts that support patient centered evidence based care, continuous quality improvements, patient outcomes, and reducing overall health expenditures. The purpose of this study is to review the literature on EHR redesign efforts for optimization and measureable outcomes.

Methods

The literature search was performed primarily in two databases, PubMed and the ACM Digital Library. Search keywords included "(Electronic Medical Record OR Electronic Health Record OR EMR OR EHR) AND Redesign." Related MeSH terms were also used to search PubMed literature. Titles and abstracts were first screened for articles related to EHR redesign. Out of those included abstracts, we looked into full-text for further information.

Results

Searches in PubMed and the ACM Digital Library yielded 140 and 30 articles, respectively. In comparison, when swapping out the keyword "Evaluation" for "Redesign," there was approximately a 3000% increase in articles. Most studies identified through the literature review included a focus on process and environmental redesign (n=34), usability aspects (n=9), review or opinion based articles (n=54) and those that were not applicable (n=64). Only nine articles measured changes of an EHR redesign. Three studies measured patient outcomes, three focused on cost savings and three on clinical reminders incorporating evidence-based guidelines.

Discussion

EHR redesign opportunities exist and have the potential to facilitate significant impacts on the future of our healthcare system in terms of cost, quality, outcomes and accountability. Small changes in an EHR, such as unbundling lab orders for individual selection or adding one additional text box can drive significant cost savings. In addition, patient outcomes can be targeted for improvement by analyzing data to identify and prioritize efforts for improvement, such as redesigning order sets. Patient safety and quality of care should always be the first priority in EHR redesign efforts. The data used to identify redesign opportunities should also be utilized as outcome metrics to determine the success of redesign efforts.[1]

Conclusion

Relatively few redesign studies exist on EHRs. The focus must now change from EHR implementation to optimization via redesign efforts. Evidence based research is needed for guidelines and prioritization in respect to system enhancements. Limitations exist in conducting randomized controlled trials with an entire system.[2] However, individual redesigned system *features* within an EHR should be evaluated using randomized controlled trial study designs.[3]

References

1. Byrne MD. Redesign of electronic health records for perianesthesia nursing. *J Perianesth Nurs.* Jun 2013;28(3):163-168. doi: 110.1016/j.jopan.2013.1003.1004.
2. Harris AD, McGregor JC, Perencevich EN, et al. The use and interpretation of quasi-experimental studies in medical informatics. *Journal of the American Medical Informatics Association : JAMIA.* Jan-Feb 2006;13(1):16-23.
3. Podgurski A. Evidence-based validation and improvement of electronic health record systems. Paper presented at: Future of software engineering research; November 7-8, 2010, 2010; Santa Fe, New Mexico, USA.

Visualizing Design Trends by Mapping the Alzheimer's Disease Trial Space

Timothy Schultz, MS[1], Chaomei Chen, PhD[1]
[1]Drexel College of Computing and Informatics, Philadelphia, PA

Abstract

We propose to enable a system in which the massive space of clinical research itself can be mapped and queried from readily-available clinical data sources, through a scalable and distributed "Big Data" graph technology called Titan. By identifying the nature of and relationships between clustered areas of clinical research within and across therapeutic areas, it is proposed that future research trends can be extrapolated to further guide clinical trial design. We provide an example of these principles by modeling the Alzheimer's Disease clinical trial landscape.

Introduction

Disconcerting trends in clinical research have counter-intuitively shown that the failure to meet desired clinical trial endpoints has come at the cost of increased spending. This suggests the need for developing new ways in which clinical trials are designed, patients are recruited, and protocols are executed. One approach for addressing this issue is to provide research organizations a mechanism to proactively identify key clinical trial design patterns and trends within a therapeutic area through the utilization of novel visualization tools. At the heart of the clinical trial design lies the need to ensure appropriate selection of participants. The selection of patient attributes noted in inclusion/exclusion criteria, which aims at minimizing comorbid conditions that may confound treatment signals, can also be detrimental to the external validity of the study. By quantifying and qualifying thematic clinical trial design trends over time, we propose a framework for providing proactive guidance in designing more efficient clinical trials. For example, there is an interesting opportunity in Alzheimer's Disease research to visualize how our understanding of the amyloid cascade has evolved over time, and the many ways researchers have since proposed the best means to engage it.

Methodology

ClinicalTrials.gov remains the definitive source for information regarding the rationale and design of clinical trial around the world. Recently efforts have aimed to extend and augment its capabilities beyond plain-text and retrieval. For example, the Clinical Trial Transformation Initiative (CTTI) was established in 2007 to enable finer-grained database querying capabilities of ClinicalTrials.gov[1]. More recently it has been demonstrated that new insights can be gained by directly analyzing certain aspects of a protocol and imputing additional endpoints, such as similarity scores between trial entries[2]. Expressing similarity sets as a graph yields interesting applications. In 2009, researchers at Los Alamos Laboratory analyzed voluminous amounts of clickstream data across web-based scientific journal websites to construct a detailed "map of science"[3]. While only visual in nature, their work demonstrated the intuitive connectivity of scientific domains. It is proposed that by analyzing ClinicalTrials.gov data, a similar network can be constructed of the clinical research space. By semantically annotating clinical trial protocols with a modular set of publicly-available ontologies optimized for conceptual coverage, similarity scores between trial entries can be calculated. By expressing similarities as weighted time-dependent edges, a massive graph is constructed. Novel hierarchical clustering techniques are deployed at different granularities and summarized by their most distinct and descriptive ontological concepts. This establishes metadata outlining thematic levels of varying granularity. Numerical values/ranges are also extracted from the text and summarized at each level. As a result, we augment ClinicalTrials.gov allowing users to not only query attributes of clinical trials, but also their overarching structure, made accessible through Aurelius' Titan platform.

Conclusion

The initial development of a framework which allows researchers to directly interrogate a "map" of a clinical research space based on graph metrics has been demonstrated.

References

1. Tasneem, A., Aberle, L., Ananth, H., Chakraborty, S., Chiswell, K.., et al. (2012). The database for aggregate analysis of ClinicalTrials.gov (AACT) and subsequent regrouping by clinical specialty. PloS One, 7(3), e33677.
2. Hao, T., Rusanov, A., Boland, M. R., & Weng, C. (2014). Clustering clinical trials with similar eligibility criteria features. Journal of Biomedical Informatics, (In Press).
3. Bollen, J., Van de Sompel, H., Hagberg, A., Bettencourt, L., Chute, R., Rodriguez, M. A., et al. (2009). Clickstream data yields high-resolution maps of science. PloS One, 4(3), e4803.

A Review of Clinical Decision Support Products in Dentistry

Kelsey M. Schwei[1], MS, Neel A. Shimpi[1], BDS, MM, Barbara Bartkowiak[2], MLIS, Zhan Ye[1], PhD, Ingrid Glurich[1], PhD, Amit Acharya[1], BDS, MS, PhD
[1]Marshfield Clinic Research Foundation, Marshfield, WI; [2]Marshfield Clinic, Marshfield, WI

Abstract

The adoption rate of health information technology is slow in dentistry and the implementation of clinical decision support (CDS) products in dentistry is even slower. This study aimed to identify CDS products that are used in dentistry and categorized the products' functionalities based off a literature review.

Introduction

With the implementation of electronic health records, the number of CDS systems is on the rise. CDS systems can help a dental clinician with patient-centered decision making as well as improve overall quality of care. Given the increasing importance of CDS in dentistry, this study evaluated the literature reviewed in a prior study[1] to identify and categorize functionalities of CDS products used in dentistry.

Method

Two informaticians reviewed the full text articles based on abstracts that were appraised in a previous study[1] to include CDS product review and classification of product functionalities. The articles were classified as relevant or non-relevant to CDS products in dentistry. Relevancy was defined based off of existing inclusion criteria[1]. The CDS products found within the articles were categorized as 'implemented and prototypical products' or 'conceptual ideas'. 'Implemented and prototypical products' were reviewed further. Each product discussed in the articles was included into a matrix. The matrix depicted the CDS products discussed in the literature on the y-axis and the CDS product functionalities on the x-axis. Product descriptions in the literature were used to determine functionality categories. The CDS products were classified as being developed in an academic or industry set-up. We also created a year of publication plot and used Google Citation to evaluate the number of citations on each article.

Results

Forty-one of the 49 articles from previous work[1] were identified to have discussions relevant to CDS products in dentistry. Implemented/prototypical (currently used or created and ready to use CDS products in dentistry) CDS products were discussed 63 times. Duplicate products were removed. Fifty-six unique CDS products were reviewed further for categorization into the functionality matrix. The frequencies of the functionalities of CDS products can be seen in Table 1. Based off of the frequencies of occurrence in the literature, the top CDS products were ORAD (frequency=3) and Logicon Caries Detector (frequency=3). Thirty-two CDS products had multiple functionalities. The top functionality was providing a diagnosis (frequency=40), followed by radiographic analysis (frequency=26). Ten CDS products were developed in academia, while 46 were developed in industry. Publication year for these 41 articles ranged from 1975-2012. The articles were cited 0-139 times in other publications.

Table 1: Frequencies of product functionalities from the articles. Products may have multiple functionalities.

Radiographic Analysis	Diagnosis	Practice Management	Prognosis	Treatment	Risk Assessment	Pathology	Database Management	Etiology	Total Products
26	40	1	4	20	1	1	1	1	56

Conclusion

This review has provided a functionality analysis of CDS products that are currently available in dentistry. The CDS products available could be useful in dental practice; however, there is always room for advancement on these CDS products. Furthermore, this review identified opportunities for further CDS development in other dental arenas, such as endodontics or oral cancer. As a next step, product reviews of the top identified CDS products will be conducted.

References

1. Schwei K, Shimpi N, Bartkowiak B, Ye H, Glurich I, Acharya A, 'Clinical decision support use in dentistry: feasibility for a literature review, AMIA 2013 Annual Symposium Proceedings, Pg 1260;

Automated Clinical Status and Treatment Prognosis tracking in Cancer Patients

Luis Selva PhD[1], Frank Meng PhD[1,2], Craig Morioka[2,3], Saiju Pyarajan PhD[1,4]
[1]MAVERIC, VA Boston Healthcare System, Boston, MA;; [2]UCLA Medical Imaging Informatics Group, Los Angeles, CA; [3]Greater Los Angeles VA Healthcare System, Los Angeles, CA, [4]Brigham and Women's Hospital, Harvard Medical School, Boston, MA.

Abstract
It has been estimated that in 2013 alone there were 1,660,290 new cases of cancer diagnosed in the United States with half a million cancer related deaths in the United States [1]. With technology advancements in the biomedical field and as treatment becomes more personalized and tailored to individual patients, tools will be needed that assist clinicians to determine the most effective treatment regimens based on all available data contained in the electronic medical record. We present a framework for clinical reports and image-based phenotyping using digital documents and medical images in order to give clinicians a rich representation of a patient's clinical condition.

Problem Description
There has been much recent progress in discovering and developing targeted therapy for cancer based on the genetic makeup of patients [2]. The Veterans Affairs (VA) Healthcare System is investing resources for developing clinical informatics infrastructures that support discovery of biomarkers for targeted therapy. The VA's Million Veteran Program (MVP) has a goal of collecting and analyzing consented bio-samples from a million veterans to build a large-scale genetic database that can be linked to the VA's electronic medical record (EMR) for enabling new discoveries. Treating patients with tailored therapies based on genetic information will require large data sets to successfully transition from general population-level medical care to targeted individualized care based on each patient's particular clinical and genetic characteristics. Precise and pertinent information on complex patient cases (e.g., cancer) is not always readily available to care providers. Here, we are building a framework to extract detailed information from unstructured imaging data and integrating with structured and unstructured clinical reports and notes that will give the physician the ability to determine a more precise clinical status using relevant health information in order to assist in optimizing treatment plans for individual patients.

Purpose and Methodology
In order to utilize targeted drugs for the treatment of cancer, clinicians must coalesce information from a multitude of sources, i.e., images, reports (pathology, radiology, molecular labs etc…) to design an effective treatment plan. Given enough information, care providers can also obtain detailed comparisons with other similar patients using historical clinical data to provide more accurate treatment response predictions [3]. We present a framework for deep phenotyping from the VA's extensive and voluminous EMR using clinical documents and integrating with radiology images to give clinicians a rich representation of patient status. Such a framework must not only scale to accommodate the extraction of large numbers of different phenotypes, it must also handle diverse data content and formats from the various regions of the VA network. Thus, the framework will have the following characteristics: 1) methods for rapid, large-scale development of phenotype extraction algorithms such as free text annotators and image feature extractors; 2) open, pluggable software architectures that enable the various algorithms to integrate seamlessly; 3) databases and warehouses that harmonize disparate data types along with provenance information that enables real-time querying and analysis (e.g., to determine discrepancies between care providers); and 4) continuous and scalable validation tools and techniques for evaluating system performance over time as changes occur both in the data as well as the system itself. Once a high quality structured database of phenotypes has been collected and curated, it can be combined with treatment outcomes data to support real-time querying by clinicians to retrieve specific characteristics of the patient or to gather groups of similar patients.

References
1. Rebecca Siegel, Deepa Naishadham, and Ahmedin Jemal, Cancer Statistics, 2013. CA Cancer J Clin 2013; 63:11-30.
2. Johnson BE, Kris MG, Berry LD, Kwiatkowski DJ, Iafrate AJ, Varella-Garcia, M., et. al. A multicenter effort to identify driver mutations and employ targeted therapy in patients with lunch adenocarcinomas: Lung Cancer Mutation Consortium (LCMC). In J Clin Oncol, Vol. 31, No. 15, 2013.
3. Sun J, Wang F, Hu J, and Edabollahi S. 2012. Supervised patient similarity measure of heterogeneous patient records. *SIGKDD Explor. Newsl.* 14, 1 (December 2012), 16-24. DOI=10.1145/2408736.2408740 http://doi.acm.org/10.1145/2408736.2408740.

Scaling Information Technology in U.S. Health Care System Reform: An Interdisciplinary Perspective

Ann Séror, PhD, MBA
eResearch Collaboratory, Quebec City, QC, Canada

Introduction

Health care system reform efforts in the U.S. have long called for large scale development of interoperable EHRs and capabilities for health information exchange across geographical locations and health care stakeholders - patients, service providers and payers. On July 21, 2004, the U.S. Secretary of Health and Human Services issued a news release calling for a ten-year program to develop and implement a national health information infrastructure, the "Decade of Health Information Technology". This infrastructure was to include EHRs for all patients as well as a national network for health information exchange. In 2009 the Health Information Technology for Economic and Clinical Health Act (HITECH), part of the American Recovery and Reinvestment Act (ARRA), further mandated the Office of the National Coordinator for HIT (ONC) and pursuit of specific objectives including certification of EHR systems and Meaningful Use of HIT as formulated in the Federal Health IT Strategic Plan (2011-2015). While significant progress has been made in adoption of EHRs by health care providers across the U.S., policies to promote health information exchange have generally failed, both in developing standards for interoperability among EHR systems as well as creation of a nation-wide health information network.

Research Problem

At the time of launch of the "Decade of Health Information Technology" research on implementation of HIT projected substantial returns on prospective investments – 81 billion dollars annually saved by improved health care efficiency and safety.[1,2] Such returns have proven elusive, in part because the multi-payer market driven foundation of the system has remained in place. This review considers the literature on the promise of U.S. health care system-level savings resulting from implementation of health information technology in general (HIT) and electronic health records (EHR) in particular. Early evaluations of HIT implementation failed to produce accurate projections of financial return on HIT investments. While some interdisciplinary studies have suggested the relevance of socio-technical models and complexity theory, no research has focused on a holistic view of the system including business model design and payers as critical stakeholders. The U.S. application of HIT continues to focus at the individual transaction level, where clinical information is consistently tied to assessment of financial transactions to be intermediated by multiple payers. The administrative costs associated with information systems to support this multi-payer business model design are excluded from assessment of return on HIT investments. Recent studies have shown that HIT investments have contributed to growth in health care costs – both at the individual provider and at the system levels of analysis.[3,4] In the U.S. health care system, HIT and EHR systems support clinical decision making, but they are designed at the same time for complex claims and billing processes – as physicians transact on average with more than 20 payers according to the American Medical Association. This administrative burden increases costs associated with development of information systems focused on increasingly granular transactions, as in the transition from ICD 9 to ICD 10, a more than five-fold increase in the number of codes for accurate billing.

Recommendations: Future Research and Policy

Conclusions of this review suggest that HIT investments will yield positive returns only with the implementation of a single payer system in the U.S.. Future research and policy should focus on redesign of the health care system business model. Such reform would contribute to improve ranking of the U.S. health care system among western nations. [5]

References

(Available on request.)

ProcessAWARE: Patient Outcomes and Resource Utilization Changes following Implementing an Electronic Rounding Checklist in the Intensive Care Unit

Ronaldo A. Sevilla Berrios MD, Sumanjit Kaur MD, Aysen Erdogan MD, Lisbeth Y. Garcia Arguello MD, John C. O'Horo MD, Adil Ahmed MBBS, Vitaly Herasevich MD, PhD, Brian Pickering MBBCH, Ognjen Gajic MD
Multidisciplinary Epidemiology and Translational Research in Intensive Care (METRIC), Mayo Clinic, Rochester, MN

INTRODUCTION: Checklists have been proven to be an invaluable tool for standardizing processes in various industries including healthcare, and is now gaining evidence in the intensive care unit (ICU) environment. Checklists can improve adherence to best practice. This should translate to more rational resource utilization and improved patient outcomes. We have developed an electronic, context-sensitive rounding tool that algorithmically selects which relevant best practice goals to be addressed for each patient on a daily basis. This was integrated into a novel electronic platform, the Ambient Warning and Response Evaluation (AWARE) system. Checklist in AWARE was deployed in April 2013 with a prescribed rounding choreography termed ProcessAWARE. We sought to test the impact of this intervention on resource utilization in the ICU.

METHODS: All patients admitted to a medical ICU undergoing ProcessAWARE implementation in April 2013 were compared to admissions in the same unit in April 2011 (this period of time was chosen to minimize seasonal variation and as the last year where no components of ProcessAWARE were available). We investigated trends in length of stay, ventilator utilization, mortality, central line-days, urinary catheter-days and antibiotic days. Each was adjusted for APACHE-III score to correct for individual patient severity of illness.

RESULTS: At total of 183 patients were evaluated in April 2011 and 199 in April 2013. There was no difference on age (mean 62 vs 61, p=0.75) or gender (52% vs 58% male, p=0.26) when compare the 2011 vs 2013 groups, however the APACHE III score at 1 hour were 42 vs 55 p <0.01 respectively. Regarding to resource utilization, adjusted central line days were significantly reduced post-AWARE implementation (-0.87 days, 95% CI -0.84--0.91, p=0.04), and antibiotic use days trended towards a decrease as well (-0.87 days, 95% CI -0.01--0.88, p=0.06). Otherwise, resource utilization was similar in the two study periods. ICU length of stay was significantly reduced after AWARE implementation (-0.92 days, 95% CI -1.62-0.23, p<0.01), as was hospital length of stay (-2.53 days, 95% CI -4.19--0.87, p<0.01). ICU mortality was also reduced (Odds Ratio [OR] 0.32, 95% CI 0.11-0.85, p<0.01), without significant change in hospital mortality (OR 0.54, 95% CI 0.25-1.13).

CONCLUSIONS: Although these results are preliminaries, the initial implementation of ProcessAWARE was associated with decreased resource utilization in the medical ICU. These findings are limited by the retrospective nature of the study and further evaluations are required.

REFERENCES:
1. Pickering BW, Gajic O, Ahmed A, Herasevich V. Novel representation of clinical information in the ICU - developing and testing user interfaces which reduce information overload. Applied Clinical Informatics. 2010; 1:116-31. DOI:10.4338/ACI-2009-12-CR-0027.
2. Ahmed A, Chandra S, Herasevich V, Gajic O, Pickering BW. The effect of two different electronic health record user interfaces on intensive care provider task load, errors of cognition, and performance. Crit Care Med. 2011 Jul;39(7):1626-34. PubMed PMID: 21478739.

An Exercise Mapping MedDRA to ICD-10-CM

Amy Sheide, RN, BSN, MPH; 3M Health Information Systems Inc., Salt Lake City, UT

Introduction

To maintain compliance with Medicare Secondary Payer Mandatory Reporting Provisions in Section 111 of the Medicare, Medicaid, and SCHIP Extension Act (MMSEA) of 2007, pharmaceutical companies currently report the diagnoses of patients participating in their clinical trials using the International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM), to be replaced by ICD-10-CM starting no later than April 1, 2015[1]. Since the pharmaceutical companies collect their clinical trial data using the Medical Dictionary for Regulatory Activities (MedDRA), they are currently meeting the reporting requirement using a MedDRA to ICD-9-CM map. The shift from ICD-9-CM to ICD-10-CM requires a new mapping from MedDRA to ICD-10-CM. Differences in the specificity and structure of the code systems presents challenges in the mapping between them.

Methodology

The MedDRA PTs from version 16.1 (released September 2013) and the 2014 release of ICD-10-CM were evaluated in the mapping project. Each MedDRA PT is associated with one or more high level MedDRA grouping concepts called a System Organ Class (SOC). The file reviewed contained 28,424 MedDRA PTs and their associated System Organ Class (SOC). The file did not contain any rows from the Surgical and Medical Procedures SOC. SOC was used to provide context to the MedDRA PT and was not mapped at this time. A total of 18,361 distinct preferred terms were eligible for mapping.

The 3M Healthcare Data Dictionary (HDD) is a terminology server that was leveraged for this mapping project. It contains multiple standard and local terminologies, including ICD-10-CM and the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT®). It also contains mappings between terminologies, such as the SNOMED CT to ICD-10-CM map set provided by the US National Library of Medicine (NLM), which was used for this project, as a previous study has shown that 9,484 MedDRA PTs have a direct mapping to SNOMED CT[2].

Results

Of the 18,361 MedDRA PTs 1,032 distinct PTs resulted in an exact text string match for an ICD-10-CM code description (5.6%). The remaining unmatched terms (17,329) generated 1,717 exact text string matches to SNOMED CT concepts that exist as 1:1 mapping to ICD-10-CM in the NLM map set, providing an additional 9.3% of automated matching to ICD-10-CM codes. The remaining 15,612 PTs required tool-assisted "manual" mapping, as the goal is to find a single "best match" to the most specific ICD-10-CM code. Some of the considerations were:

- Level of specificity – the best match for a PT could be a higher level, less specific "parent" code in ICD-10-CM which is not a "billable" code. For example, 'Tibia Fracture' would map to S82.209 'Unspecified fracture of shaft of unspecified tibia' rather than one of the 16 "billable" codes under it, since MedDRA does not capture laterality, episode of care, healing status or classification for fractures.
- Laterality – MedDRA PTs do not capture laterality except in specific disease instances. In general, the only time the terms 'right' or 'left' appear within MedDRA PTs is in the instance of ventricular cardiac disease. One of the key features of ICD-10-CM is the presence of laterality. The differences between the code systems decrease the chances of an exact equivalent match between MedDRA PT and an ICD-10-CM code.
- Gender – MedDRA may capture diagnosis both as a pre coordinated term 'breast cancer female' or 'breast cancer male' but also the genderless 'breast cancer' which does not have an corresponding genderless match in ICD-10-CM. Therefore, for the PT 'breast cancer,' one must choose between C50-919 malignant neoplasm of unspecified site of unspecified female breast and C50.929 malignant neoplasm of unspecified site of unspecified male breast.

Conclusion

Both MedDRA and ICD-10-CM are needed within the pharmacology industry to meet different reporting requirements. The mapping between the two code systems is feasible but presents challenges around mapping multi-level of terms with varying levels of attributes. Mapping from MedDRA PTs to ICD-10-CM is valuable in this specific use case but may not present the same relevant benefits outside the scope.

References

1. http://www.cms.gov/Medicare/Coordination-of-Benefits-and-Recovery/Mandatory-Insurer-Reporting-For-Non-Group-Health-Plans/Downloads/New-Downloads/S111ICD10Alert.pdf
2. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2815504/

Electronic Detection of Inpatient Diagnostic Error: A Scoping Review of Available "Triggers"

Edna C. Shenvi, MD, Robert El-Kareh, MD, MS, MPH
University of California San Diego, La Jolla, CA

Abstract

"Trigger" tools are specific criteria used to identify clinical records most likely to contain errors. We performed a scoping review of the literature to determine what such criteria may be used for automated detection of inpatient diagnostic error. Little automation or validation has been done specifically for this category of healthcare error.

Introduction

Diagnostic errors are a significant patient safety issue, but research into these errors is limited by difficulty in reliably measuring their incidence. While chart review remains the gold standard for detection, specific "triggers" may be useful screening tools to identify charts that are most likely to contain errors. Several established sets of criteria are designed to detect general adverse events, including nosocomial infections, adverse drug events, and procedural complications, and are not specific to diagnostic error. We sought to compile such explicit screening criteria that may indicate diagnostic error in hospitalized adult patients and are measurable in electronic health records, therefore making them amenable to automated detection.

Methods

Using a mixture of MeSH terms and keywords, we searched PubMed, Web of Science, and CINAHL databases and retrieved 8861 citations. After initial screening of titles and abstracts, we reviewed the full text of 146 articles. Citation tracking identified an additional five references to ultimately lead to 33 included articles that either developed or validated such tools. We extracted study setting, methods, and the criteria used, and we also examined for the presence of automation and if the study reported validation of criteria. We had developed a model of potential "signals" of outcomes of diagnostic error; some were specific to the patient status, such as death, whereas others were categorized as clinician assessment or management, like a care transfer. We determined which of the elements of our model were represented in actual studies.

Results

Several studies employed or validated variations of screening criteria developed in early studies on healthcare errors. Four of these criteria are associated with same-admission inpatient diagnostic error: transfer to a higher level of care, death, cardiac or respiratory arrest, and long length of stay. Subsequent trigger tools added four other clinical management indicators: calling multiple consults, change of physician in charge, change in procedure, and abrupt medication stop. Multiple studies used the trigger of calling a code or emergency response team, which often overlaps with cardiac or respiratory arrest.

Some studies reported validation metrics for their criteria such as positive predictive values or kappa coefficients for manual reviewers. However, only one study validated a specific criterion's performance in detecting diagnostic error. Only two studies used automated methods: comparing automated with traditional manual triggers and electronic screening of discharge summary text, although these did not report proportion of errors that were diagnostic in nature. Death, transfer to special or intensive care unit, and arrest or code team activation are the most widely used triggers to indicate error; however, none is specific for diagnostic error.

Discussion

Inpatient diagnostic errors may be more easily detected and studied using available electronic "triggers" to facilitate chart review. While this approach has been taken for adverse events in general, we did not find this approach applied specifically to diagnostic errors in the inpatient setting. Validation of these criteria is needed to identify those that are the most useful for these errors. Our preliminary model identified some additional criteria that may be useful for detecting diagnostic error, such as changes in primary diagnosis and multiple diagnostic procedures, which have not been used in either manual or automated methods for detecting error and may warrant future evaluation.

An Approach to Self-management of Metabolic Syndrome in Japan Using an Internet-Based System

Akiko Shibuya, RN, PHN, PhD[1], Tsuneo Yamada, MSc[2], Yukihiro Maeda[1], Yoshimasa Umesato, PhD[1], Yoshiaki Kondo, MD, PhD[1]
[1]Nihon University School of Medicine, Tokyo, Japan; [2]MEDIS, Tokyo, Japan

Abstract
We developed a healthcare monitoring and management system for individuals with metabolic syndrome (MetS) using an Internet-based system. Users can record health information, including daily blood pressure, waist circumference, and weight, in the system via a portal server at Tono City, Japan, and can obtain individually tailored advice and necessary medical information from healthcare professionals via their mobile phones or desktop computers. This system may have a significant impact on self-management for individuals with MetS.

Introduction
Metabolic syndrome (MetS) is linked to an increased risk of developing cardiovascular disease (CVD), type 2 diabetes mellitus (T2DM), and hypertension. Despite the potential to prevent T2DM and reduce CVD risk through lifestyle changes, individuals with MetS often have difficulty losing substantial weight and reducing waist circumference (WC) through self-management. In a previous study, poor decision making and a lack of both knowledge and tools for lifestyle interventions were considered important barriers for individuals with MetS[1, 2]. To help address these issues, we developed a healthcare monitoring and management system that can be accessed through mobile phones or desktop computers.

Methods and results
Figure 1 shows the architecture of our web-based healthcare monitoring and management system, which operates via the Internet. Users can record health information, including blood pressure, body weight, WC, and exercise, in the portal server and then view the data in graphical form on their mobile phones or desktop computers (Figure 2). Furthermore, through email or mobile phones, users are able to obtain personal health check-up data, including HbA1c, total cholesterol, and body mass index, as well as individually tailored advice from healthcare professionals regarding lifestyle factors such as diet, physical exercise, and other important medical information that can help reduce MetS risk factors.

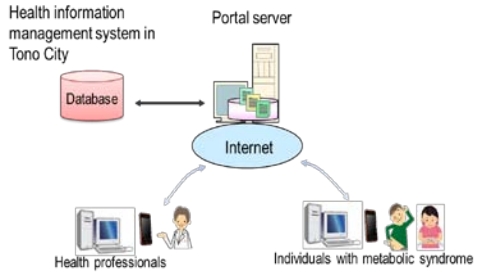


Figure 1. Overview of the system

Figure 2. The website as seen on a desktop computer monitor

To verify the efficacy of the system, 31 individuals (15 men, 16 women; mean [±SD] age, 56 ± 7.6 years) with MetS in Tono City, Japan, participated in this trial (response rate 81.6%). Participants first received lectures regarding the health site, after which they recorded their health status information and viewed the results in graphical form on their mobile phones or desktop computers. We then assessed their satisfaction and system usability using a questionnaire survey. Overall, 81% of the subjects rated the usefulness of the system as excellent or good. This indicates that the healthcare monitoring and management system for individuals with MetS was positively received. It is believed that this system may be useful for guiding behavior changes and lifestyle modifications for individuals with MetS.

Conclusion
The present study suggests that a healthcare monitoring and management system using mobile phones or desktop computers may help meet various support needs of individuals with MetS.

References

1. Goutham Rao, Lora EB, Bonnie JS, Linda JE, Melanie T, Alice H, et al. Circulation. 2011;124:1182-1203.
2. Dunkley AJ, Charles K, Gray LJ, et al. Diabetes, Obesity and Metabolism. 2012;14:616-625.

Developing an Electronic Health Record for Google Glass: Challenges and Use Cases

Karandeep Singh, MD[1,2]; Joseph V. Bonventre, MD, PhD[1,2]; Adam Wright, PhD[1,2]
[1]Brigham and Women's Hospital, Boston, MA; [2]Harvard Medical School, Boston, MA

Abstract

The mobile use of electronic health records (EHRs) has been primarily confined to smartphones and tablets, which require a user to hold the device. The Google Glass provides a unique platform by which a healthcare practitioner can access EHR data hands-free in real-time. This poster summarizes lessons learned during the development of a Google Glass-based EHR that interfaces with the Longitudinal Medical Record EHR at Brigham and Women's Hospital.

Introduction

There are several challenges that arise in designing an EHR for Glass, including but not limited to lack of a traditional input device, a fairly small display, a short battery life during activities requiring prolonged activation of the display or wireless connection, and absence of WPA2 enterprise wireless support. Though the Google Glass can be paired with a Bluetooth keyboard through the installation of the Android "Settings.apk" file using the Android Debug Bridge, the native input is confined to the touchpad, camera, accelerometer, and microphone. The touchpad is located on the right side of the Glass and can capture swipes and taps. Though this is effective for menu traversal and selection, it is quite slow for entering alphanumeric data. The camera is an effective way to capture secure information (such as a medical record number) using a QR code scanner, and it may also be an effective method to traverse menus through detection of hand gestures. The accelerometer is an effective way to tell to what extent the wearer is looking up or down and provides an effective way to scroll through free text data such as laboratories or clinical notes. Using speech recognition is an effective way to query the EHR but must be weighed against the noise pollution created by such voice commands. Another challenge to using the device in the exam room relates to the uncertain effect that the device may have on the patient-physician relationship: the display is not shared with the patient, the patient may feel their privacy violated due to the front-facing camera, and the novelty of the device itself may result in a feeling of unease. As such, initial use cases for interacting with the EHR must respect these social constraints.

Prototyped Use Cases

The following use cases have been prototyped in clinical scenarios using live patient data through the Glass EHR application. Additional proposed use cases include real-time critical results reporting, which could prove especially valuable in the emergency department, intensive care unit, and on inpatient hospitalist services.

1. Rapid Queries on Rounds: Using voice commands to ask for specific information (e.g. "what was the last creatinine?") on rounds may allow for faster information retrieval than traditional input-based methods. Secondly, for a supervising physician, having direct access to laboratory data on rounds while a trainee is presenting may allow for silent review of information that was not available prior to rounds.

2. A Second Screen for Clinic: While the use of a second screen has been associated with increased efficiency for many EHR-related tasks, using a Glass as a second screen has the added benefit of being always available, allowing a provider who is typing a patient note to have ready access to patient data for charting purposes without needing to occupy additional desk space through a second desktop monitor.

3. Glass as a Bluetooth Phone Accessory: The ability of Glass to make calls through a connected phone allow for voice commands and alerts to trigger the option to make a phone call (e.g. "call the primary care doctor").

Conclusion

Through the use of camera, speech, accelerometer, and touchpad-based inputs, traversing the EHR quickly and securely is technically feasible with Google Glass. Social barriers should be respected when designing a workflow in which Glass is used in clinical practice.

Using Data Mining as a Model for Behavior Change

Scott M. Sittig, MHI, RHIA[1], Amy Franklin, PhD[1]
[1]School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX

Abstract
In order to manage mounting health care costs for chronic conditions such as diabetes, innovative solutions for wide spread behavior change are needed. One challenge is tracking the success of broadly implemented interventions. Medical claims databases may provide a novel solution to measuring the success of such projects. Linking changing data from diagnoses, medication adherence, lab tests ordered, and visits to physicians may provide a means to track the success of behavior change programs.

Introduction
Medical claims databases contain raw aggregate data linked to diagnoses, medication usage, types of lab tests ordered, types of physician visits and related information. If used carefully, with full assessment for completeness and appropriate statistical transformation, these data can inform the function of the larger health care system[1]. In this poster, we extract components of medical claims database to determine their potential for the development and evaluation of a diabetes behavior change program. This effort is part of a larger project to show the use of claims database data to reflect individual behaviors such as self-efficacy.

Methods
Using a combination of data mining and evidence based rules to extract data from a hospital medical claims database (01/01/2013 – 12/31/2013), standard SQL and proprietary queries were used to access and claims related to a subgroup of diabetic patients. Building from a diagnosis code of 250.00 – 250.9, 120 adults were selected and their claims queried for medication adherence based on prescription filling patterns with an 80% or greater possession ratio, dates for HbA1c and LDL reported tests along with screenings for diabetic retinopathy and diabetic nephropathy.

Results
Adherence ratios for the subgroup diabetic patients were categorized into three classifications: diabetes care patterns, disease management, and medication adherence. Aggregate data from diabetes care patterns (e.g. frequency of physician visits and appropriate testing) and medication adherence (e.g. proportion of on time medication refills) were identified as potential indicators of behavior change. Findings from the historical record of 2013 indicate, that contrary to the recommendation of bi-annual HbA1c[3], only 66% of our sample had a HbA1c within this time frame. Less than 27% met the requirement for two HbA1c tests in the last 12 months. Medication adherence was also found to be sub-optimal in three-medication classifications that fell below the 80% threshold for adherence. These classifications included thiazolidedione-containing medication (75%), sulfonylurea (74%) and exenatide (29%). These data provide points of evaluation by which a successful targeted behavior change program may be evaluated.

Conclusion
Aggregation of health information into very large data sets and repositories offers extremely valuable opportunities and benefits[2]. A medical claims database contains valuable aggregate data on behavior and diagnosis specific adherence ratios, which provides a baseline for establishing a behavior change program.

References

1. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PRO, Bernstam EV, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care. 2013 Aug;51(8 Suppl 3):S30–37.
2. Bloomrosen M, Detmer D. Advancing the framework: use of health data--a report of a working conference of the American Medical Informatics Association. J Am Med Informatics Assoc Jamia. 2008 Dec;15(6):715–22.
3. Standards of Medical Care in Diabetes. Dia Care. 2005 Jan 1;28(suppl 1):s4–s36.

Clinical Decision Support Integrated in the Electronic Medical Record to Reduce Overuse of Blood Product Transfusion in the Cardiac Surgery ICU

Joseph H. Skalski, M.D., James R. Stubbs, M.D., Andrew A. Higgins, R.N., Lance A. Trewhella, R.N., Robert R. Bleimeyer, Kristy H. Doan, James M. Naessens, Sc.D., Pedro J. Caraballo, M.D., Mark H. Ereth, M.D.
Mayo Clinic, Rochester, MN

Abstract
Inappropriate transfusion of blood products has been associated with adverse clinical outcomes. Our institution implemented a comprehensive patient blood management program specifically for patients undergoing cardiac surgery. Clinical decision support rules mine patient data from the EMR including recent laboratory data and uses our transfusion algorithm to calculate recommended blood product needs individualized to the patient. We present preliminary data after a successful implementation.

Introduction
A conservative blood product transfusion strategy has been demonstrated to improve patient outcomes in multiple clinical settings. Cardiac surgery patients often have high blood transfusion requirements. A conservative transfusion strategy is not consistently followed by clinicians practicing in the cardiac surgery OR's and ICU resulting in overuse of blood products. Our institution has developed and implemented a comprehensive patient blood management program to guide transfusion for these patients by using a transfusion algorithm guided by coagulation tests. This program uses multiple interventions including education, CPOE-based clinical decision support (CDS) and reports. We present preliminary data showing provider compliance and initial impact of the CDS rules integrated in the medical record, specifically in the computerized physician order entry system (CPOE).

Methods
Our institution uses GE Centricity Enterprise as the main component of the electronic medical record (EMR) including the CPOE system. We used several CDS interventions to implement a previously published transfusion algorithm based on chest tube output and recent coagulation tests after cardiac surgery: 1) A new order set that simultaneously shows the orderable blood products (red blood cell, platelets, fresh frozen plasma, cryoprecipitate and granulocytes), and recent related lab orders and results (CBC with differential, prothrombin time, activated partial thromboplastin time, fibrinogen, and thromboelastogram maximum amplitude). 2) If a provider attempts to order a blood product for a cardiac surgery patient, but the patient lacks necessary recent test results, then the expert system displays an alert advising the order of clinically indicated tests before ordering the transfusion. 3) If a provider attempts to order a blood product and the required tests results are up-to-date, the expert system uses the transfusion algorithm on the EMR data to calculate recommended blood product needs individualized to the patient. It also displays a pop-up alert explaining the selection and advice to pursue based on clinical judgment. The provider has the option to change the preselected orders before finalizing. The system captures transactional data to support analytics.

Results
The CDS rules were implemented in 2011. The initial post-implementation acceptance rate of CDS rule recommendations was over 95%. Overall, blood product use for cardiac surgery has decreased approximately 40% since implementation without negative impact on clinical outcomes. However, the rule was implemented alongside multiple other interventions to encourage conservative blood product transfusion, so this reduction in blood product use cannot be attributed entirely to the CDS rules but the reduction is likely sustained over time.

Conclusion
Integrating CDS rules with individualized defaults into CPOE to tailor transfusion recommendations based on specific patient parameters is an effective intervention that is well accepted by providers. A comprehensive implementation program, including education and analytics, facilitates implementation of advanced CDS interventions. Additional evaluation of the long-term impact of these CDS interventions is underway.

The Financial Costs Associated with Implementing Electronic Health Records in U.K. Hospitals.

Sarah P. Slight MPharm, PhD, PGDip[1,2,3], Casey Quinn, PhD[4], Anthony J. Avery, MD, FRCGP[5], David W. Bates MD, MSc[1,6], Aziz Sheikh MD, MB, BS[1,3,6].

[1] The Center for Patient Safety Research and Practice, Division of General Internal Medicine, Brigham and Women's Hospital, Boston, MA, USA; [2] School of Medicine, Pharmacy and Health, The University of Durham, UK; [3] eHealth Research Group, Centre for Population Health Sciences, The University of Edinburgh, Teviot Place, Edinburgh, EH8 9AG, UK; [4] PRMA Consulting Ltd., Fleet, Hampshire, UK; [5] Division of Primary Care, University of Nottingham, UK; [6] Harvard Medical School, 250 Longwood Ave, Boston, MA, USA.

Abstract: *With increasing numbers of healthcare institutions considering implementation of electronic health record (EHR) systems, we sought to identify and categorize the costs associated with implementation and the factors that can influence these costs. We conducted 41 semi-structured interviews with members of implementation teams at 12 U.K. hospitals. The cost categories identified in this study can assist hospitals in the development of their business plans.*

Introduction: EHR systems hold the promise of improving the safety, quality and efficiency of health care.[1] U.K. hospitals have been slow to implement and adopt such systems, however, due in part to their costs and to uncertainties about whether they can achieve a return on investment. With more and more health care institutions considering implementation of EHR systems worldwide, this study aimed to categorize the costs associated with implementation and the factors that can influence these costs.

Methods: After obtaining the necessary ethical and institutional approvals, we conducted 41 semi-structured interviews between February 2009 and January 2011 with a diverse range of staff involved in the implementation of three different, centrally procured, EHRs at 12 U.K. hospitals. These hospitals were among the first to receive these systems as part of the National Programme for IT. Informed consent was obtained from all participants and interviews were audio-recorded with permission. We adopted an iterative approach to analysis, which enabled us to refine questions, investigate specific cost categories in greater depth, and pursue emerging themes and concepts during subsequent data gathering. A workable list of main- and sub-themes was developed inductively and applied systematically to these data with the aid of the computerized qualitative data analysis software QSR N-Vivo.[2]

Results: We identified four overarching cost categories associated with implementing EHR systems, namely: infrastructure (e.g., hardware and software), personnel (e.g., project management and training teams), estates or facilities (e.g., furniture and fittings), and other miscellaneous costs (e.g., consumables and training materials). A number of factors were felt to impact on these costs. The amount and type of infrastructure chosen by different hospitals appeared to be dependent on the stage of hardware maturity within the hospital; the requirements of the software application being implemented; the products currently available on the market; the budget (if predetermined); and the physical requirements of the wards or office rooms. The amount of resources spent on training clinicians and administrative staff to use the new EHR system depended on the number of users at each site; training methods employed; decision to backfill staff; and level of support provided to clinical users.

Conclusions: Healthcare organizations faced hard compromises relating to cost. For example, the infrastructure implemented often did not satisfy the demands of ward staff at peak times, but having more infrastructure was perceived as too expensive. With cost being one of the most significant barriers to EHR adoption worldwide, it is important for hospitals and governments to be clear from the outset as to the categories of costs involved and the factors that may impact on these costs, as well as the potential benefits which may result from the investments. The cost categories identified in this study can assist hospitals in the development of their business plans.

References:

1. U.K. Clinical Research Collaboration Select Committee on Health. 2007.
2. QSR International. 2011. http://www.qsrinternational.com

A Comprehensive Simulation Modeling Methodology to Reduce Health Care Process Redesign Risk

Rita Snyder, PhD, RN[1], Kevin Bennett, PhD[2], Bo Cai, PhD[3], Nathan Huynh, PhD[4], José Vidal, PhD[4], Bridgette Parsons, PhD(c)[4], Kolby Redd, MHA[1]
[1]College of Nursing; [2]School of Medicine; [3]Arnold School of Public Health; [4]College of Engineering & Computing; [1-5]University of South Carolina, Columbia, SC, USA

Problem. Redesign of high-risk health care processes can lead to unanticipated consequences. Ash, et al.[1-2] highlighted the significance of understanding clinical practice processes prior to implementation of CPOE systems. Others[3-4] have emphasized the need for better methods to assess EMR impact on clinician teamwork, and clinical communication processes. The persistent implementation failure of health information technology (HIT) innovations has escalated the need for better methods to assess pre-implementation workflow, and ongoing HIT implementation impact on workflow processes.[3-4] A comprehensive process redesign methodology is needed to examine the potential impact of process redesign interventions in advance of implementation.[4]

Methodology Description. A comprehensive three-phased health care process redesign methodology that integrates statistical and computer simulation modeling approaches is described (Figure 1). The goal of the methodology is to reduce process redesign risk in advance of actual intervention implementation. ***Phase 1*** supports multi-source contextual data collection. ***Phase 2*** uses Phase 1 data to iteratively model baseline workflow processes that are validated by subject matter experts until the most representative process model is identified. In ***Phase 3***, stakeholders examine the impact of simulated process redesign interventions on health care outcomes to determine the most feasible redesign intervention.

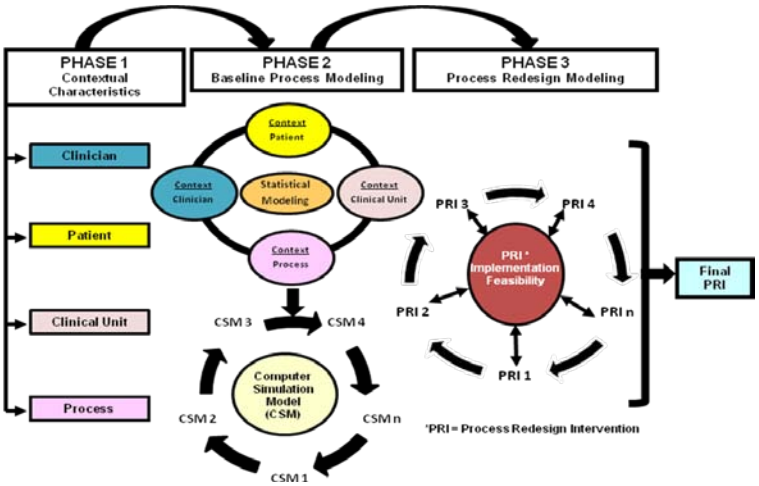


Figure 1. Comprehensive health care process redesign methodology.

Methodology Pilot Field Testing. Phases 1 and 2 were pilot tested on 2 comparable medical-surgical units using the medication administration process (MAP). Contextual data, consisting of RN characteristics, 305 RN MAP observations, and medication administration errors warnings corresponding to MAP observations, were examined using statistical and computer simulation modeling approaches. No statistically significant contextual differences were found between the 2 study units. Computer simulation modeling of MAP observations identified 2 types of RN MAP workflow patterns that corresponded to RN experience level, i.e., bundled=more experienced; unbundled=less experienced.

Conclusion. Field test findings supported Phase 1 and 2 data collection procedures, metrics, and statistical and computer simulation modeling approaches. Additional work is underway to field test expanded subject matter process model validation, and Phase 3 stakeholder redesign intervention feasibility assessments.

References

1. Ash, J. (1997). Organizational factors that influence information technology diffusion in academic health sciences centers. *Journal of the American Medical Informatics Association, 4*(2), 102-111.
2. Ash, J., Sittig, D., Poon, E., et al. (2007). The extent and importance of unintended consequences related to computerized provider order entry. *Journal of the American Medical Informatics Association, 14*(4), 415-423.
3. Gabriel, P. (2010). "Meaningful Use" means process redesign. *CHEST, 138*(3), 472-474.
4. Bowens, F., Frye, P., Jones, W. (2010). Health information technology: Integration of clinical workflow into meaningful use of electronic health records. *Online Research Journal: Perspectives in Health Information Management, 7*(Fall), Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2966355/

Exploration of Potential Drug Off-label Uses in Clinical Practice

Sunghwan Sohn, PhD[1], Hongfang Liu, PhD[1]
[1]Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN

Abstract

Off-label drug use is common in clinical practice but most of these off-label uses lack rigorous scientific evidences to support them. However, off-label drug use is legal and often useful. Analyzing actual clinical practice of off-label uses will be beneficial for patient treatments, patient safety, and quality improvements. In this study, we explored actual clinician-asserted drug uses from a large patient cohort in Mayo Clinic and compared them with the public drug usage database in order to investigate potential drug off-label uses.

Introduction

Off-label drug use is an intended drug use that is not approved by the FDA. It is very common in clinical practice. One in five prescribed medications in the U.S. is off-label. Although there are some public drug resources, they contain primarily labeled drug uses. A study that analyzed actual off-label drug uses in real clinical practice was limited. In this study, we compiled clinician-asserted drugs and their intended uses from clinical notes of 140K patients whose primary care is at Mayo Clinic and explored potential off-label uses.

Materials and Methods

We examined the Current Medication sections in clinical notes where most drugs reside and some of these drugs occur with clinician-asserted indications (i.e., drug uses) in a semi-structured way. We extracted the drugs using an open-source medication extraction and normalization system, MedXN (published in sourceforge) and indications using regular expressions. We then normalized the drug-indication pairs to standard terminologies—i.e., drugs to RxNorm ingredients and indications to SNOMED-CT concepts using RxNorm relation and MedTagger (published in sourceforge), respectively. Finally, we aggregated all indications along with their frequency for a given drug. These drug-indication pairs were compared with the public drug-indication resource, MEDI that is compiled from RxNorm, SIDER2, MedlinePlus, and Wikipedia. MEDI also contains a marker, 'possible label use,' which denotes whether a given indication is highly likely on-label or not.

Results

We investigated statistics at a different level based on the volume of patients who have a given drug indication. Mayo's drug-indication pairs match MEDI approximately 58% to 68% of the time, depending on patient volume. Figure 1 shows the portions of potential off-label candidates when $\log_2$(#patients) is >=5. Table 1 contains potential off-label uses.

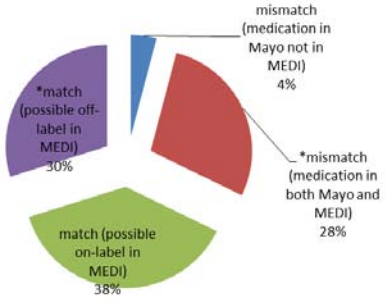


Figure 1. Off-label candidates (starts with *)

Table 1. Examples of potential off-label indications

Medication	Potential off-label
Brimonidine	photophobia
Bupropion	smoking cessation assistance
Albuterol	cough
Doxepin	allergies, hives
Azithromycin	traveler's diarrhea
Gabapentin	hot flashes
Trazodone	difficulty sleeping
Methylphenidate	depression

Discussion

Mayo's drug-indication pairs demonstrate that they have both off- and on-label medication uses. The ratio of potential off-label candidates, when restricted to candidate pairs that occurred in >=32 patients, is approximately 58%. However, it should be noted that this ratio in our results does not represent the actual off-label uses because we only investigated included indications. A well-managed database of actual off-label uses would likely prove beneficial to clinical practice. Our study may serve as a foundation for further investigation and the eventual development of such a database.

Google Glass for clinical procedures reference. Perception of optimal UI (user interface) and functionalities.

Yauheni Solad[1], MD; Nitu Kashyap[2], MD; Allen Hsiao[2], MD
[1]Yale University, Yale Center for Medical Informatics (300 George Street, Suite 501, New Haven CT); [2]Yale New-Haven Hospital (20 York Street, New Haven, CT 06510)

Background

Clinical procedures learning is one of the crucial steps in medical education. During residency, especially in small community hospitals, lack of required supervision can potentially limit the trainee's procedural exposure. Despite the abundance of reference material, not all references provide a quick procedure review, especially at the point of care. Wearable devices provide the opportunity to address this issue by providing crucial information at the point of care without clinical workflow interruption. Google Glass is an Android-powered wearable device with a consumer grade HUD (head-up display).

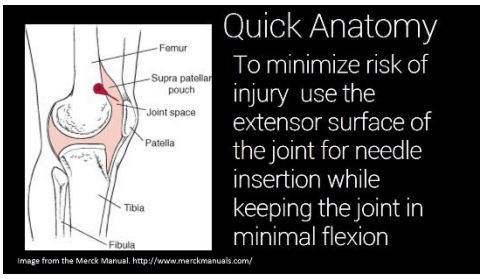
Goal

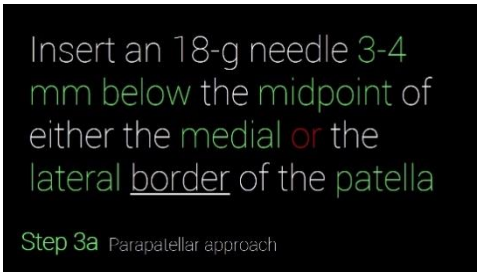
To evaluate perceptions of the different types of Google Glass user interfaces for clinical reference systems.

Methods

For demonstration purposes, we developed several versions of user interfaces (UI) for the Google Glass procedures reference system. Knee arthrocentesis was selected as a sample procedure and all versions utilized similar step-by-step procedural instructions. Demonstrated UI included one of the following: textual information, textual information and audio ("read aloud" function), graphic information, a combination of textual and graphic information, or video fragments.

To assess the UI perception, we surveyed medical students and residents at Yale New-Haven Hospital. After the demonstration, respondents answered a series of questions about the UI usability and potential applicability in clinical care. Examples of the Google Glass views:





Results

Preliminary data showed high levels of acceptance of the technology among medical students and residents. The combination of visual and auditory data in UI was found to be the most useful. Given the relatively small size and low-contrast level (especially in bright environments) of Google Glass display, textual information was found useful only in conjunction with audio ("read aloud"), especially if the view included significant amounts of text. Smaller video fragments, focused on every procedural step, got higher acceptance levels versus full procedural video. The majority of respondents supported the idea of potential live video procedure streaming to a supervisor. Many respondents liked the overall concept, but a significant number of questions were raised concerning patient privacy and degree of device acceptance in the patient population. Also technical limitations like: device security, heating problems and low battery time were found to be potential barriers to widespread device implementation.

Modeling Propensity for Readmissions with Claims Data

Gerardo Soto-Campos, PhD, MS,[1] Elisabeth L. Scheufele, MD, MS,[1,2] Aditya Sane, MS,[3] Santosh Narayanan,[1] Matvey B. Palchuk, MD, MS[1,2]
[1]ConvergeHealth by Deloitte, Deloitte Consulting LLP, Newton, MA; [2]Harvard Medical School, Boston, MA;
[3]Deloitte Advanced Analytics, San Jose, CA.

Abstract

We built a predictive model for hospital readmissions using claims data. The probability of hospital readmission was computed as a logit model. The coefficients for the logit model were trained with a Lasso regression on a set of 237,129 patients and tested on 70,465. The model's discriminative power evaluated via ROC was 0.813 and its calibration had p <0.05 assessed using the Hosmer-Lemeshow test.

Introduction

Health care organizations have been working to lower hospital readmission rates. Existing patient care models have been shown to improve readmission outcomes,[1] but solutions must be developed to help focus efforts on patients most amenable to readmission preventive management. Well validated predictive models offer a great alternative towards this solution. However, currently published readmission models report modest discriminative power (c-statistics varying between 0.55 and 0.75).[2] Hence we built a patient level predictive model for hospital readmissions with claims data using Lasso regression[3] which allows the inclusion of hundreds of covariates in a single model, thus taking advantage of the multitude of attributes available in patient data.

Methods

The probability for readmissions was obtained via a logit model with specific coefficients provided by the Lasso regression. MarketScan[4] data containing information on demographic, administrative, and some clinical data was used to build and validate the model. The model's discriminative power was estimated via a receiver operating characteristic (ROC) curve[3] and its calibration was evaluated with the Hosmer and Lemeshow's approach.[5]

Results

The original Lasso regression used 403 covariates. The model was trained with a random sample of 237,129 patients. The coefficients describing the probability of readmission in this model were applied to each patient of an independent testing set of 70,465 random cases. The ROC curves, with c-statistic and 95% CI and the Hosmer-Lemeshow p-values were computed (see Table 1).

Set	N	c-stat (95% CI)	Hosmer-Lemeshow p
Testing	70,465	0.813 [0.809 , 0.817]	<0.05
Training	237,129	0.820 [0.818 , 0.822]	<0.05

Table 1: Lasso Model of Readmission Propensity

Discussion

Our model has better c-statistic than most of the currently published readmission models. It separates readmission cases from non-readmission cases more than eighty percent of the time (c-statistic = 0.813). The low p-values found in the recalibration (Hosmer-Lemeshow, HL) may suggest that the predicted probabilities of readmission overestimate the actual observed readmission risks in the data. However, there is evidence[5] that for N larger than 50,000 a Hosmer-Lemeshow test with p-value smaller than 0.05 does not necessarily imply that the model's predictive capabilities are suspicious or that the goodness of fit is incorrect. Next steps include using different non-linear approaches taking into account variations in readmission risk as a result of geographic location and incorporating data from the clinical environment (e.g., labs, radiology results, etc.) into the model.

References

1. Coleman, E. et al, The Care Transition Intervention. *Arch Intern Med*, 2006, 1822-1828.
2. Kansagara, D. et al Risk Prediction Models for Hospital Readmission. A Systematic Review. *JAMA*, 2011, 1688-1698.
3. Hastie, T., et al. *The Elements of Statistical Learning 2nd Edition.* Stanford CA: Springer Verlag. 2008
4. Truven Health MarketScan® Commercial and Medicare Databases.
5. Kramer, A.A., and Zimmerman J.E., Assesing the calibration of mortality benchmark in critical care: The Hosmer-Lemeshow test revisited, *Crit Care Med*, 2007, 2052-2056.

Specifying Initial Requirements and Architecture for the CUPID System

Karen Sousa, PhD[1], Blaine Reeder, PhD[1], Mustafa Ozkaynak, PhD[1], John Welton, PhD[1]
[1]University of Colorado College of Nursing, Aurora, CO

Abstract

There is a need for distributed data sharing networks that enable analysis of patient data aggregated from multiple hospitals to answer operational research questions, especially those related to nursing work and patient outcomes. The University of Colorado (CU) College of Nursing (CON) is leading the effort to develop such a system in collaboration with regional hospitals and other stakeholders. This poster describes our efforts to define requirements and architecture for the CUPID system.

Introduction

One of the primary goals of the Colorado Collaborative for Nursing Research (CCNR) at CU CON is to create a nursing data analysis center that enables near real-time nursing quality metrics from electronic health record data and delivers these metrics and trends back to hospitals for operational decision-making. This goal is being realized through the work of a CCNR technical team with efforts to design and develop the University of Colorado Patient Initiated Data (CUPID) system. The purpose of the CUPID system is to improve efficiency and effectiveness of quality patient care by changing the paradigm of how we use data to drive care processes.

Methods

The CCNR technical team has deep experience in nursing research, informatics and system design. Our current goal is to specify the requirements for CUPID and evaluate candidate systems that will meet these requirements. Through regular meetings we have specified data aggregation use cases that CUPID must support. Our design approach is informed by a reusable design philosophy[1] through which we seek to leverage existing systems and infrastructure to meet CUPID system goals. As part of this approach, we have initiated collaboration with CU Colorado informatics colleagues who recently completed the requirements and evaluation process for the Colorado Health Observations Regional Data Service (CHORDS)[2] project. Regular meetings of the CCNR team have resulted in a timeline for deliverables linked to requirements, a working list of use cases and candidate systems for evaluation.

Results

A partial list of use cases that CUPID must support includes: 30-day hospital readmissions; nursing work as it relates to patient outcomes; verification of patient symptom documentation by nurses; and workflow analysis through EHR documentation. In addition, we have created a general process flow for CUPID related to these use cases for data publishers and subscribers from participating organizations (Figure 1).While due diligence efforts to evaluate candidate systems must still be conducted, a comparison of CHORDS functional requirements with early CUPID requirements shows promise that CHORDS infrastructure aligns with system needs.

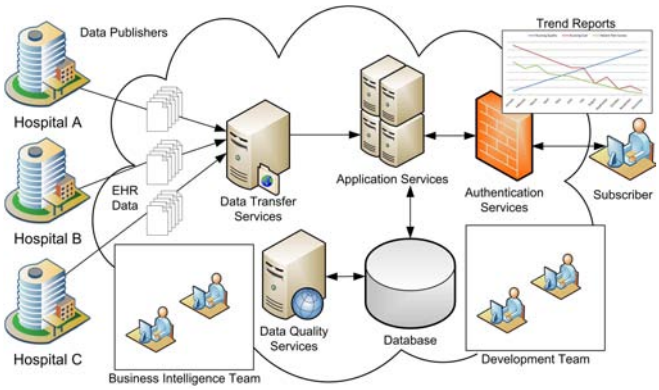


Figure 1. CUPID system process overview

Conclusion

Full CUPID requirements and results of evaluation efforts will be presented in addition to progress of the CUPID project at the AMIA Annual Symposium in November 2014.

References

1. Reeder B, Hills R, Demiris G, Revere D, Pina J. Reusable design: A proposed approach to Public Health Informatics system design. *BMC Public Health.* 2011;11(1):116.
2. Ames MJ, Bondy J, Johnson SC, Wade TD, Davidson A, Kahn M. Analysis of Federated Data Sharing Platforms for a Regional Data Sharing Network. AMIA 2013 Summit on Clinical Research Informatics; March 20, 2013 - March 22, 2013, 2013; San Francisco.

A System Usability Study Assessing a Machine-Assisted Interactive Interface to Support Annotation of Protected Health Information in Clinical Texts

Brett R. South, MS[1,3], Danielle L. Mowery, MS[1,4], Jianwei Leng, MS[1,2]
Stéphane M. Meystre, MD, PhD[1,3] , Wendy Chapman, PhD[1,3]
[1]Salt Lake City VA Health Care System, Departments of [2]Internal Medicine & [3]Biomedical Informatics, University of Utah, Salt Lake City, UT; [4]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA

Abstract: *Redaction of HIPAA-specified protected health information (PHI) from clinical texts remains an important research topic in the clinical NLP domain. Reducing the manual workload required to adequately de-identify clinical texts is one area that could benefit from machine-assisted annotation methods. Using a publicly available clinical corpus, we assessed usability of a machine-assisted approach that combined outputs from an existing de-identification system called BoB and outputs from an interactive interface called the "Oracle Module" from an annotation tool, extensible Human Oracle Suite of Tools (eHOST). From our annotation experiment, we report annotator perceptions from a self-reported questionnaire obtaining System Usability Score (SUS) ratings.*

Introduction: Generating a reference standard that identifies PHI types defined by HIPAA[1] is a challenging task. More efficient approaches to this task could be used to reduce the workload associated with manual review. We assess the usability of a machine-assisted approach that integrates pre-annotations from a de-identification system with outputs from an interactive annotation interface. The SUS developed by Brooke[2] is a simple, ten-item attitude Likert scale that provides a global view of subjective assessments of usability. System usability can be measured by taking into account the context of the system being used by those who use it for a specific task– in this case de-identification of clinical texts. Key dimensions of usability integrated with SUS include ***system effectiveness***, ***user satisfaction,*** and ***system efficiency*** which were assessed for this usability study.

Methods: We randomly sampled 1,535 unique texts from a sample of 2,330 clinical reports comprised of 40 subtypes from the publicly available MTSamples corpus (www.mtsamples.com). To generate pre-annotations, we applied a de-identification system called BoB - "Best of Breed" - originally developed in the Consortium for Healthcare Informatics Research (CHIR) De-identification project[3]. Seven reviewers independently annotated this subsample for all annotation types applying the CHIR annotation guidelines and schema. For the annotation experiment, we implemented a nonequivalent control group design with replication switching. Annotators alternated an experiment condition of reviewing document batches using the machine-assisted approach (BoB+eHOST Oracle) with a control condition of annotating raw documents (no BoB+eHOST Oracle).

Results: ***System effectiveness:*** Usability ratings obtained from SUS were slightly above average (69%). False positive annotations were introduced for <1% of eHOST Oracle annotations versus 67% introduced by BoB. ***User satisfaction***: Although annotators indicated they preferred using the eHOST Oracle module alone, eHOST Oracle only accounted for 640 (3.6%) out of 17,643 annotations generated by all 7 annotators. ***System efficiency:*** No statistically significant gains in efficiency were observed comparing the machine-assisted approach with the control.

Conclusion: Further research is required to assess under what conditions machine-assisted approaches result in above average usability ratings and efficiency gains.

Acknowledgements: This study was supported by the integrating Data for Analysis, Anonymization and Sharing (iDASH) NIH (U54 HL 1084) and ShARe project NIGMS (R01GM090187). Permission was obtained to use the MTSamples data set for our research from the UCSD IRB and the MTSamples transcription company.

References

1. Standards for privacy of individually identifiable health information: final rule. 67 Federal Register 53181 (2002) (codified at 45 CFR 160 and 164).
2. Brooke J. SUS: A "quick and dirty" usability scale. In: Jordan PW, Thomas B, Weerdmeester BA, McClelland (Eds.) Usability Evaluation in Industry pp. 189-194. Taylor & Francis, London UK (1996).
3. Ferrández O, South BR, Shen S, Friedlin FJ, Samore MH, Meystre SM. BoB, a best-of breed automated text de-identification system for VHA clinical documents. J Am Med Inform Assoc (2013) 20(1), 77–83.

A universal software-tool supporting proper continuity of care data handling

Basile Spyropoulos, PhD[1], Maria Botsivaly, Msc[1], Vasileios Pierros, Msc[2], Mamakou Vassiliki, MSc, MD[3], Aris Tzavaras, PhD[1]
[1]Technological Educational Institute (TEI) of Athens, Athens, Greece, [2]University of Athens, Athens, Greece, [3]Dromokaiteion Psychiatric Hospital, Athens, Greece

Abstract

It is the aim of this project to present the work progress of a multifunctional web-based framework, designed to support the continuity of health care information management, as the patient moves between different points of care. The system consists of a Patient Management module, a non-Hospital Scheduling and a non-Hospital Supervision module, serving proper continuity of care handling. It complies with E2369 (CCR)[1] and ISO 13606-1:2008[2] Standards and it combines symmetric and asymmetric cryptographic algorithms' data-protection.

Introduction

Healthcare provision will soon move towards outpatient, community and homecare schemata. This development is expected to provide for the upcoming demand for patient-centered care, while reducing at the same time the level of reimbursement for healthcare organizations and providers.

Methods and Resources

We have developed a complete platform, to ensure Continuity of Medical Care among Primary Health-care Agencies, Hospitals and Home Care, according to existing International and European regulations and standards, together with the emerging relevant Greek National ones. The platform is web-enabled and modular, consisting of three applications, the Patient Management, the Non Hospital Scheduling and the Non Hospital Supervision. The system utilises all the appropriate sets of clinical terminologies, coding systems and vocabularies[3]. It is also designed in such a way as to provide for both multilingual user interface and for the use of different terminology and coding systems as well, depending on the user's preferences. The system has some additional special features as it is the formation of CCRs according to different standards and the transmission of referral letters using a semantically annotated web service. The Non-Hospital Scheduling application offers a very flexible and easily adoptable module for the physician to schedule the homecare activities that are recommended for each particular patient, while the Non Hospital Supervision applications provides for both the surveillance and the monitoring of the homecare plan.

Results

The system in progress is currently being tested in our laboratory with anonymised data provided to us by the physician participating in our team. Both the services and the data bases are stored in our laboratory's server, which hosts a Windows 2008 Server System with IIS 6.0 (Internet Information Services). The various coding systems, clinical terminologies and vocabularies that are used in different countries have different annotations and hierarchical structures, making it difficult to create a unitary user interface, applicable to different countries. Semantic Web technology appears to be a feasible solution. Nevertheless, significant efforts still have to be made towards the realization of semantically reach distributed networks in healthcare domain. Such technologies require the development and adoption of well-structured and annotated vocabularies and the use of appropriately designed formal ontologies. The non-Hospital Supervision may be operated as a web-based and a stand-alone application, supporting a local database stored in home computer and allowing for the decision supporting services to be executed locally.

Conclusion

Regarding the system's architecture, we consider that the use of web-based techniques are appropriate for such applications, since web-based models allow for numerous distributed users; while reducing the maintenance task to a group of web servers at the same time.

References

1. ASTM www.astm.org: E2369-12, Standard Specification for Continuity of Care Record.
2. ISO 13606 – 1:2008. Health informatics – Electronic health record communication – Part 1: Reference model.
3. American academy of Allergy, Asthma and Immunology Practice Management Resource Guide, 2012 edition.

Developing and Testing a Web-based Interdisciplinary Patient-centered Plan of Care

Diana L. Stade[2], Kelly McNally[2], Anuj K. Dalal MD[2,3], Kumiko Ohashi RN, PhD[2,3], Sarah A. Collins RN, PhD[1,2,3], Conny Morrison[2], Jae-Ho Lee MD, PhD[1,2,3], Frank Chang MSE[3], Katherine Robbins RN[2], Anthony F. Massaro MD[2,3] David W. Bates MD, MSc[1,2,3], Patricia C. Dykes RN, PhD[2,3]
[1]Partners Healthcare Systems, Wellesley, MA; [2]Brigham and Women's Hospital, Boston, MA; [3]Harvard Medical School, Boston, MA

Abstract: *The goal of an interdisciplinary, patient-centered plan of care is to actively engage patients, nurses, and physicians in meaningful collaboration. Currently, patients and family caregivers are not formally involved in developing their plan of care while recovering from acute illness in the hospital. Diagnoses, problems, goals, and safety measures are documented in several places across disciplines. To meaningfully engage patients in developing their plan of care through the use of bedside technology, information documented by nurses, physicians, physician assistants and sub-specialists must be reconciled. In this study, we describe our methodology for developing a web-based interdisciplinary, patient-centered plan of care.*

Introduction: Patients want easy access to meaningful medical information while in the hospital and ways to effectively communicate with their providers to engage in recovery. We previously identified a core set of information needs to engage patients in their plan of care using web-enabled devices at the bedside.[1] Our first prototype offered provider identification, medications, test results, and dietary information. However, a method of providing information such as medical problems, goals, and safety precautions to the patient at the bedside is not available due to current methods of documentation. Furthermore, there is limited ability to capture the patient's goals and concerns and communicate them back to the care team. In this study, we develop and test web-based interdisciplinary plan of care tools for providers and configure them to interact with the patient portal at the bedside.

Methods: We conducted observations in medical intensive care and oncology units at Brigham and Women's Hospital to identify workflow processes and methods for documenting the plan of care by nurses and physicians. We reviewed 17 nursing plan of care worksheets where problems and goals were documented. Subsequently, we conducted qualitative focus groups with healthcare providers facilitated by rapid prototype development of 10 + versions in PowerPoint and InfoPath to simulate electronic versions of the web-forms. Usability testing was performed on provider and patient facing tools.[2] Bedside interviews were conducted to obtain feedback on the patient facing tool with accessible plan of care content.

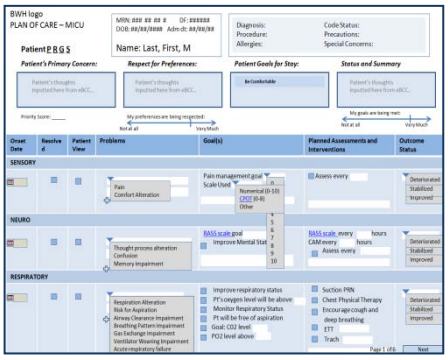


Figure 1- Prototype of a web-based interdisciplinary patient-centered plan of care. Patients' goals and preferences for care are displayed for clinician view and medical problems and goals are selected for patient view at bedside.

Results: Plan of care documentation methods and workflows such as morning rounds and family meetings were identified. Structured problems and goals were adopted from the Clinical Care Classification system and integrated into a web-based plan of care worksheet (Figure 1).[3] Tailored plan of care information was made available on the bedside devices with infobuttons linking to Medline Plus to provide patients self-managed education. A rounding checklist tool was developed to inform patients of safety precautions related to preventable harms and to provide clinicians with risk assessment data. The patient plan of care at the bedside was optimized for the iPad Air.

Conclusion: We developed innovative web-based tools that promise to facilitate development of an interdisciplinary, patient-centered plan of care. We plan to refine and implement these tools for patients, family caregivers, and providers to use in the medical intensive care and oncology units at Brigham and Women's Hospital.

Acknowledgements: The BWH PROSPECT study is funded by the Libretto Consortium supported by the Gordon and Betty Moore Foundation.

References

1. Dykes PC1, Carroll DL, Hurley AC, Benoit A, Chang F, Pozzar R, Caligtan CA. Building and testing a patient-centric electronic bedside communication center. J Gerontol Nurs. 2013 Jan;39(1):15-9.
2. U.S. Dept. of Health and Human Services. The research-based web design & usability guidelines. Washington: U.S. Government Printing Office, 2006. Available from: http://www.usability.gov/.
3. Saba VK. The clinical care classification (CCC) system, version 2.5 [Internet]. Pending Copyright 2012. Available from: http://www.sabacare.com/.

Facilitating Reconciliation of Inter-Annotator Disagreements

Johann Stan, PhD, Dina Demner-Fushman, MD, PhD, Kin Wah Fung, MD, MS, Olivier Bodenreider, MD, PhD,
Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, National Institutes of Health, DHHS, Bethesda, MD

Abstract

Development and evaluation of Natural Language Processing methods often requires text annotation. To gauge the difficulty of the task and increase the reliability and quality of annotations, researchers often recruit at least two annotators. The discrepancies in annotations by multiple annotators need to be identified and reconciled. We present a tool that identifies and helps reconciling and validating annotations in a widely used annotation tool Brat.

Introduction

In the process of annotating a corpus of drug package inserts for drug-drug interactions, we were faced with a problem of reconciling differences in fairly complex annotations of interactions between drugs, drug classes and substances. Our goal was to annotate interactions for training supervised machine learning (ML) algorithms and evaluating the results. Annotated corpora are most useful for training ML tools if they are consistent. To ensure consistency, two experts annotated the interactions and two senior annotators adjudicated the disagreements. To facilitate annotation, we used Brat[1] that is fairly convenient for annotation, but does not provide mechanisms for reconciliation of disagreements. Therefore we have developed functions that allow reconciling disagreements and ensure consistency of annotation across similar interactions mentioned multiple times in the drug package inserts.

Methods

The *Disagreement Reconciliation module* compares annotations by two annotators and uses line numbers in the files loaded to Brat for annotation to indicate the location of disagreements. The following details about the disagreements are printed to a text file: 1) absence of an annotation, for example, *"119 MISSING FROM FILE 1 Type: Drug Span: Amplodipine"* indicates that the first annotator did not highlight Amlodipine as drug in the 119th line of the package insert; 2) different labels assigned to the same span, for example, if one of the annotators labeled *alcohol* as substance, and the other one as drug. The *Sentence Validation module* identifies similar sentences and checks if they have been annotated consistently. Sometimes different package inserts for drugs in the same drug class or different sections of an insert contain almost identical sentences. These sentences can be used to quantify intra-annotator agreement over the entire annotation process, as well as consistency in reconciling disagreements. Sentence similarity was computed using an implementation of the *Jaccard* similarity measure (threshold 0.75). Annotated biomedical entities as well as those identified by Metamap were replaced with standard names, e.g. "drug" in order to capture similar syntactic constructions used for different representatives of drug classes with similar pharmacodynamic and pharmacokinetic properties.

Results

We tested the annotation reconciliation tools on 176 manually annotated package inserts (8081 biomedical entities and 4841 interactions). The tools identified 2584 discrepancies, of which 1200 were in entity annotation and 1384 in interaction annotations. 320 similar sentences were annotated inconsistently (e.g. different types attributed to same drugs or different interaction types).

Conclusion

The tools helped us improve annotation guidelines and detect and reconcile discrepancies. To the best of our knowledge, this is the first publicly available extension for assisting with discrepancies and assuring consistency of annotations using Brat. The toolkit can be downloaded from http://lhce-brat.nlm.nih.gov/disagreementAnalyzer.htm

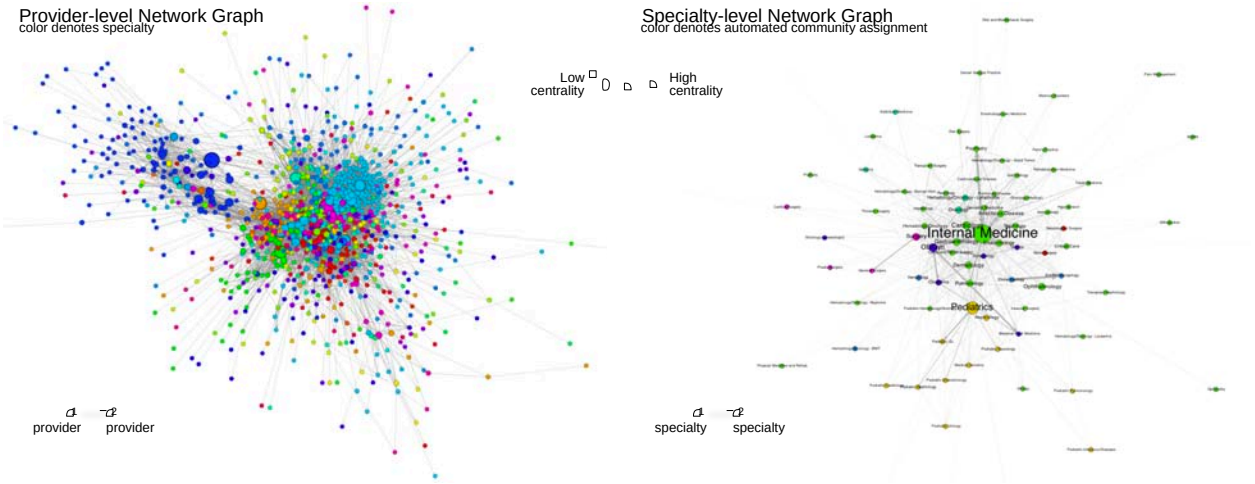
References

1. Stenetorp P, Topic G, Ohta T, Ananiadou S, Tsujii J. Brat: a Web-based Tool for NLP-Assisted Text Annotation, In Proceedings of the Demonstrations Session at EACL 2012, 2012.

Social Network Analysis of EHR-Based Provider Communication

David E. Stark, MD[1], Zachary M. Grinspan, MD, MS[1,2], Daniel M. Stein, MD, PhD[2]
[1]Division of Pediatric Neurology, Department of Pediatrics and [2]Department of Healthcare Policy and Research, Weill Cornell Medical College, New York, NY

Introduction: Chronic illness drives a substantial percentage of national health expenditures. Effective care coordination is essential to longitudinal management of this population, and has been a key motivator driving large investments in health information and communications technology. Despite the substantial cost and uptake of communication tools to facilitate team-based collaborative care, their usage patterns remain relatively understudied.
Methods: We conducted a retrospective cross-sectional analysis of EHR-based communication within the ambulatory care network of a large academic medical center. We obtained all patient-linked provider-to-provider messages generated over an 8-month interval (8/2012-4/2013) and employed social network analysis to construct network graphs at the level of individual providers and provider specialties. In addition to performing an exploratory analysis of electronic communication patterns, we sought to determine the extent to which measurable differences in communication patterns might exist among primary care practices. In particular, we contrasted a hospital-based practice largely serving a public insurance population with hospital-affiliated practices caring predominantly for privately insured patients. Recognizing that these practices differ in many respects, we expected to measure quantifiable differences in communication patterns among the practices. Specifically, we hypothesized that within their respective networks, PCPs for privately insured patients would exhibit greater relative centrality as compared with PCPs for publically insured patients, reflecting a greater tendency for care management within the former group to conform to a 'hub-and-spoke' model in which the PCP (as the hub) communicates directly with specialists.
Results: We collected and analyzed 40,753 messages between 1,060 providers (89% MDs) regarding 12,393 patients. Providers exchanged on average 2.6 messages/week (range 0.06 – 73.9), typically during business hours (weekdays, 8a-5p). As in other domains, social network analysis recapitulated formal organizational structures and unveiled other informal communication patterns. Automated community detection (Walktrap algorithm) revealed four major clusters consisting of medical, surgical, pediatric, and OB/Gyn subspecialties. Further inspection also revealed interdisciplinary subnetworks; for example strong ties between specialists within internal medicine, infectious disease, and psychiatry involved in the care of HIV patients. As hypothesized, quantifiable differences emerged between public- versus private-insurance practices. Messages regarding private-insurance patients were more likely to be exchanged between a PCP and a specialist (47% of private-insurance messages versus 27% of public-insurance messages) and less likely to be between PCPs (35% private vs. 50% public) or between specialists (18% private vs. 23% public), X^2 (2, N = 14,036) = 526.37, $p < 0.001$. Network analysis at the provider level demonstrated that PCPs in the private-insurance network had substantially greater relative betweenness centrality than PCPs in the public-insurance network (4.0 versus 2.1, Wilcoxon rank sum test, W = 7694, $p = 0.003$). At the specialty level, overall graph centralization was greater for private- versus public-insurance patients (0.71 versus 0.64, respectively).
Conclusion: Social network analysis permits quantitative measurement of EHR-based provider communication. Our findings are foundational for ongoing work developing novel methods to quantify care coordination. Such measurements will support (1) research to identify factors associated with care coordination and (2) initiatives to develop more advanced social computing tools within EHRs for facilitating coordinated care.



Collaborative mHealth Tools for Diabetes Management

Bryan Steitz, Erika Poole PhD, Madhu Reddy PhD
College of Information Sciences and Technology
The Pennsylvania State University, University Park, PA

Mobile health interventions are an important part of the self-management of diabetes. The use of mobile technologies to support collaboration among stakeholders is important for effective self-management. We surveyed eighteen healthcare professionals in order to better understand the collaboration and data needs for the successful implementation of mobile health tools for diabetes self-management.

Background

It is estimated that 382 million people worldwide are living with diabetes.[1] Mobile health (mHealth) technologies are at the forefront of diabetes self-management and have shown promising results in the management of diabetes-related complications. Patients must undergo a lifetime of treatment and self-management, much of which can be assisted by the use of mHealth tools in collaboration with their care providers. Current applications offer a variety of self-management tools such as educational materials, nutrition tracking, activity monitoring, and blood glucose management. However, the extent to which collaboration is supported in such applications has been relatively unexplored.

Methods

We surveyed a cross-section of participants (industry and academia) of the National Science Foundation's Center for Health Organization Transformation's (CHOT) fall meeting to identify potential opportunities and barriers for using collaborative mHealth diabetes management applications. The brief survey included six open-ended questions dealing with collaboration or the use of such applications for collaborative management with external stakeholders. Eighteen surveys were distributed and twelve (66.7%) were completed. Survey results were compiled into an Excel spreadsheet and analyzed for common themes.

Results and Discussion

The majority of respondents indicated that collaboration between the patient, family, and physicians is most important for diabetes self-management tools. For the applications themselves, patient compliance with self-management interventions and medications is among the most important for the surveyed healthcare professionals, followed by the BG levels and results of other tests. When designing diabetes mHealth applications, they should be simple to use so that older patients can operate them. The development of diabetes self-management applications should focus on the collaboration between stakeholders, especially when dealing with adolescents and elderly users.[2, 3] The increased use of diabetes mHealth tools with a collaborative focus will provide additional benefits to all users and will allow the physician to provide higher quality and more accurate care.

Conclusion

The impact of specific features for diabetes self-management mHealth applications is not easily determined, but our findings support the need for additional collaboration. Present applications offer a variety of services for self-management, but without additional collaboration features mHealth will not reach its maximum potential.

Acknowledgement

This research was supported by a grant from Penn State University's Center for Integrated Healthcare Delivery Systems (CIHDs).

References

1. *IDF Diabetes Atlas*. 6th ed. 2013: International Diabetes Federation.
2. Cafazzo, J.A., et al., *Design of an mHealth app for the self-management of adolescent type 1 diabetes: a pilot study*. J Med Internet Res, 2012. 14(3): p. e70.
3. Li, Y., et al., *A Design to Empower Patients in Long Term Wellbeing Monitoring and Chronic Disease Management in mHealth.* Stud Health Technol Inform, 2013.

An Integrative Framework for Drug Target Prediction and Repurposing

Jingchun Sun, Ph.D[1], Cui Tao, Ph.D[1], Kevin Zhu, BMath[1], W. Jim Zheng, Ph.D[1], Junjie Chen, Ph.D[2], Hua Xu, Ph.D[1]
[1]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA
[2]Department of Experimental Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA

Discovery of the safe and effective drugs and their targets play important roles to improve the health care. However, the process is very time-consuming and expensive, especially for complex diseases. Recently, computational approaches for target prediction and drug repurposing have become more common and effective compared to traditional observation-based drug discovery. However, since the data about underlying molecular mechanisms of drugs distribute among different knowledge domains and different databases, it is very challenging to design effective strategies to discover novel drug targets and propose successful drug repurposing. To alleviate this problem, we propose a computational framework to integrate complex relationship among different types of data and infer the potential drug targets by using the semantic web technology, and to improve performance through network neighborhood effect modeling. In this study, we utilize the colorectal cancer (CRC) as a proof-of-concept use case to evaluate the approach.

We collected the FDA-approved drugs to treat CRC from the NCI Cancer Drug Information website. Using these CRC drugs, we constructed a CRC ontology including drugs, diseases, genes, pathways, SNPs, and their relations from the database PharmGKB. From the CRC ontology, we inferred potential CRC drug targets using specified Web Ontology Language description logic rules. Starting from these potential targets, we first ranked them based on the fraction of CRC disease genes in their neighborhood at the first, second, and third shortest-path distances and then integrated the three sets of ranking using a robust rank aggregation (RRA) method. These CRC disease genes were collected from the Cancer Gene Census, the Online Mendelian Inheritance in Man (OMIM), and the Genetic Association database (GAD). Finally, from the significantly ranked genes, we inferred the drugs for CRC treatment based on the drug-target association from the DrugBank and the therapeutic target database (TTD).

Starting from eight FDA-approved CRC drugs and the CRC ontology, we inferred 113 potential CRC drug targets. The 113 genes were significantly enriched with CRC disease genes when compared to the 20,737 human protein-coding genes, which was more than expected by chance (Hypergeometric test *P*-value: 4.96×10^{-10}). Additionally, they were significantly enriched in the pathways related to drug metabolism by KEGG pathway enrichment analyses. Among the 113 genes, 15 were selected as the promising drug targets based on their neighborhood of CRC disease genes in the context of one human PPI network. For example, three of them encode known CRC drug targets (*EGFR*, *TOP1*, and *VEGFA*). Considering that the drugs targeting the promising CRC target genes could be used to treat CRC, we inferred 73 potential drugs for CRC treatment by mapping the 15 promising drug targets to drug-target association from the DrugBank and TTD data. We validated these drugs by comparing their overlap with the drugs studied in CRC clinical trials. We found that, among the 73 drugs, 18 have been investigated in at least one CRC clinical trial. For example, the drug celecoxib has been investigated in the 35 clinical trials for CRC either alone or in combination with other drugs. Celecoxib is an anti-inflammatory drug and a selective COX-2 inhibitor used for the treatment of osteoarthritis, rheumatoid arthritis, acute pain, painful menstruation, and menstrual symptoms. It has been regarded as a promising drug for the treatment of familial adenomatous polyposis (FAP), an inherited disorder characterized by cancer of the large intestine (colon) and rectum. The results indicate that our computational framework is promising for drug target prediction and drug repositioning.

Therefore, in this study, we developed a unique computational framework to integrate the ontology technology and network neighborhood modeling for drug target prediction and drug repurposing. The results demonstrate that this framework indeed identifies the novel targets and drug repurposing. Besides, we see many opportunities to improve upon the basic design of this integration, including integrating more relationship from multiple data sources during the ontology construction, application of score strategies in the ontology reference, and the combination of more network properties into the gene ranking.

Journaling and Journal Retrieval: An Information Management Tool to Assist Chronic Patients' Everyday Self-Management

Si Sun, MS
School of Communication and Information, Rutgers University, New Brunswick, NJ.

Abstract

An information management tool incorporating journaling and information retrieval features is designed to help people living with chronic diseases understand their health conditions and make health-related decisions. The study is carried out in two stages. In stage one, a prototype system is designed and five experienced type 2 diabetes patients provided feedback regarding the system during a semi-structured interview. In stage two, a smartphone application will be developed and its usability tested with 20 diabetes patients. This poster presents the results of stage one.

Introduction

Information management is a major barrier for the self-management of chronic conditions. Information overload, information fragmentation, and the ephemeral quality of daily information make it particularly difficult for patients to keep track of their health condition and act upon its development. In this circumstance, two methods – journaling and mapping tools – have attracted increased attention as important venues to help patients garner their daily health information. This study attempts to develop a tool with the combined advantages of both methods.

Background

Journaling that uses unstructured data and visualization of structured health information are two major methods that could promote healthful behaviors. For example, journaling can achieve this through increasing patients' awareness toward their health conditions[1]. Also, visualizing information can provide patients with health records for self-education[2]. However, both methods have issues that prevent them from reaching their potentials. For example, information overload is a common issue with journaling that make it hard for patients to recall the existence of useful information[3]. Also, a problem with structured health information is that it does not support the vast diversity of patient behaviors that can differ between individuals[3]. This study presents a smart phone based application that can potentially offer the benefits of both journaling and visualization and compensate for their respective issues by enabling the visualization of patients' health journals.

Methods

The initial prototype is a series of computer aided drawings on paper that feature the major functions of the app as they would appear on an actual smart phone. These features include (a) journaling, where patients can enter journals in free text; (b) browsing, where patients can review previous journals sorted by date and time; (c) searching, where patients can search for previous journals using keywords (fuzzy search) and dates; and (d) visualization, where the journal search results can be displayed in word usage frequency charts (when searching with keywords) or word cloud (when searching by date, using specialized thesaurus including AGROVOC for food and UMLS Metathesaurus for health-related terms). For each text box and button, a short floating text regarding their utilities is displayed.

Five diabetes patients were recruited from the author's personal network to participate in a semi-structured interview regarding the utility of the prototype and possible improvements. These patients all lived with diabetes for more than ten years and self-reported to manage diabetes information on a daily basis using electronic devices. Before the interview, patients reviewed the functions of the prototype system and watched the author's demonstration.

Preliminary Results and Discussions

An improved prototype system is developed based on patients' feedback. One of the major improvements is to include a planning module that utilized the journal retrieval function to persuade short-term behavior change. This new module enables patients to enter plans for the day. When this happens, the app retrieves and displays relevant previous journals based ranked by the number of words matched between the new entry and previous journals. This new module allows patients to easily reflect on their self-management records and make better informed decisions. Other suggestions including offering social sharing and a reputation system that would further engage patients through social support and peer pressure are also recorded for future developments. During the poster presentation, the author will demonstrate the improved prototype system on a poster with personas and user scenarios included.

This system could manifest patient experiences that are perhaps not easily visible in generic journal records, and help patients gain better understandings of their health conditions through engaging in a rich collection of their personal health records. Further, this system could provide a new basis for decision making and information sharing in both home and clinical settings.

Reference

1. Brown B, Chetty M, Grimes A, Harmon E. Reflecting on health: a system for students to monitor diet and exercise. In: *CHI EA '06*. ACM Press; 2006:1807.
2. Ikeda Y, Tsuruoka A. Self-monitoring of blood glucose, as a means of self-management. *Diabetes Res Clin Pract*. 1994;24:S269-S271.
3. Kobsa A, Chen Y, Wang T. Discovering personal behavioral rules in a health management system. In: *PersvasiveHealth '12*.; 2012:224-227.

Bridging the Gap in Transfusion Medicine: From Data to Decision

Nazanin Tabesh-Saleki, MS[1, 2], Mary E. Herwig, MBA[2], Julie F. DeLisle, MSN, RN[2], Nanci L. Fredrich, BSN, RN, MM[2], Sandra J. Butschli, BS[2], Timothy B. Patrick, PhD[1], Kathleen E. Puca, MD[2]

[1]Univerisity of Wisconsin-Milwaukee, Milwaukee, WI; [2]BloodCenter of Wisconsin, Milwaukee, WI

Abstract

Forty to sixty percent of blood transfusions are considered inappropriate[1], providing either no benefit, or heightened risk of complications, and increased cost of care. Addressing the variability in transfusion practice requires evidence-based guidelines, focused education, and meaningful presentation of information to drive appropriate decision-making and interventions. A focus on tight integration of the above elements bridges the gap that exists between knowledge and practice, and thereby can alter behavior.

Introduction: Historically, transfusion of two units of Red Blood Cells (RBC) has been the norm[2]. This traditional practice has come under intense scrutiny lately due to an increased awareness of the association of RBC transfusions with complications and worse outcomes, which are dose-dependent[3, 4]. Patient Blood Management (PBM) is a multidisciplinary, patient-centric, data driven approach designed to optimize utilization of blood products in patients who may require a transfusion in an effort to improve outcomes. Implementation of a robust hospital-wide PBM program is challenging; it requires leadership support, involvement of diverse groups of care providers, re-education of providers with ingrained transfusion practices, and standardization of blood utilization guidelines. Capturing relevant data plays a critical role in supporting the program, as derived information provides a baseline for current practice, both department and provider-specific, and can identify areas for improvement. For this study, efforts were focused on capturing data to reflect provider transfusion practices, specifically on ordering and transfusing one versus two units of RBC. **Methods:** Through a joint effort between transfusion medicine and information technology experts in collaboration with a local hospital, we have developed a set of PBM key performance indicators (KPI) to monitor transfusion practice. The development of PBM-focused KPIs included identification of data sets for extraction from the hospital's information systems. The data files were transferred to a central repository via a secure connection. Through use of third party software, we defined a unified metadata layer and mapped data elements for context sensitive views of information. The information was presented back to the hospital via a secure web-based dashboard, which depicted provider ordering practice at multiple levels including, by facility, service line, and provider. We coupled our efforts with provider education in the form of PBM newsletters and quarterly reports to department chairs highlighting RBC ordering practice. **Results:** Since our implementation of PBM the percent of one-unit RBC orders increased and simultaneously the percent of two-unit RBC orders decreased. Figure 1 represents the gradual divergence in provider ordering practice. In addition, an ongoing decline in total RBC unit transfused was noted. Using first quarter in 2013 (Q1-13) as the baseline, there has been a steady reduction in percent RBC units transfused – 3.99% (Q2-13); 8.31% (Q3-13); and 11.8% (Q4-13). **Conclusion:** Through capturing relevant data from disparate source systems and transforming it into meaningful information, we were able to effectively focus PBM strategies. The data from this one KPI along with focused education has resulted in lower RBC utilization. Our approach provided relevant metrics on provider RBC ordering practices and raised awareness around opportunities associated with PBM.

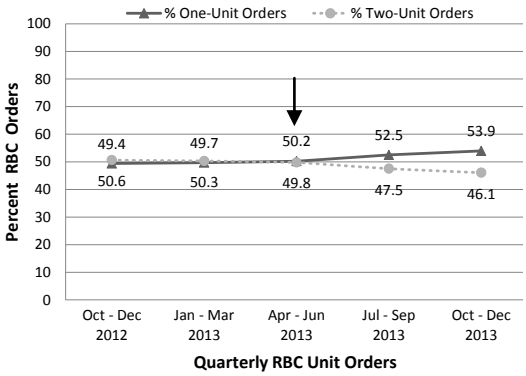


Figure 1. Trend in providers' ordering practice; representing normalized percent one-unit versus two-unit RBC orders. The arrow marks May and June 2013, when the first reports were distributed to hospital department chairs and the first PBM newsletter was disseminated to the providers, respectively.

References

1. Shander A, Fink A, Javidroozi et al International Consensus Conference on Transfusion Outcomes Group. Appropriateness of allogeneic red blood cell transfusion: the international consensus conference on transfusion outcomes. Transfus Med Rev. 2011; 25(3):232-246.e53.
2. Berger MD, Gerber B, Arn K, Senn O, Schanz U, Stussi G. Significant reduction of red blood cell transfusion requirements by changing from a double-unit to a single-unit transfusion policy in patients receiving intensive chemotherapy or stem cell transplantation. Haematologica 2012;97:116-22.
3. Marik PE and Corwin HL. Efficacy of red blood cell transfusion in the critically ill: A systematic review of literature. Crit Care Med 2008;36(9):2667-74.
4. Ferraris VA, Davenport DL, Saha SP, et al. Surgical outcomes and transfusion of minimal amounts in the operating room. Arch Surg 2012;147(1):49-55.

Genetic Variant Databases: Current Practices for Development and Curation

Keeon Tabrizi, MEng[1], Robert Coopersmith, Ph.D.[1], David Hardison, Ph.D.[1], Elisabeth L. Scheufele MD, MS[1,2], Matvey B. Palchuk, MD, MS[1,2]
[1]ConvergeHEALTH by Deloitte, Deloitte Consulting LLP, Newton, MA; [2]Harvard Medical School, Boston, MA

Abstract
Disease-associated genetic variant databases play an important role in utilizing and interpreting variant data in a clinical context. We have performed an analysis to better understand best practices regarding evidence standards for variant classifications, curation practices, and to assess overall quality systems for gathering and maintaining high integrity data. The analysis included public locus specific databases (LSDB's), internal databases at commercial entities (clinical labs and biopharmaceutical companies), ClinVar at NCBI, and those at academic medical centers. A representative subset of databases was then chosen for structured interviews to obtain information that was not publically available. The interview was comprised of a series of questions developed from criteria to evaluate evidence quality and curation practices. Interviews were summarized respective to the evaluation criteria and an assessment was completed across the databases and against several highly-regarded and referenced databases.

Introduction
Rapid advances in genomics and informatics are augmenting the scope of personalized medicine to incorporate genomic data into medical decisions. By linking data obtained from next generation sequencing and other high throughput technologies with preclinical data, clinical outcomes, and other available patient information (clinical, demographics, etc.), it is possible to distinguish clinically actionable genetic variants from the larger collection of variants of unknown significance.[1] This process requires aggregation of good quality evidence from multiple sources. However, many of the ultra high-throughput genomic tests identify a significant number of variants where clinically meaningful phenotype or outcome has not been well-established. Thus, potential risk is introduced to patients if medical decisions are based on genomic information of unclear or ambiguous significance.[2]

Problem
Currently, issues with data quality and classification standards are not well-understood by the broader community of academic and commercial researchers, clinical laboratorians, physicians, patients, and regulators. These issues often limit the utility of databases when used to inform important decisions regarding medical practice. This need is underscored by research initiatives such as a recent $25 million NIH grant to bridge the various issues and develop a framework for evaluating variants relevant to patient care in disease.[3] While biomedical databases can collect data from a wide variety of sources and facilitate the capture and aggregation of information, the reliability of the data within any particular database relies on several factors, including the quality of the genomic data stored in the database, the amount and type of preclinical and clinical information associated with that data, and the curation practices instituted by that database.

Solution
We have conducted a study to survey the current landscape of disease variant databases and in particular to investigate curation practices and the use of evidence to classify variants. We included public LSDB's, internal databases at commercial entities (diagnostic and biopharmaceutical companies), and those at academic medical center diagnostic units. Information from publically available sources was used to develop criteria (Figure 1) for evaluating the evidence quality and curation practices. A representative subset of databases was then chosen for structured interviews based on a series of questions developed from the criteria. We found that parameters varied widely among databases, both across and within classes, while others were more consistent. Based on our review we propose a set of leading practices based on the examined database attributes.

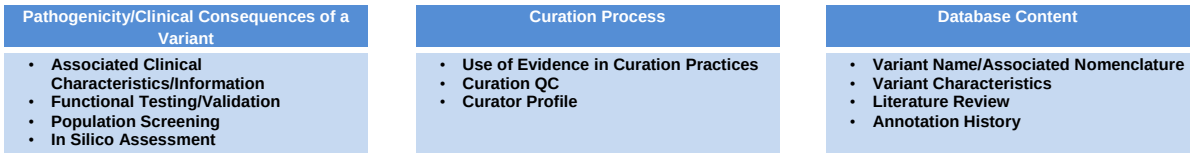


Figure 1: Evaluation Criteria

References
1. Facio, FM, et. al. (2013). A Genetic Counselor's Guide to Using Next-Generation Sequencing in Clinical Practice. Journal of genetic counseling. doi:10.1007/s10897-013-9662-7
2. Duzkale, H, et. al. (2013). A systematic approach to assessing the clinical significance of genetic variants. Clinical genetics, 84(5), 453–63. doi:10.1111/cge.12257.
3. New NIH-funded resource focuses on use of genomic variants in medical care. (2014, September 25). U.S National Library of Medicine. Retrieved February 9, 2014, from http://www.nih.gov/news/health/sep2013/nhgri-25.htm

Information Use and Performance in an Accountable Care Organization

Douglas A. Talbert, PhD[1], Harold Chertok, DO[2], Kenneth R. Currie, PhD[1], Rebecca L. Turpin, MSN[1], Steve Talbert, PhD, RN[3]
[1]Tennessee Technological University, Cookeville, TN;
[2]Cumberland Center for Healthcare Innovation, Cookeville, TN;
[3]University of Central Florida, Orlando, FL

Introduction

An Accountable Care Organization (ACO) is a group of health care providers working together to coordinate care with the goal of delivering high quality, cost effective care.[1] When an ACO is able to satisfy the Medicare-defined quality standards at a cost lower than the threshold set by Medicare, the achieved savings are shared with the ACO.

Cumberland Center for Healthcare Innovation (CCHI) is an ACO consisting of 40 primary care physicians in 27 practices across the 14-county rural Upper Cumberland region of Tennessee, an area with documented higher health care costs and lower health care quality than the national average. CCHI's Medicare costs (per patient cost) and quality (percent of quality metrics satisfied) vary widely across member practices. Identifying the causes of these variations could help CCHI achieve the quality and costs needed to improve care and realize shared savings.

Methods

Studies[2] and national initiatives such as the HITECH Act[3] suggest that the availability and use of information in the clinical decision-making process influence health care cost and quality, but to the best of our knowledge, the connection between information use and performance across an ACO has not been studied. To improve our understanding of the extent to which information use explains the variation across the practices in CCHI, we surveyed the practices regarding their use of information in preparing for patient visits. Table 1 shows the high-level measures on which the survey and this study focused.

Table 1. Survey measures related to clinical information use.

A. Information captured during initial contact for a visit	D. Clinical info. updated by nurse or med. asst during initial patient review
B. Nature of the use of that initial info. for visit planning	E. Clinical info. reviewed by physician or physician extender prior to visit
C. Clinical info. reviewed by nurse or med. asst prior to visit	F. Does visit planning occur prior to or after patient arrival?

Of the 27 practices surveyed, four are new to the ACO and do not have Medicare-reported cost and quality metrics, five other practices did not return a complete survey, leaving 18 practices for which we had both cost and quality metrics and information-use metrics.

Results

We identified the most and least expensive practices as outliers. An analysis of the 16 remaining practices showed the correlation of measure *E* with *cost* (r = -.536) was significant (p = .023). Backward regression found that the model for *cost* that contained measures *D* (β = -.251) and *E* (β = -.470) was also significant (p = .049). The *quality* correlation matrix showed moderate but not significant correlation with measures *D*, *E*, and *F*, and while regression did not identify any independent predictors of *quality*, measure *E* had the highest standard coefficient (β = .537).

Discussion

With such a small sample size, it is not surprising that we found so few significant factors, but these initial findings suggest that, of the factors we examined, information use by the physician (or physician extender) in preparing for the visit had the most influence on cost and quality. We are continuing to study the impact of these and other factors (such as visits per day) on the cost and quality of health care.

References

1. http://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/ACO/
2. Chaudray B, *et al*. Systematic review: impact of health information technology on quality, efficiency, and cost of medical care. *Ann Intern Med* 2006; 144:742-52.
3. http://www.healthit.gov/policy-researchers-implementers/hitech-act

A Time-and-motion Study of Clinical Trial Eligibility Screening in a Pediatric Emergency Department

Huaxiu Tang, MS, Melanie Hounchell, BA, CCRC, Judith W Dexheimer, PhD,
Stephanie Kennebeck, MD, Imre Solti, MD, PhD, MA Yizhao Ni, PhD
Cincinnati Children's Hospital Medical Center (CCHMC), Cincinnati, OH

Abstract

Determining patient eligibility is a major barrier to clinical trial enrollment. We conducted a time-and-motion study on the clinical trial eligibility screening workflow in a pediatric emergency department (ED). We observed the workflow activities and the major findings are presented. Most (50%) of the time is spent screening patients and performing procedures; the time spent walking and waiting suggests that rearranging work locations could save time.

Introduction

Eligibility screening (ES) is widely applied to determine the suitability of candidates for clinical trials. Time-and-motion has been used to evaluate efficiency of clinical activities[1]. The goal of this study is to investigate the workflow of clinical trial ES in a pediatric ED to identify inefficiencies that may be streamlined to improve workflow.

Methods

Fifteen 120-minute observations were conducted from 12/12/2013 to 1/3/2014, including 5 morning 5 afternoon and 5 evening observations. The workflow of eight research assistants (RA) was tracked. One RA was shadowed in each session and their activities were recorded in 30-second increments. The activity categories are listed in the table, with associated sub-categories. Time spent was summarized with mean and standard deviation (STD). The numbers of patients screened, approached and enrolled were also recorded.

Table. Time and percentage over all time spent on workflow activities.

Category	Subcategory	Minutes per Section	Percent
Patient Screening (31.62%)	Reading EHR content on screen	36.0 (±20.18)	29.4%
	Discussing patient's eligibility with clinical staff	2.67 (±2.01)	2.18%
Patient Contact (18.67%)	Ask patient to join in research	4.03 (±2.72)	3.30%
	Review/Confirm patient's eligibility	0.07 (±0.25)	0.05%
	Consent procedures	4.73 (±3.65)	3.87%
	Record data	8.67 (±11.29)	7.09%
	Waiting for patient to finish study procedure	3.83 (±5.13)	3.13%
	Other activities in patient contact	1.50 (±2.42)	1.23%
Performing Procedures (17.63%)	Performing procedures/ Documenting eligible patients	18.5 (±12.39)	15.2%
	Documenting ineligible patients	3.03 (±4.10)	2.48%
Physician Contact (1.09%)	Waiting for physician to finish study procedure	1.33 (±2.05)	1.09%
Other Activities (30.99%)	Task administration	14.2 (±7.46)	11.6%
	Emails/Web browsing	2.97 (±5.03)	2.43%
	Waiting	7.70 (±9.34)	6.30%
	Walking	12.6 (±5.87)	10.3%
	Personal time	0.47 (±1.50)	0.38%

Results

The *Patient Screening* and *Performing Procedures* activities (e.g. documenting eligible patients) occupied 50% of the RAs' time. The RAs spent 10.3% of their time *Walking* and 6.3% *Waiting*. The RAs approached 30 out of 197 screened patients and enrolled 20 subjects (enrollment 66.7%). Among the 10 cases in which patient declined enrollment, 8 occurred in evening, 1 in morning and 1 in afternoon. The decline rate for the evening is 57.1%, which is statistically significantly ($p<0.001$) different from 14.3% for the morning and 11.1% for the afternoon.

Discussion & Conclusion

Patient Screening and *Performing Procedures* occupy half of the RAs' time. This can be potentially reduced by computerized approaches such as automated ES and documenting. RAs spent a notable amount of time in *Walking* and *Waiting*, suggesting that relocating the workspace could be helpful. Patients declined more in evening shifts; the result may not be significant due to limited observation sections; future work will be to investigate the difference of decline rate. The findings suggest workflow improvement areas to increase the efficiency of clinical trial ES.

References

1. Yen, K., Shane, E. L., Pawar, S. S., et al. Annals of emergency medicine. 2009; 53(4): 462-468.

Reducing Readmissions and Intermountain Discover App: Stratification of Interventions based on Risk Assessment

David P. Taylor, PhD, Farrant H. Sakaguchi, MD, MS, Adam Kraft, MD, Kira Wagner, BS
Intermountain Healthcare, Salt Lake City, UT

Abstract

Hospital readmission reduction is a health policy priority. Interventions, such as ProjectRED, ProjectBOOST, Care Transitions Interventions, and Camden Health have shown efficacy at decreasing readmissions. However, these interventions would likely benefit from personalized risk assessments and stratifications of risk. We seek to marry risk prediction tools, such as LACE, HOSPITAL, and LACE + with current interventions through development of a web application and iPad app within our clinical environment. This application will improve flexibility to evaluate different algorithms, provide greater collaboration among care team members and highlight the most useful data.

Introduction

Reducing hospital readmissions is a national health policy goal with financial penalties for readmissions within 30 days. ProjectRED, BOOST, Care Transitions Interventions, and Camden's Care Management Program are interventions to decrease readmissions and have demonstrated promising results. However, tailoring the interventions to the individual risk and needs of each patient will likely be more effective. Tools such as LACE, HOSPITAL, and LACE+ have been developed to predict the risk for readmission. Stratification and tailoring of interventions based on personalized risk calculations are necessary to help manage constrained resources.

While the electronic health record (EHR) may be an ideal platform to standardize risk assessments and to present appropriate protocols, most current EHRs lack the flexibility to implement and rapidly evaluate different risk prediction tools, and intervention strategies and changes can be resource intensive. A flexible knowledge discovery tool where end-users can easily create questionnaires and checklists while pulling data from the EHR facilitates the comparison of different prediction tools and interventions.

Approach

The Homer Warner Center for Informatics Research (HWCIR) has developed Intermountain Discover, a HIPAA compliant form tool that can pull information from the EHR to prepopulate forms. The tool is available as both a web application and an iPad app, has dashboard capabilities and basic data visualizations. The interface can support workflows, guiding users through series of questions and acting as a dynamic checklist of appropriate interventions.

We have created a series of forms in Intermountain Discover to perform risk assessments based on LACE, HOSPITAL, an adapted LACE+, and Camden's Care Management Program. Services will extract data from the EHR as appropriate. We have tailored different interventions based on components of ProjectRED and BOOST to be based on predicted risk for readmission as well as specific risk factors. We will use the enterprise data warehouse to prospectively evaluate readmissions. We will evaluate which risk assessment tool has the greatest predictive value for our patients and which interventions have the greatest impact to reduce readmissions. We anticipate beginning to gather data for the hospitalist services at two local hospitals within the coming month.

Conclusion

Flexibility is needed to evaluate the performance of different risk stratification strategies and personalized interventions to reduce readmissions. Developing protocols to be implemented directly within the EHR poses potential disruption and user frustration. The use of a tool that can be viewed by all members of the care team is likely to provide greater collaboration and insight, while clarifying what data are most useful.

While the use of standard terminology services will be a part of future development, there is an immediate and great need for research tools that are able to pull data directly from the EHR and that facilitate flexible documentation.

Acknowledgement: Matt Ebert, Developer; Dr. Linda Venner, Study Co-PI, and Clinical Coordinators Krystal Allan, Lori Stohler, and Pualani Kros.

A Qualitative Assessment of CPOE and Variation in Drug Name Display

Thu-Trang Thach, MPH[1], Alexandra Robertson[1], Arbor J. L. Quist[1], Lynn Volk, MHS[2], Adam Wright, PhD [1,2,3], Shobha Phansalkar RPh, PhD[1,3,5], Sarah Slight MPharm, PhD, PGDip[1,4], David W. Bates MD[1,5], Gordon D. Schiff MD[1, 5].

[1]Brigham and Women's Hospital, Boston, MA; [2]Partners HealthCare, Wellesley, MA; [3]Wolters Kluwer Health, Indianapolis, IN; [4]School of Medicine, Pharmacy and Health, University of Durham, UK; [5]Harvard Medical School, Boston, MA.

Abstract

Medication safety is an important goal of electronic health record (EHR) implementation. Refining the delivery and content of existing EHRs is a key regulatory initiative of the current US healthcare administration. This study aims to evaluate the impact of drug-name display on the potential for medication errors in computerized provider order entry (CPOE).

Background

Previous work suggests that prescribing errors account for the highest proportion of medication errors[1] and that 1 in every 4 reported in the U.S. can be linked to drug name confusion.[2] CPOE has been widely proposed, implemented, and shown to reduce errors associated with traditional, handwritten prescribing. Still, CPOE has not fully lived up to its predicted benefits, and many attribute these shortfalls to suboptimal system design. The FDA Brigham and Women's Hospital CPOE Medication Safety Project was funded by the FDA to assess drug name display issues and understand their potential to contribute to medication errors.

Methods

We assessed look-alike-sound-alike pairs, adjacency errors, font, and visual display in 10 CPOE systems at 6 U.S. medical centers via remote walk-throughs and on-site visits. We created a CPOE Assessment Tool incorporating medication ordering, review, and deletion scenarios to guide walk-throughs with a series of "regular" and "expert" users of these systems. Analysis of screen shots, video, and audio transcripts from these evaluations provided guidance for visits at each site to collect additional system information from medical leadership, IT experts, and CPOE users. Our examination of inter- and intra-system characteristics revealed patterns, themes, and variations.

Results

We identified a lack of standardization in the display of drug names in CPOE systems, which could contribute to medication errors. Issues and themes that emerged included: *(a) Brand versus generic name* – some systems limited search by brand or generic names, some systems displayed results either by either brand, generic or both names, and some system displays encouraged prescribing of preferred (typically generic) products. *(b) Combination products* – searches for combination drugs often yielded the brand name with the ingredients listed parenthetically and rarely provided explanations of ingredient strengths. In some systems, searching the brand name yielded only the brand name, without its generic ingredients. *(c) Drug-name modifiers* – drug name modifiers had some consistency, appearing primarily at the end of the drug name, in uppercase, and were rarely truncated but were often abbreviated (e.g. XR, DS). *(d) Extraneous additions to drug name fields* – we observed non-systematic population (jerry-rigging) of name fields with additional information ranging from drug indication to facility-specific availability. *(e) Font, text size, capitalization* – these varied widely with no standardized approaches seen either across or, often, within systems.

Conclusions

There is little consistency in the ways that CPOE systems search for and present drug names. This lack of standardization was seen both across and within systems. Brand vs. generic, combination products and drug name modifiers were particularly inconsistent, each presenting the potential for user frustration, confusion, and erroneous orders.

References:

1. Lisby M, Nielsen LP, Mainz J. Errors in the medication process: frequency, type, and potential clinical consequences. Int J for Quality in Health Care 2005; **17**(1): 15-22.
2. Lambert BL, Lin SJ, Chang KY, Gandhi SK. Similarity as a risk factor in drug-name confusion errors: the look-alike (orthographic) and sound-alike (phonetic) model. Med Care 1999; 37(12): 1214-25.

The impact of an automated response function in an electronic checklist on checklist accuracy: an observation from a simulation-based study

Charat Thongprayoon, MD; Andrew M Harrison, BS; John C O'Horo, MD, MPH; Ronaldo A Sevilla Berrios, MD; Brian W Pickering, MBBCh, Herasevich V, MD, PhD
Multidisciplinary Epidemiology and Translational Research in Intensive Care (METRIC), Mayo Clinic, Rochester, MN

Background:. With increasing prevalence of electronic medical record (EMR) systems, the utility of a checklist integrated into the EMR with decision support tools has been increasing apparent. The automated response function in a smart checklist is designed to individualize the checklist for each patient, which can potentially reduce the time to checklist completion. However, it may simultaneously increase checklist errors if providers are less engaged when using a smarter checklist. In a secondary analysis of a simulation-based study, we aimed to investigate the effect of the automated response function in an electronic checklist for checklist accuracy, as completed by intensive care unit (ICU) physicians.

Methods: We conducted a simulation-based study at an academic tertiary center. A total of 21 ICU physicians participated in this study. Each participant completed checklists for 3 ICU patients using an electronic checklist in a controlled usability lab setting with full access to the EMR system.. The automated response function was programmed with a detection algorithm using patients' real time data to select only applicable checklist items for participants to complete. Two independent clinician-reviewers served as the reference standard for checklist error calculation. The association between automated response errors and checklist errors per participant was assessed using the Chi-squared test.

Results: The number of checklist errors by ICU physicians and the number of automated response errors for each checklist item are summarized in Table 1. Although participants were advised to disregard the automated response, , the automated response errors in the "lung protective ventilation", "spontaneous breathing trial", and "adequate source control" items still resulted in an increased risk of checklist errors by participants.
Table 1. The number of checklist errors by ICU physicians and the number of automated response errors.

Checklist item	# of Checklist errors	# of Automated response errors	p-value*
Adequate source control	16	33	<0.001
Lung protective ventilation	9	15	<0.001
Spontaneous breathing trial	3	6	<0.001
Remove urinary catheter	15	3	0.07
Sedation break indicated	6	3	0.56
Continue vasoactive medication	5	9	0.70
Treat pain	1	2	0.86
Need for antimicrobial reviewed	1	1	0.90
Treat delirium	0	1	N/A
Remove vascular device	6	0	N/A
Head of bed elevated	0	6	N/A

*p-value from Chi-Squared test to assess association between checklist errors and automated response errors

The discrepancy in defining the applicability of checklist items between the detection algorithm and the reference standard was the main reason for checklist errors.

Conclusion: In previous studies, the electronic checklist was shown to significantly reduced provider workload and overall checklist errors However, the automated response function equipped in electronic checklist might inadvertently increase checklist errors in some items. Therefore, these automated response functions must be well validated and refined to achieve optimal accuracy before implementation in a real electronic checklist used in the ICU setting.

Value Set Management to Enable Interoperable Clinical Decision Support: Development, Use, and Initial Evaluation of the OpenCDS Value Set Manager

Tyler J. Tippetts[1], Phillip B. Warner, MS[1], David E. Shields[1], Salvador Rodriguez-Loya[2], Catherine J Staes, BSN, MPH, PhD[1], Kensaku Kawamoto, MD, PhD[1]
[1]University of Utah Department of Biomedical Informatics, Salt Lake City, UT;
[2]University of Sussex, School of Engineering & Informatics, East Sussex, United Kingdom

Abstract
To enable interoperable clinical decision support (CDS), clinical concepts referenced within CDS knowledge resources must be mapped to standard terminologies. While value sets from standard terminologies are increasingly becoming available, additional value sets are still required to support CDS. To address this problem, a Web-based Value Set Manager was developed by the OpenCDS initiative that leverages the Unified Medical Language System. Experience with implementing and using this resource, including challenges and future directions, is described.

Introduction
Value sets are lists of specific values derived from one or more standard terminologies to define concepts.[1] The definition of value sets is an essential requirement for a CDS knowledge resource to be semantically interoperable across institutions. Resources such as the National Library of Medicine's Value Set Authority Center (VSAC) [1] and the Public Health Information Network Vocabulary Access and Distribution System (PHIN VADS) [2] include value sets that can be used for CDS purposes. However, there is often a need to create additional value sets for CDS purposes. The objective of this effort was to develop an open-source, Web-based value set management tool that leveraged available terminology resources and enabled the efficient definition of value sets relevant for CDS.

Methods
This project was conducted within the context of the OpenCDS initiative (www.opencds.org), which is a multi-institutional collaborative effort to develop open-source, standards-based tools and resources for CDS. OpenCDS uses internally defined concepts which are then mapped to value sets. To facilitate the development of value sets for mapping to OpenCDS concepts, a Web-based Value Set Manager for CDS was developed using the Grails Web application development framework, the Google Web Toolkit user interface designer, and the National Library of Medicine's Unified Medical Language System (UMLS). System performance was evaluated for typical use scenarios.

Results
The UMLS database was instantiated locally. An application programming interface (API) was developed to allow various terminology operations on the UMLS, such as searching for codes within a code range (e.g., ICD9CM 250-250.99) and for codes subsumed by a parent concept (e.g., codes that are descendants of ICD9CM 250). Leveraging this API, the Web-based Value Set Manager enables value sets to be defined within a given code system (see Figure) and mapped to OpenCDS concepts. Identifying codes within a range is essentially instantaneous, and identifying codes subsumed by a parent concept typically takes less than five seconds to complete. Challenges included differences in how various terminologies are represented in the UMLS and the need for performance optimization for some terminology operations. Future directions include supporting multiple terminologies within a single value set and facilitating the editing of externally defined value sets.

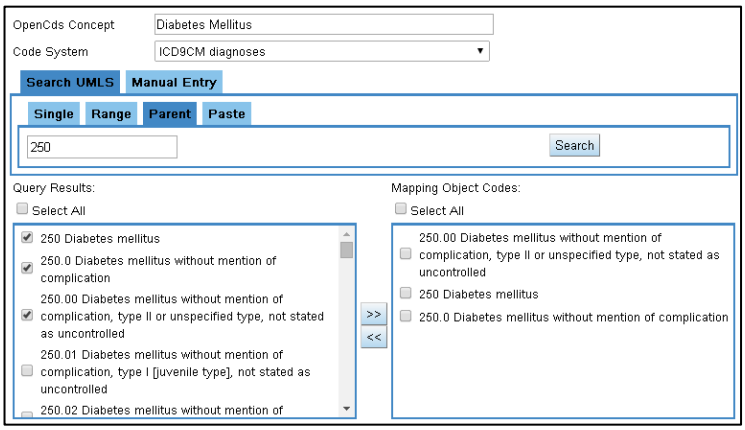


Figure. Screenshot of OpenCDS Value Set Manager

Conclusions
Specifying value sets with standard terms is essential to enabling interoperable CDS. The OpenCDS Value Set Manager fulfills this need efficiently and will be made freely available as a part of the OpenCDS initiative.

References
1. National Library of Medicine. Value Set Authority Center (VSAC). 2014. Available at: https://vsac.nlm.nih.gov.
2. CDC. Public Health Information Network Vocabulary Access and Distribution System (PHIN VADS). Available at: https://phinvads.cdc.gov.

Contextualization of Comparative Assessment Reports and System-wide Decision Making for Drugs and Devices Reimbursement

Michele Tringali, MD[1], Silvia Vecchio, PhD[2], Emanuele Lettieri[3], Luca Romano[2]
[1]Regione Lombardia, Milano, Italy, MA; [2]ASL Pavia, Pavia, Italy; [3]School of Engineering, Milano, Italy

Abstract

Streamlining assessment and appraisal, for both emerging and established health care technologies, can be difficult and inconsistently applied within socialized systems like Italy's NHS as well as and by insurance-driven health care plans elsewhere. An information framework incorporating state of the art models from European HTA assessment reports and tools from the EVIDEM (Evidence and Values) Collaboration was successfully developed and used in Lombardy region for coverage decision on 13 health care technologies.

Background

Comparative Effectiveness Research (CER) and similar Health Technology Assessment (HTA) products for drugs, medical devices and other health care technologies require localization, i.e. adaptation of global information to contextual factors, in order to be used for informing reimbursement from payers. Real world processes of decision making are seldom formally modeled, and Multiple Criteria Decision Analysis (MCDA) approaches and tools can help legitimize robust appraisal practices and streamline communication to providers, doctors and patients.

Methods

Lombardy region's (10 million people) health care authority designed, tested and implemented an information framework with the aim to guide local experts in contextualizing national and European HTA reports prepared by AgeNaS (agency for medical devices) and EunetHTA[1] (the European Network of HTA that encompass all 27 Member States in the EU) to be efficiently used by a Priority (for emerging) and an Appropriateness (for diffusing or obsolete technologies) Committees to prepare value-based policy decisions, i.e. managed adoption or refusal and maintenance or delisting of drugs, devices, diagnostic and screening procedures. The EVIDEM Framework and set of MCDA tools[2] were implemented with minor modifications[3]. Data from four EUnetHTA Reports (Genomic tests for breast cancer adjuvant therapy, EndoBarrier for obesity, Renal cancer drugs, Renal denervation for drug resistant hypertension) were semi-automatically extracted using two templates (Model and Report files) and imported into an Access database, extracted in XML format then imported (FlexiC Import extension) into a Joomlaa 2.5 CMS organized around a hierarchical set of 9 Dimensions, 20 Criteria and a variable number of Subcriteria and Issues (example in Table 1). Domain experts revised and added content (epidemiology, financial impact, clinical practice, organizational issues areas).

Table 1. A sample Issue for Criteria *C08 - Improvement of efficacy/effectiveness* for the Dimension *D4 - Clinical effectiveness*.

D4 - Clinical effectiveness
C08 - Improvement of efficacy/effectiveness
S01 - Mortality and morbidity
D0002 - What is the expected beneficial effect on the disease-specific mortality?

The localized report was then used by the Appraisal Committee to structure complete decisions through MCDA methods (weights elicitation and discussion, scores and judgments collection, qualitative analysis of comments and elaboration of an Appropriate Use Index for each technology). A simplified version of the process, limited to an assessment of the first layer (dimensions) of the hyerarchy, was used by the Priority Committee with nine new and emerging technologies for which Horizon Scanning (HS) reports has been made available by Age.Na.S.

Results

The regional authority was able to elaborate contextually relevant coverage decisions on several health technologies with both initial (HS) or more complete (HTA) information.

Conclusion

An integrated information framework built around a 4-layers simple hierarchy of operational meanings for technical assessment, localization and decision analysis, adapted from established models for HTA and MCDA, can be applied to a wide range of health care technology coverage problems with consistency.

References

1. Kristensen FB et al for the EUnetHTA Collaboration: practical tools and methods for health technology assessment in Europe. EUnetHTA. Int J Tech Ass Health Care 2009;25:Supplement 2 (2009), 1–8
2. Goetghebeur MM et al: Applying the EVIDEM framework to medicines appraisal. Med Decis Making. 2012 Mar-Apr;32(2):376-88.
3. Radaelli G et al: Implementation of EUnetHTA Core Model in Lombardia: the VTS framework. Int J Technol Assess Health Care. 2014 Jan 22:1-8.

Address for correspondence: Dr. Michele Tringali, Regional Health Care Office, via Pola 9/11, 20124 Milano, Lombardy - Italia.
Email: michele_tringali@regione.lombardia.it

Mobile Education: Reaching Homeless PTSD Veterans

Stephanie R. Tucker, MS[1], Sriram Iyengar, PhD[1], Amy Franklin, PhD[1]
[1]University of Texas Health Science Center, Houston, TX

Abstract

Smart phone technology has immense potential for changing health education delivery. Mobile education has the ability to reach individuals outside traditional channels. This poster describes the application of smartphone-based instruction for PTSD training in homeless veterans. Such technology provides multimedia education, improves access, and may minimize social anxiety for this population. This project contributes evidence of the ways in which mobile education may increase the use of HIT and access to healthcare for underserved populations.

Introduction

Mobile devices are increasing dramatically in use. In the United States, cell phone ownership by adults exceeds 90%[1]. Mobile technology has many advantages over paper and computers for delivering health education content. It can be used without the connectivity often mandatory in streaming web-based systems, offers portability enabling anywhere/anytime use, provides a greater degree of privacy with the small screen than is available with large books, manuals, or computer screens, supports emotional engagement[2] and social cohesion[3], and delivers a high level of interactivity.

Methods

Here, we describe the benefits of mobile educational for homeless veterans with PTSD highlighting the gains from a mhealth platform. Presentation of materials to homeless veterans is often in to face-to-face verbal only exchanges. In such circumstances, multi-methods of learning are not available and situation factors such as crowding and social anxiety may overwhelm participants. Mobile education is highly adaptable to learning theories, and has the potential to present similar content making use of multiple adaptive forms of presentation including text, video, visual aids, and games.

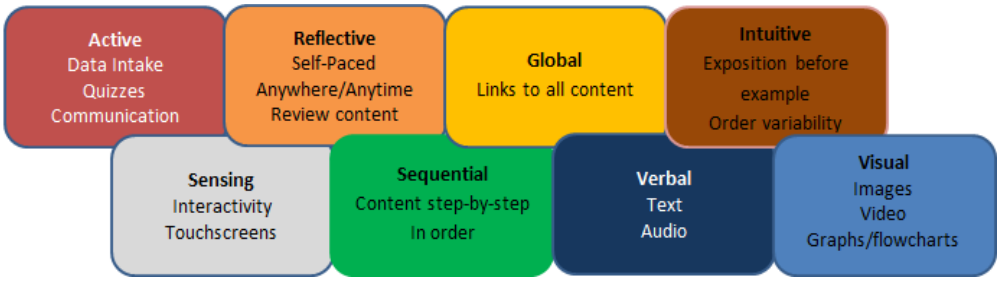


Figure 1 Model of mobile features adapted to Felder-Silverman Learning Style Theory[4]

Results

We believe that mobile education has the potential to reach populations currently unable or unwilling to access more traditional methods such as classrooms and office visits. In our example case study of homeless veterans, social, physical, mental, and even financial constraints may limit access to necessary care. We anticipate our intervention will achieve these aims and recognize potential drawbacks such as the difficulty of monitoring a homeless population, and charging mobile devices.

Conclusion

In an extension of telemedicine through mobile devices, we seek to explore ways in which education can be transformed. This study provides growing evidence of the ways in which mobile education may increase the use of HIT and access to healthcare for non-traditional populations.

References

1. Rainie, L. Cell phone ownership hits 91% of adults. Pew Research Center. 2013. Available at: http://www.pewresearch.org/fact-tank/2013/06/06/cell-phone-ownership-hits-91-of-adults. Accessed 3/7/ 2014.
2. Vincent, J. Emotional attachment to mobile phones: An extraordinary relationship. In Hamill, L., Lasen, A., & Diaper, D, eds. Mobile World. London, England: Springer; 2005: 93-104.
3. Rainie, L, Wellman, B. Networked: The new social operating system. Cambridge, MA: MIT Press; 2012.
4. Felder, R.M, Silverman, L.K. Learning and teaching styles in engineering education [Electronic Version]. Engineering Education. 1988. 78(7), pp 674-681. Available at http://www4.ncsu.edu./unity/lockers/users/f/felder/public/Papers/LS- 1988.pdf. Accessed 3/13/2014

Title: The State HIE Program Evaluation: A Typology of State HIE Approaches

Theme: Health Information Technology

Authors: Petry Ubri, Felicia LeClere, Sai Loganathan, Michael Latterner, Prashila Dullabh

Research Objective

The State Health Information Exchange (HIE) Cooperative Agreement Program ("the Program") was created by the American Recovery and Reinvestment Act of 2009 to expand the secure movement of electronic health information within the health care system. As part of a multi-year evaluation of the Program, we developed a method of organizing the strategies pursued by states to promote HIE services, a typology to help us determine 1) how states implemented the Program and 2) the methods by which states enabled exchange.

Study Design

Data was collected from all states and territories on various aspects of their program design to understand how these approaches vary by state and whether there are commonalities among state that accelerate the development of HIE. Contextual variables were assembled from secondary sources that characterize both the health care and economic climate of the state and the measures of HIE both before and after the initiation of the Program. The data analyses identify key program factors that may contribute to the viability of information exchange (e.g., legal and policy-related activities, governance structure, and technical aspects).

Principal Findings

This presentation will share descriptive statistics on trends in three areas: leadership and organization; technical approach; and legal/policy aspects of the program. Also discussed will be limitations in the typology and how these will be accounted for in the interpretation of key findings. Important trends include: Most states (70%) used a single technical entity for technical services, although many planned to build a network that might involve different vendors. Directed exchange is more commonly offered than query-based exchange (79% and 68%, respectively). Fifty-nine percent of states have enacted legislation supporting HIE or HIE and electronic health records, and many are bolstering its use with additional initiatives and financial incentives. Finally, a majority of states (68%) use an opt-out consent model.

Conclusions

The typology has allowed us to identify three primary categories and a core set of factors to best describe HIE implementation. While these factors are not necessarily associated with programmatic progress or lack therefore, there is a notable, clear trend towards certain implementation models and approaches.

Implications for Policy, Delivery or Practice

By assessing state progress and identifying national trends in HIE implementation strategies, we hope to contribute to the knowledge base guiding future HIE planning and implementation efforts.

Does size matter? A comparison of a large web corpus and a smaller focused corpus for medical term extraction.

Klaar Vanopstal, PhD[1], Els Lefever, PhD[1] and Véronique Hoste, PhD[1]
Ghent University, Belgium

Abstract

This paper investigates the cost-effectiveness a large corpus versus a smaller, more focused corpus for medical term extraction in English and Dutch. An error analysis will be presented, and we will study the extent to which these comparable – as opposed to parallel - corpora can be used in bilingual term extraction.

Introduction

Efficient acquisition and management of terminology has become increasingly challenging in the past few decades, as the explosion of information makes it difficult to keep term bases up-to-date. Although larger quantities of information are freely available, it is still difficult to find parallel texts for specialized domains, especially for underresourced languages. Automatic term extraction (ATE) is especially useful in the biomedical domain, which is highly productive when it comes to the creation of new terms. In this paper, we will investigate the cost-effectiveness of using a larger corpus versus a smaller, more dedicated corpus for ATE for English and Dutch, and their potential for use in bilingual term extraction from comparable texts in English and Dutch.

Description of corpora used

Large corpus of medical web data

The first corpus we used in this experiment was compiled using BootCat[1], a freely available web crawler which uses seed terms to automatically construct Google queries. We used the Wikipedia page for *heart failure* as a source for seed terms (n = 229). This resulted in a corpus of 2.5M words. The same procedure was followed for the Dutch web corpus, which contains 1.2M words.

Smaller, focused corpus of MEDLINE titles

The assumption here was that titles of scientific articles about a specific subject have a relatively high concentration of specialized terms. This smaller corpus was compiled using the PubMed interface to MEDLINE. We constructed a query for *heart failure* (*"Heart Failure"[MeSH Major Topic]) AND "heart failure"[Title/Abstract]*), which yielded about 45,000 results, from which we only retained the titles in our corpus. This resulted in a limited yet highly focused corpus of 613k words. The corresponding – smaller - Dutch corpus will be collected from the *Netherlands Journal of Medicine* and *Belgian Journal of Medicine.*

Gold standard

A gold standard was constructed by selecting the *heart failure*-related terms from MeSH. It consists of 172 concepts (and their Dutch translations as provided in the UMLS) and a total of 737 variant terms to express these concepts.

Results

Term extraction was performed on the larger and smaller corpora using TExSIS[2]. A total of 87,728 and 43,693 single and multiword terms were extracted from the English large web corpus, and the smaller corpus of titles, respectively. The web corpus contained 275 gold standard terms (37%), 234 of which were found by our term extraction system. The smaller corpus of MEDLINE titles contained 179 gold standard terms (24%), 138 of which were detected by TExSIS. In our poster, we will present a more detailed error analysis.

Conclusions

Although the difference in size between the two corpora was 75%, the difference in gold standard terms that were covered, was only 13%. This raises the assumption that titles (and by extension also abstracts) are an interesting and cost-effective source for terminology extraction. We will make the same comparison for a Dutch corpus of titles and a Dutch web corpus.

References

1. Baroni, M. and S. Bernardini. *BootCaT: Bootstrapping Corpora and Terms from the Web*. in *Proceedings of the 4th Language Resources and Evaluation Conference (LREC)*. 2004.
2. Macken, L., E. Lefever, and V. Hoste, *TExSIS: bilingual terminology extraction from parallel corpora using chunk-based alignment.* Terminology, 2013. **19**(1).

Disease/Disorder Semantic Template Filling – Information Extraction Challenge in the ShARe/CLEF eHealth Evaluation Lab 2014

Sumithra Velupillai, PhD[1], Danielle Mowery, MS[2], Lee Christensen, MS[3], Noemie Elhadad, PhD[4], Sameer Pradhan, PhD[5], Guergana Savova, PhD[5], Wendy W Chapman, PhD[3]
[1]Stockholm University, Sweden; [2]University of Pittsburgh, PA; [3]University of Utah, UT; [4]Columbia University, NY; [5]Childrens Hospital Boston and Harvard Medical School, MA

Abstract

We characterize data and describe the ShARe/CLEF eHealth 2014 challenge – Task 2.

Introduction

The meaning of disease/disorder (DD) mentions in the clinical narrative is often affected by semantic modifiers, such as negation. Such information is critical for accurate automated information extraction (IE). The ShARe/CLEF eHealth 2014 Task 2[1] is organized to advance the development of IE methods for detecting such DD semantic modifiers.

Methods and Challenge Corpus Characteristics

We released the extended ShARe corpus[2], a subset of clinical narratives from the MIMIC-II database[3], manually annotated by domain experts for DDs and 10 attribute/value semantic modifiers. Each DD is mapped to a SNOMED CT identifier from the UMLS (CUI), and each modifier has a normalized value:

Attributes Types	Example Sentence	Norm Slot Value	Cue Slot Value (span offset)
Negation Indicator (NI)	***Denies*** numbness	yes	*0-5*
Subject Class (SC)	***Son*** has schizophrenia.	family_member	*0-2*
Uncertainty Indicator (UI)	***Evaluation*** *of* MI.	yes	*0-9*
Course Class (CC)	The cough ***worsened*** over the next two weeks.	worsened	*11-18*
Severity Class (SV)	He noted a ***slight*** bleeding.	slight	*12-17*
Conditional Class (CO)	Return ***if*** fever.	true	*8-9*
Generic Class (GC)	pain ***while*** standing	true	*6-10*
Body Location (BL)	Patient has ***facial*** rash.	*C0015450: Face*	*13-18*
DocTime Class (DT)	Patient had tumor removed.	before	*--No cue annotated/no slot--*
Temporal Expression (TE)	The rash was present ***for 3 days***.	duration	*22-31*

In total, 433 clinical reports were released to task participants (300 training, 133 test). These are the richest semantic annotations of any previous clinical IE challenge. Results from participant submissions will strengthen the state of the art for deeper semantic IE from clinical reports.

Acknowledgments. This work is partially supported by CLEF Initiative, ShARe (R01GM090187), SHARP (90TR0002), Swedish Research Council, Swedish Fulbright Commission and THYME (R01LM010090). We thank the ShARe/CLEF eHealth 2014 chairs Lorraine Goeuriot, Liadh Kelly.

References

1. ShARe/CLEF eHealth 2014 website: http://clefehealth2014.dcu.ie/task-2
2. Elhadad N, Chapman WW, O'Gorman T, Palmer M, Savova G. The ShARe Schema for the Syntactic and Semantic Annotation of Clinical Texts. In preparation.
3. Lee J, Scott DJ, Villarroel M, Clifford GD, Saeed M, Mark RG. Open-access MIMIC-II database for intensive care research. Proc 33rd IEEE EMBS. 2011; 8315–8318.

Novel insights into disease interactions from Medicare claims data using Berg Interrogative Biology™ Informatics Suite

Vijetha Vemulapalli, Jiaqi Qu, Leonardo O. Rodrigues, Vivek K. Vishnudas, Rangaprasad Sarangarajan, Ely Benaim, Viatcheslav R. Akmaev, Niven R. Narain
Health Analytics Division, Berg, Boston, MA

There has been a quantum explosion in healthcare data availability since the drive to adopt electronic health records. This data can be leveraged to open up new avenues in advancing healthcare. Increased understanding of molecular and clinical interactions in patients that causes variation in outcomes is paramount to creation of precision models in medicine.

Big data efforts in healthcare have predominantly been focused on better management/curation of health care data and mining to test hypotheses. Though these types of big data efforts advance the field by improving access to data and speeding up research, they are limited by current knowledge and long-held biological or clinical phenotype hypotheses. Alternatively, knowledge discovery that is data-driven and unbiased by current knowledge will create a paradigm shift in healthcare by leading to discovery of new and often surprising trends in disease outcomes. Identification of non-obvious comorbidities, development of optimal treatment strategies and protocols will lead to practice-changing events that may mitigate side effects, thereby reducing healthcare costs.

Mathematical/statistical learning tools developed in Artificial Intelligence are well adapted to decipher complex interaction patterns in Big Data. The Berg Interrogative Biology™ Informatics Suite is a computational workflow for integration of varied data modalities and inference of causal effects in a data-driven manner using Artificial Intelligence (particularly Bayesian Networks (BN)). We propose that use of BNs in healthcare Big Data analytics has a significant impact on enhancing patient care and improving healthcare/hospital efficiency.

To demonstrate this idea here, we use low-resolution, publicly available data to make novel discoveries that directly inform care and lead to novel hypothesis. We built a data relationship network of DRG codes by applying Berg Informatics Suite to publically available billing data from CMS[1]. The data includes discharge counts for top 100 codes for 1600 providers in 2011. Data was prepared for AI model building using rigorous methods of filtering, normalization and missing data imputation. A cause-and-effect relationship network representing interactions between the numbers of discharges for DRG codes was built by using a BN learning algorithm.

Heart and renal failure sub-networks (**Error! Reference source not found.**) were chosen from the entire network for further exploration since these conditions rank high as leading causes of death. Most interactions identified in these sub-networks are recognized, hence serve as method validation. Identification of interaction between "bronchitis/asthma" and "renal failure" is new and unexpected based on current medical knowledge. Further exploration of this connection has resulted in a novel hypothesis where some treatment choices for bronchitis could have adverse effects on renal function. Testing of this hypothesis in a more granular dataset will potentially lead to findings that can directly be applied to improving patient care.

This analysis shows strong support for use of AI methods in hypothesis generation which on further exploration will result in clinically actionable information. Such network analysis strongly supports use of purely data-driven methodology on Big Data in healthcare to expedite medical research by providing unique insights unbiased by current knowledge.

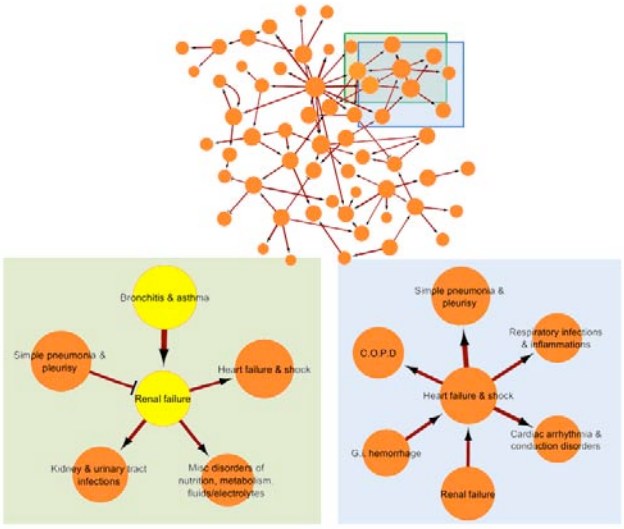


Figure 1: Heart and renal failure sub-network. Novel interaction is highlighted.

[1] https://data.cms.gov/Medicare/Inpatient-Prospective-Payment-System-IPPS-Provider/97k6-zzx3

Feasibility of Improving Health Literacy Using an Internet Medication Application Implemented in a Diabetes Clinic in India

Jonathan S. Wald, MD, MPH[1,3], Madhu Shrestha, BS[1], Christine Poulos, Ph.D[1], N.S.Vishwanath, MBBS, MPH[2]

[1]Research Triangle Institute, Research Triangle Park, NC; [2]St. John's Research Institute, Koramangala, Bengaluru, India; [3]Harvard Medical School, Boston, MA

Abstract

Limited patient health literacy and barriers to provider-patient communication are common throughout the world and reduce the effective use of medications. The objective of this pilot study was to determine the feasibility of offering tailored printouts for diabetes medicines (Meducation®) to patients at a diabetes clinic in Kerala, India. Preliminary results showed that patients and medical staff found the medicine handouts to be useful and identified implementation challenges to address in future work.

Introduction

Non-adherence to medication contributes to poorer health outcomes and increased utilization of health resources[1]. Meducation® (Polyglot, Inc., Morrisville, NC) is a software product that provides tailored medication information to patients of all literacy levels using an easy-to-read format, pictograms, large font, basic reading level and 16 language choices. This preliminary report focuses on user acceptance and implementation challenges.

Methods

Diabetes medication name, dosage, and frequency were translated into Malayalam for use in the application. Patients newly presenting to the Diabetes Care Centre at the Medical Trust Hospital (MTH) in Kerala, India were recruited as participants, and completed multiple structured surveys between October 2014 and February 2015. Sixty-five participants assigned to an intervention group receiving medicine handouts and answered survey questions covering self-reported medication adherence, diabetes knowledge, self-care activities, and satisfaction with the Meducation® handouts. Semi-structured interviews of MTH staff were conducted. Approval was obtained from Institutional Review Boards at RTI and at MTH to perform this study.

Results

Survey data 4-6 weeks after the initial patient visit showed that most participants (a) found the handouts to be helpful and referred to them for information on their medicines(34/40, 85%); (b) said they would not be likely to view, download, or print the medicine information themselves (24/35, 69%); and (c) reported they would most prefer obtaining the handout from their physician (28/32, 86%). Challenges noted by staff included: (1) recurring internet slowdowns and power outages prevented access to the website at times; (2) interruptions to work flow; (3) that Malayalam translations sometimes lost the original meaning of the English text.

Discussion

While these results are preliminary, Meducation® handouts were generally viewed positively by participants, but they were not preferred in an electronic format or via self-service. Staff utilized workarounds, such as using pre-saved PDF files or having colleagues with a faster connection generate and email back handouts. Staff recommendations for improving implementation focused on utilizing the pharmacist to distribute the handouts and considering an approach that supported a "sometimes connected, sometimes powered" workflow.

Conclusion

Applications such as Meducation have the potential to change the way patients receive medication information, but more work is needed to match component applications with suitable health IT platforms in under-resourced settings.

References

1. Haynes RB, Ackloo E, Sahota N, McDonald HP, Yao X. Interventions for enhancing medication adherence. Cochrane Database of Systematic Reviews 2008, Issue 2. Art. No.: CD000011. DOI: 10.1002/14651858.CD000011.pub3.

Effective Strategies For Encouraging Provider Utilization of Mental Health Measurement-Based Care Software (COMMEND)

Dan Y. Wang, PhD[1], Justin G. Chambers[1], Sara J. Landes[1], Eve B. Carlson[1], Josef I. Ruzek[1], Steven E. Lindley, MD[1, 2]
[1]VA Palo Alto Health Care System, Palo Alto, CA; [2]Stanford University, Stanford, CA

Abstract
Adoption of measurement-based care software by mental health providers is not automatic nor common practice. For our system which is known as COMMEND, some success came with a modified interface enabling data to be saved in less than a minute and managerial acceptance and regular review of collected data.

Introduction
An integral part of delivering the best mental health care to returning veterans within the VA Health Care System (VA) is measuring outcomes and tracking outcomes against the types of evidence-based therapies and therapeutic drugs employed. We developed a software tool known as COMMEND which facilitates the saving of such data and their graphical display versus time for providers including psychiatrists and therapists. COMMEND works in synchrony with the VA's Computerized Patient Record System (CPRS), automatically displaying when a provider logs into CPRS. COMMEND's patient panel page shows most recent and baseline outcomes, most recent and next appointments, etc. to aid provider planning. At a click of the patient's name, the provider can also bring up graphs showing measured outcomes and therapies used and drugs prescribed versus time for the last 3 years. However, adoption by providers has not been enthusiastic. We describe some major barriers and initial strategies for overcoming some of these barriers.

Methods
Starting from March 1, 2013, about 25 providers were serially trained to use COMMEND over about 9 months. After training, each provider's use was monitored: all major user interactions with COMMEND were logged into a database. We also held many teleconferences with individual providers to discuss their concerns, many of which were addressed by modified features.

Results
The largest barrier to adoption stems from the fact that it is not standard practice for providers to use graphical outcome trends to guide or modify patient treatment. While outcomes are important for managers and administrators evaluating the performance of a facility with respect to the therapies employed, at the individual patient level there is considerable skepticism about the added value. The second barrier stems from providers' reluctance to do data recording tasks beyond the required writing of the session note.
After many rounds of iterative redesign, we achieved some success in provider adoption. The number of COMMEND notes written per week markedly increased in Aug-Sept. 2013. There are two main causes: one, creation of a vastly simplified session note template called Quick Note allowing providers to record session characteristics in about 30 seconds per patient. Two, we started giving bimonthly summaries of COMMEND data collected to the providers' manager, who began using it to evaluate and review group performance, in particular regarding the use of evidence-based therapies.

Conclusion
To encourage mental health providers to adopt measurement-based care software, one effective strategy was to streamlining the COMMEND software further, reducing the time taken for data entry to no more than a minute per patient. Equally important was gaining managerial acceptance and regular review of collected data.

Building medical informatics data system for quality improvement in Children's hospital setting

Haijun Wang[1] Ph.D. MPH, Terri L Brown MSN, RN, CPN[1], Charles G. Macias MD, MPH[2]
1. Quality and Clinical Systems Integration, Texas Children's Hospital, 1102 Bates Street, Houston Texas 77030. 2. Center for Clinical Effectiveness and Emergency Medicine, Texas Children's Hospital, 1102 Bates Street, Houston Texas 77030.

Quality improvement (QI) is now a driving force in health care and is an essential aspect of service delivery at all level. Medical data not only enables us to accurately identify problems, it also assists to prioritize quality improvement initiatives and enable objective assessment of whether changes and improvement have indeed occurred. In order to better server the QI needs and improve the process of quality improvement, we built a data system which made our QI initiatives more effective.

In addition to clinical data, QI need many different kinds of data, so we first built an enterprise data warehouse (EDW) which integrated EPIC Clarity, PeopleSoft and patient satisfaction and many different types of data together. SQL server 2008 R2 database was used to build EDW and ETL package was run daily, so the data will be one day late. The subject data mart (SAM) is built based upon metrics defined care process (CP) team for each QI project. The QI projects were determined based on the key process analysis (KPA) results and our CP team is multidisciplinary team consisting of clinicians, nurses, quality specialists, data architects, clinical data specialists and BI developers.

The QlikView was used as dashboard development and data visualization tool which consumes data in the EDW. In addition to dynamically display the metrics, we also developed dynamic statistical process control charts (SPC) to monitor the quality improvement project's process. The automated SPC chart allows us to observe the effective ness of our interventions. This automated feature eliminated all of the manual process we used before.

Up to know we have formed 8 CP teams, built more than 10 SAMs and corresponding dashboards. By using this system, we have successfully reduced the chest x-ray rates for asthma patients by more than 35%, and tremendously increase mono antibiotics rate to over 92 percent. We also fixed many data issues existing in out medical record system so our data quality have greatly improved.

Improving the Review of Individual Patients in a Clinical Data Repository

Nich Wattanasin, MS[1], Michael Mendis[1], Kenneth D. Mandl, MD[3], Isaac S. Kohane, MD, PhD[3], Shawn N. Murphy, MD, PhD[1, 2]
[1]Partners HealthCare, Boston, MA; [2]Massachusetts General Hospital, Boston, MA; [3]Boston Children's Hospital, Boston, MA

Abstract

The recruitment of a sufficient number of patients is a crucial part of clinical trials research. We have developed components and leveraged community projects in i2b2, an open source platform for secondary-use of clinical data for research, to accelerate the process of identifying and reviewing suitable patients for a study.

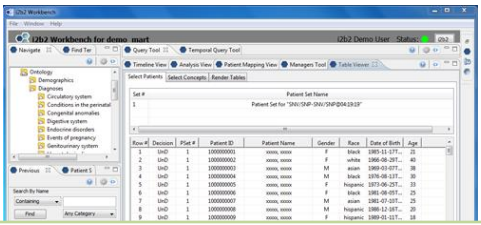
Background

Informatics for Integrating Biology and the Bedside (i2b2) is one of the sponsored initiatives of the NIH Roadmap National Center for Biomedical Computing. The primary goal of i2b2 is to provide clinical investigators with a cohesive set of software tools necessary to collect, host, and manage clinical data from the EMR, and enable the secondary-use of that data for research. The i2b2 platform is designed in a modular fashion consisting of interoperable "cells", or software modules that communicate through web services, which fosters the invention of additional plugins contributed by the i2b2 community that extends the functionality to a new "i2b2-CT" platform, which supports end-to-end identification of patients for clinical trials research. Today, within the i2b2 web client, an investigator can render views of patient centric data for review, as well as utilize newly developed tools, such as our clinical trial suite of apps, to make the review process more efficient for identifying patients qualifying for clinical trials.

Methods

We have developed a number of new components to enhance the i2b2 platform to support the patient recruitment process for clinical trials research. These elements include 1) a web service based ETL cell to retrieve the latest patient data from an institution's enterprise web services, 2) an identity management cell to manage encrypted patient lists and authorizations, and 3) a new workbench plugin to support a tabular display for selection of patients matching phenotypic criteria. In addition, we leverage community-driven projects such as the SMART-i2b2 platform (Substitutable Medical Apps Reusable Technologies) to further augment the determination of trial suitability workflow and allow for investigators to view PHI in a customizable patient-centric view. Utilizing the SMART application programming interface (API), we designed and developed a suite of clinical trials related SMART apps that run inside the i2b2 web client and allow one to manually specify eligibility criteria for a clinical trial or automatically import the criteria from ClinicalTrials.gov. The apps comb over the patient's data on a patient-by-patient basis, for example, looking for a certain medication or problem, or a text string in a note, and displays a matching score based on the defined criteria, allowing the investigator to flag the patient as a potential match.

Results



You can launch the **Patient Centric View** in a pop-up web browser for each patient directly from the i2b2 Workbench, or drag a row to save a patient of interest

We have implemented and deployed a limited release of the work described herein at Partners HealthCare for select i2b2 disease-based driving biology projects and as part of the Scalable Collaborative Infrastructure for a Learning Healthcare System (SCILHS) project, a Patient Centered Outcomes Research Institute (PCORI) grant at Harvard. After identifying a set of patients matching user-defined characteristics such as diagnoses, medications, procedures, etc., the set can be viewed in a tabular format and refined further. Next, a patient's record can be reviewed one-by-one in the SMART Patient Centric View, and the investigator given the opportunity to bookmark any patients of interest in the i2b2 Workplace.

Implementing an end-to-end patient identification workflow for clinical trials in a robust and extensible framework such as i2b2 allowed us to not only leverage existing community-driven projects, but also look forward to emerging initiatives based on this groundwork to support recruitment for multi-site clinical trials. This work was sponsored by Harvard CTSA, NIH U54LM00874, ONC 90TR0001/01, and PCORI.

Profiles Research Networking Software: An Open Source Community

Griffin M Weber, MD, PhD[1]
[1]Harvard Medical School, Boston, MA

Abstract: *Profiles Research Networking Software (RNS) is a social networking website for scientists used at more than 30 institutions across the country (http://profiles.catalyst.harvard.edu). As an open source platform, it has benefitted greatly from software contributions from other universities, partnerships with commercial companies, and feedback from users and developers. This poster reviews how these pieces have come together to create a successful open source platform, and it presents a timeline for future enhancements to the product.*

Introduction: Profiles RNS automatically generates online expertise profiles for investigators using a variety of data sources, including local administrative databases and public repositories such as publications from PubMed. Profiles are linked together using automated algorithms that discover investigators who are prior coauthors, who perform similar research, who are in the same organization, who are physically close, etc. Interactive data visualizations and social network analysis metrics help users understand and navigate through these networks.

Academic Partnerships: Although Profiles RNS was originally at Harvard in 2008, its freely available code and BSD open source license has enabled other universities to contribute to the product: Profiles RNS is a Semantic Web application that uses the VIVO ontology, which was developed by Cornell and University of Florida. VIVO provides a standardized way of describing the scholarly activities of researchers (http://vivo.org). University of California San Francisco (UCSF) incorporated the OpenSocial framework into Profiles RNS, which lets developers (e.g., Wake Forest and others) easily create modular social networking "gadgets" that plugin to the website (http://orng.info). Boston University added a feature to Profiles RNS that automatically synchs publications with ORCID, a non-profit organization that assigns unique identifiers to (http://orcid.org).

Commercial Partnerships: Recombinant Data Corp (now part of Deloitte) was the first Authorized Service Provides for Profiles RNS. In addition to providing commercial support for institutions using Profiles RNS (installation, customization, training, etc.), they played a significant role in managing the open source codebase by moving Profiles RNS to GitHub, reviewing and testing contributed code, writing documentation, and adding new features to the software. Symplectic (http://symplectic.co.uk) is picking up where Recombinant left off, as a new Authorize Service Provider for Profiles RNS. KNODE created ORNG gadgets that lists investigators' collaborators outside their own institutions (http://knodeind.com). Elsevier and Thomson Reuters, through Scopus (http://info.scival.com) and Web of Science (http://researchanalytics.thomsonreuters.com/incites), have provided commercial publication data for institutions' Profiles RNS websites.

Collaboration: Several mechanisms exist for institutions to learn about Profiles RNS or participate in its development. A Google Group (http://groups.google.com/group/profilesrns) provides a mailing list, which has more than 200 members, who frequently share tips, bug fixes, feature requests, etc. A monthly Users Group webinar hosted by Harvard discusses long-term plans for Profiles RNS and often has invited speakers. A separate Developer meeting hosted by UCSF focuses on technical issues. Conferences (such as AMIA) are an important venue to disseminate information Profiles RNS and collaborate with developers of other research networking tools.

Next Steps: Harvard's five year plan for Profiles RNS includes: automatic import of grants, patents, and medical licensures; recommendation tools that provides personalized search results; and bibliometric and social network analysis reports that enable individuals and institutions to review their scholarly portfolios. Though, many other new features will be added in parallel by the numerous academic and commercial developers.

Conclusion: Profiles RNS was built to foster collaboration. It therefore seems fitting that the software itself is result of many contributors working together. Close relationships with commercial partners has been essential, and this should be considered by other institutions selecting among open source license types. Frequent communication among geographically separated development teams, for example, though regularly scheduled conference calls, is important for coordinating efforts and maintaining enthusiasm and momentum. An open model for Profiles RNS ultimately helps the end users, who benefit from the ideas and efforts of not one but many development teams.

References: Kahlon M, Yuan L, Daigre J, et al. The use and significance of a research networking system. J Med Internet Res. 2014 Feb 7;16(2):e46.

Funding: NIH Awards 8UL1TR000170 and 1UL1TR001102, Harvard University and its affiliated health centers.

A Keyword Suggestion Strategy Based on Citation Networks

Wei Wei, MS, Shuang Wang, PhD, Xiaoqian Jiang, PhD, Lucila Ohno-Machado, MD, PhD
University of California, San Diego, La Jolla, CA

Motivation

In order to write review articles, it is necessary to select effective keywords to conduct the queries, and this is a very time-consuming task. In traditional literature searches, keywords are manually prepared based on experts' knowledge. However, these manually selected keywords might miss some other essential ones. We propose a strategy to suggest related keywords using machine-learning techniques.

Strategy

We infer new related keywords from pre-selected keywords. Our machine learning algorithm is based on a corpus of relevant literature and its corresponding citation network. We are preparing a corpus of biomedical informatics literature using the article collecting system we developed. We have collected over one thousand articles in the biomedical informatics domain and are developing tools to build their citation networks. We will use ParsCit[1] to extract reference information from articles and calculate Levenshtein distances to match related articles and construct the citation network. We will then collect keywords to retrieve the top *k* cited articles from PubMed. Each target article will be mapped to the pre-built citation network. Then, based on the topology among the citation relations, we will generate a local graphical model (GM), where each article node contains a list of MeSH term weights. We will then build a global GM by combining all local GMs. A non-leaf node aggregates the weights of MeSH terms from child nodes; it also propagates the aggregated weights to the parent nodes, where message passing methods can be used to efficiently learn the weights of MeSH terms at the root node K (see Figure 1)[2].

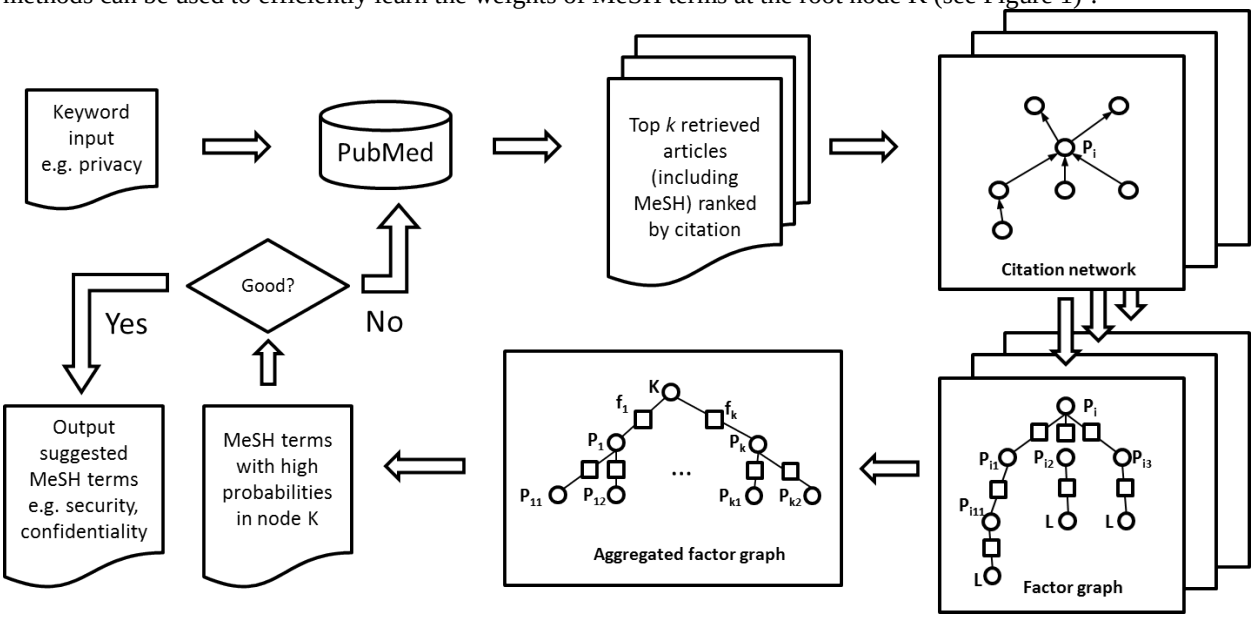


Figure 1. Procedures of the keyword suggestion system. A circle P represents an article; a square represents a factor; a circle L denotes a prior. Each node P maintains a list of weights for the entire MeSH term set. If a term appears in the article, its prior might be assigned a large weight; otherwise, it might be assigned a small but non-zero weight. A factor node contains a pre-defined voting rule and other information such as the article's metadata. A factor node collects partial weights of a term from its child article nodes and then passes the weights to its parent node. For example, P_1 to P_k are *k* articles; f_1 to f_k calculate weights w_1, …, w_k for the term "privacy" from their child nodes. The final weight of "privacy" at node K is proportional to $\prod_{i\in[1,k]} w_i$. The suggested MeSH terms are calculated by collecting MeSH terms associated with high probabilities in node K.

Acknowledgement: This work is supported in part by NIH grants U54HL108460.

References

1. Luong M-T, Nguyen TD, Kan M-Y. Logical structure recovery in scholarly articles with rich document features. *International Journal of Digital Library Systems (IJDLS).* 2010;1(4):1-23.
2. Yedidia JS, Freeman WT, Weiss Y. Understanding belief propagation and its generalizations. *Exploring artificial intelligence in the new millennium.* 2003;8:236-239.

Developing an online social network platform to collect family health history

Brandon M. Welch, MS, PhD[1], Robert Lario, MSE, MBA[2], Joshua D. Schiffman, MD[2]
[1]Medical University of South Carolina, Charleston, SC; [2]University of Utah, Salt Lake City, UT

Abstract

Family health history (FHx) is one of the most important and consistent risk factors for disease. Unfortunately, FHx is often poorly collected in the clinic. To improve the ability of individuals to collect and record their own FHx, we're developing an online social platform that helps family members connect and share health information with each other. We believe that a social network will help individuals collect a more comprehensive and accurate FHx from their family.

Introduction

Family health history (FHx) is one of the most important tools available for genetic diagnosis and disease risk assessment,[1] and is the most consistent risk factor for almost all human diseases. Half of all families display a positive FHx for one or more common chronic diseases.[2] Unfortunately, FHx is often inadequately collected in a clinical setting as clinicians are typically hindered by limited time, inadequate reimbursement, and lack of expertise related to genetics and FHx.[1] As a result, FHx is discussed with only half of new patients and less than a quarter of return patients.[3] As such, a comprehensive FHx data collection and documentation may be best accomplished outside the timeframe of a physician office visit.[1] As a result, several researchers and organizations, such as the Surgeon General, have built patient-oriented tools to help individuals collect and record their own FHx. However, these tools tend to be difficult to use, disease specific, collect limited structured data, have limited interoperability, and lack useful clinical decision support. Furthermore, these FHx tools require the patient to recall all their relatives' health information, which results in inaccuracies and data gaps, particularly in 2nd and 3rd degree relatives.[4] In general, patient-oriented FHx collection tools have had limited adoption and impact outside of research settings.

Our approach

To overcome the challenges of current patient-oriented FHx collection tools, clinicians and informaticists at the University of Utah and the Medical University of South Carolina are currently developing an online social network platform called ItRunsInMyFamily.com. This website will allow individuals to record their own health information and link up with relatives to exchange health information to create a more comprehensive and accurate FHx. ItRunsInMyFamily.com will leverage both popular social networking techniques and health information technology approaches to help individuals and their relatives collect and share health information with each other. At its core, ItRunsInMyFamily.com is being developed as a Meaningful Use-aligned, standards-based personal health record (PHR). The platform will use social network techniques to generate a family pedigree of available health information. ItRunsInMyFamily.com will be a free resource that also provides basic care recommendations (e.g. talk to your doctor about your risk for disease X) to individuals. Clinicians would also be able to access their patients' FHx information and receive personalized risk assessments and care guidelines for their patients. In the future, we are anticipating the incorporation of genomic information into the platform. In this poster, we will describe our architecture approach, standards used, the user interface, and approaches to data privacy & security being incorporated in the development of ItRunsInMyFamily.com.

Conclusion

We believe that an online platform using social networking technology which helps individuals share and discuss their health information with relatives will create a more comprehensive and accurate FHx than current approaches.

References

1. Rich ECEC, Burke W, Heaton CJCJ, et al. Reconsidering the family history in primary care. *J Gen Intern Med*. 2004;19(3):273–280.
2. Scheuner MT, Wang S, Raffel LJ, Larabell SK, Rotter JI. Family History : A Comprehensive Genetic Risk Assessment Method for the Chronic Conditions of Adulthood. *Am J Med Genet*. 1997;324(February):315–324.
3. Acheson LS, Wiesner GL, Zyzanski SJ, Goodwin MA, Stange KC. Family history-taking in community family practice: implications for genetic screening. *Genet Med*. 2000;2(3):180.
4. Murff H. Accuracy of family history information for risk assessment in clinical care. In: *Family History and Improving Health, NIH State-of-the-Science Conference*.; 2009.

BCMA use in the psychiatric population: Challenges in the inpatient setting

Kandace M. Whiting, MSN, RN1, Linda Harrington, PhD, DNP, RN2, Pat Matos, DNP, Constance Johnson, PhD, RN1
1Duke University, Durham, NC; 2Texas Christian University, TX; 3UCLA Resnick Neuropsychiatric Hospital, Los Angeles

Abstract
Consistent use of BCMA safeguards improves medication safety. Psychiatric illness symptomology presents additional challenges to BCMA use. BCMA use in the psychiatric setting was evaluated to identify factors that negatively impact BCMA compliance. Although adherence to BCMA protocols remains integral to medication safety, patient specific considerations need to be considered when developing and implementing BCMA protocols.

Introduction

Barcode Medication Administration (BCMA) effectiveness is contingent upon consistent use of embedded safeguards and compliance with best-practice protocols[1]. Nonetheless, noncompliance is well documented in the literature[1-2]. The psychiatric population presents additional challenges to adhering to BCMA best-practice recommendations that are not present in other patient populations. The purpose of this project is to evaluate BCMA use in the psychiatric setting to identify BCMA workarounds and address issues associated with BCMA policies, organizational resources, workflows, and end-user training that negatively impact BCMA compliance.

Methods

We evaluated BCMA use in a 24-bed in-patient psychiatric unit by comparing unit policies to best-practice using the Harrington BCMA Checklist[2] and observing medication administration episodes (MAE). All steps of each observed MAE were documented and we created visual diagrams depicting MAE workflows. These visual diagrams were compared to the BCMA Checklist[2] and deviations were deemed workarounds. Workarounds were categorized and analyzed based on Koppel et al (2008) workaround classifications and corresponding causes. Discrepancies between current practice and best-practice dictated policy revisions, workflow redesigns, and end-user education programs.

Results

Twenty-four workaround types were identified in the observed MAEs. Of the 24 identified workaround types, six were related to patient identification challenges associated with psychiatric symptoms. For example, patients who were severely disorganized, delusional, or paranoid were frequently observed not wearing wristbands and/or refusing to be scanned. Thus, patient wristband barcodes were either scanned while not on the patient or not scanned at all. Other identified challenges included use of workstations on wheels when caring for aggressive patients and burdensome workflows related to the mobile nature of psychiatric patients.

Conclusions

Behavioral compliance with BCMA best-practice is impacted by factors related to tasks, technology, patients, organizational policies, and the environment[1]. Psychiatric symptoms including paranoia, disorganized thought patterns, aggression, and delusions present additional challenges that impact best-practice BCMA use. Although adherence to BCMA protocols remains integral to medication safety, patient specific considerations need to be considered when developing BCMA protocols. Additional research is needed to determine how to account for these challenges to assure medication safety in the psychiatric population.

References

1. Koppel R, Wetterneck T, Telles JL, Karsh BT. Workarounds to barcode medication administration systems: their occurrences, causes, and threats to patient safety. J Am Med Inform Assoc. 2008 Jul-Aug;15(4):408-23.
2. Harrington L, Clyne K, Fuchs MA, Hardison V, Johnson C. Evaluation of the use of bar-code medication administration in nursing practice using an evidence-based checklist. J Nurs Adm. 2013 Nov 43(11):611-7.

Embedding a medical search engine within an electronic health record

Jayne A. Williams, MA[1], Robin L. Kruse, PhD[2], Patricia E. Alafaireet, PhD[3], Jeffery L. Belden, MD[2], Karl M. Kochendorfer, MD, FAAFP[1,4]

[1]MedSocket, Columbia, MO

[2]University of Missouri Department of Family and Community Medicine, Columbia, MO

[3]University of Missouri Department of Health Management and Informatics, Columbia, MO

[4]University of Illinois Hospital and Health Sciences System, Department of Family Medicine, Chicago, IL

Abstract

This study aims to investigate the impact on and user satisfaction with embedding an information retrieval tool into electronic health record (EHR) system used by family physicians and residents at an academic medical center. The tool, 1-Search, provides a single search box for retrieving information from a variety of content sources and types. A pre and post implementation survey were used to gather users' awareness and feedback about their experience with 1-Search. Initial results suggest that physicians support incorporation of the search engine in the EHR.

Introduction

Physicians' information needs are often complex and require access to a variety of information sources, such as guidelines, primary research, patient education materials, and their clinic's internal algorithms.[1] Physicians working at the point of care need quick access to high-quality, specialized information from varied sources to treat patients as effectively as possible. 1-Search is a federated search engine built using Microsoft SharePoint and FAST Search Server. It retrieves results from evidence-based clinical resources as well as an organization's intranet. In 2013, to help meet this need, 1-Search was embedded in the electronic health record system at a large Mid-Western academic health center and made available to members of the department of Family and Community Medicine. We surveyed physicians to determine likelihood and frequency of use, physician acceptance, and usability. Specifically, this study funded by the National Institutes of Health (1R43LM0115909-01) sought to answer the following questions:

1. What type of information sources do physicians use at the point of care?
2. What is the impact and user satisfaction of a medical search engine embedded in an electronic health record?

Methodology

The initial product was developed from a multi-source knowledge base composed of literature and previous student findings as well as substantive input by practicing physicians and health science library professionals. Post development and prior to full deployment, a short training was provided to all potential users. Surveys were administered before (N=97) and after (N=98) implementation to measure usefulness and impact of 1-Search. The pre-survey determined the extent of physicians' information needs and their knowledge about 1-Search. The post-survey addressed user satisfaction and perceived impact of the tool on clinical decision making. Search time, number of clicks, and selected links were recorded by the system.

Results

Response rates to the pre- and post-surveys were 70% and 59%, respectively. Forty-one people responded to both surveys. The proportion of 1-Search users did not differ between pre- and post-surveys (42% and 53%, p=.23). Among 1-Search users on the post-survey, respondents reported that they used 1-Search primarily to address a question about a specific patient (25 of 29), look up treatment information (17 of 29), or to confirm thinking (14 of 29). Of 56 respondents who completed the question, 26 reported being more likely to use 1-Search now that it is embedded in the EHR.

Conclusion

This study provided the necessary evidence to support the continued use of 1-Search as well as its further development. Tools, such as this type of embedded information retrieval tool, hold promise to improve care though reduced physician cognitive workload, through more compressed reference information delivery and through timely access to information within a typical clinical workflow.

References

1. Hersh W. Ubiquitous but unfinished: grand challenges for information retrieval. Health Info Libr J [Internet]. 2008 Dec [cited 2013 May 8];25 Suppl 1:90–3. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19090855

Title: Important information to communicate between clinicians and families in the intensive care unit

Michael E. Wilson MD, Sumanjit Kaur MD, Alice Gallo De Moraes MD, Brian W. Pickering MD, Ognjen Gajic MD, Vitaly Herasevich MD PhD
Multidisciplinary Epidemiology and Translational Research in Intensive Care (METRIC), Mayo Clinic, Rochester, MN

Introduction: Family members of patients in the intensive care unit (ICU) play an integral role in medical decision making and supporting patients. Communication with families is often a difficult and complex task for ICU clinicians. Few studies have attempted to systematically identify information important for clinicians to know about family members in the ICU. Our objectives were to:

1) Identify important information for ICU clinicians to know about family members
2) Identify important information for ICU clinicians to communicate with family members

Methods: A literature search and semi-structured interviews identified 21 items perceived to be important for clinicians to know about family members and 32 items perceived to be important for families to know about patients from clinicians. Family members and clinicians of two ICUs from a single institution were asked to rate the identified information as necessary to ICU decision making using a seven point Likert scale. (7=necessary information, 4=neutral, 1=unnecessary information).

Results: Fifty-four family members and 50 ICU clinicians completed the survey (response rate 64%). Subjects rated each identified piece of information as necessary (mean ratings 5.1-6.6 for family members, 4.7-6.8 for clinicians), although significant variability existed regarding necessary and unnecessary information. Themes important for clinicians to know about family members include: family background, questions, understanding, goals, concerns, wellbeing, and requests for additional help. Themes important for families to know about the patient's medical care include: diagnosis, treatments, prognosis, clinical status, schedule, comfort, goals of care, medical team, and family participation. Family members rated 25 of the 53 pieces of information as more necessary than did clinicians (p values ranging from <.0001 to .05).

Table. Family member trust of information from the following sources

	Family members (n=54) mean ± SD
Doctor in the ICU	6.5±1.00
Nurse in the ICU	6.3±0.79
Doctor outside of the ICU (such as primary care provider)	5.4±1.29
Family or friends	4.7±1.50
Brochures, pamphlets, etc.	3.7±1.48
Internet	3.7±1.60
Religious leader or spiritual advisor	3.4±1.25
Newspapers, magazines, etc.	2.9±1.44

Discussion and Conclusions: We identified information important for clinicians to know about family members and information important for family members to know about the patient from clinicians in the ICU. There is variability between information that clinicians and family members considered as important. This variability may contribute to miscommunication, decreased family and clinician satisfaction, and poor quality decision making. Our findings and survey results can be utilized to design tools to improve information exchange between family members and clinicians in the ICU. Successful communication between family members and clinicians in the ICU is a key to improve mutual understanding and shared decision making.

Evaluation of Real-Time Use of the EHR in Documenting Pain Scores

Tamara J. Winden, MBA, Jennifer M. Krueger, MS, Karl Fernstrom, MPH, Kai G. Hanson, MS, Michelle Ophaug, BSN. Allina Health, Minneapolis, MN

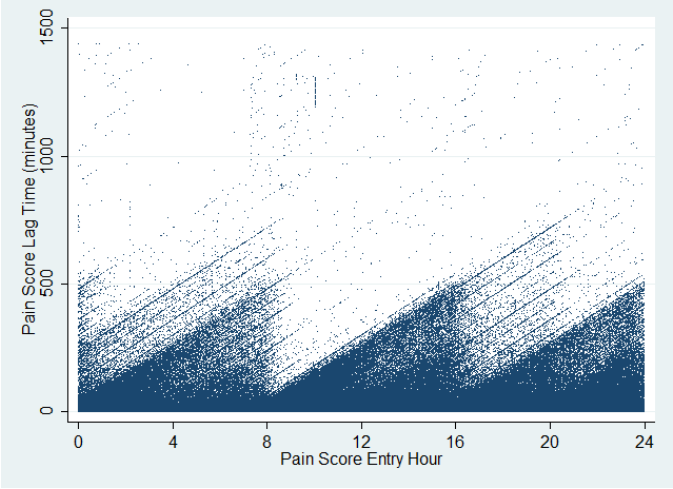
Abstract: ***Real-time use of Electronic Health Record (EHR) systems by clinicians is a vital component of the current and future healthcare provision in the U.S. Delay in documentation into the EHR could influence patient satisfaction, safety, and healthcare decisions. The goal of this mixed methods study is to examine the lag time between observation and data entry of pain scores into the EHR and identify barriers to real-time documentation. This project will present an analysis of current practice in documentation of pain scores, identify barriers to real-time EHR data entry, and evaluate impact to patient care.***

Introduction: One challenge with EHRs is making sure documentation is as current as possible. This requires that EHR tools are appropriately incorporated into the clinician's workflow. To increase accessibility and reduce documentation lag times, many institutions have invested significantly in improving EHR efficiencies. They have also invested in additional computer hardware for patient rooms, hallways, as well as moveable carts to increase EHR availability to staff. However, documentation appears to still lag behind observation times and this latency has the potential of impacting patient care and safety. This, combined with current federal initiatives around the control of pain in the inpatient setting, was the impetus for this project. The objectives of this study are to examine the latency in documentation of pain scores in the EHR, understand the current documentation practices and barriers to real-time documentation of pain scores, and examine the impact to patient care. This study was approved by the institutional review board.

Materials and Methods: This project is a mixed methods study of real-time documentation practices of nursing and other clinical staff in 10 inpatient hospital settings. The primary focus is on data entry of pain scores into the EHR. Retrospective secondary data from the EHR will be examined along with a cross-sectional survey of clinical staff that enters pain scores into the EHR and observations on hospital units in an effort to describe barriers to real-time documentation.

Results: Preliminary analysis of the secondary data shows that most of the variation in the pain score lag time between hospitals may be due to hospital size. Larger hospitals have an increased median pain score lag time and increased interquartile range. There seems to be little variation in lag time between shifts as a whole. Though when we looked at common transition times there was both an increase in the median lag time and its variability. This may be a symptom of clinical staff waiting to complete documentation until the end of their shift (Figure 1). Finally, there does seem to be a negative association between pain score lag time and the level of pain intensity. As pain intensity increased both the median lag time and its variability decreased significantly.

Figure 1. Pain score documentation over a 24 hour period



Creating Clinically Homogeneous Groups of Prostate Cancer Patients

Janusz Wojtusiak, Che Ngufor, Lorens Helmchen, Jack Hadley
George Mason University, Fairfax, VA

Introduction

Comparative effectiveness analyses that use observational data require methods, such as propensity score matching and instrumental variables, that "simulate" randomization to avoid bias associated with the effects of unobserved factors that influence both the treatment selected and the clinical outcome. Despite such adjustments, the results of these methods apply essentially to an "average" patient. Personalized medicine requires finding the treatments most appropriate for a specific patient, not on the average. A promising approach to address this goal is creation of homogeneous groups of clinically similar patients and performing comparative effectiveness analysis within the clusters. Several methodological and practical issues need to be addressed before such groups are created.

The presented project focuses on using machine learning methods to create homogeneous groups of prostate cancer patients using a combination of classification learning and clustering to create groups that are clinically similar. Construction of the groups uses information typically available to urologists making treatment decisions: prostate cancer characteristics (PSA, Gleason score, stage) and predicted survival based on conditions existing prior to prostate cancer diagnosis.

Method

The presented method works in two stages. First, it creates models of the competing mortality risk based on pre-cancer clinical conditions. We compared parametric and machine learning-based classification models applied to predict short-term mortality (1 year) and long-term (7 years) mortality. In order to select right models several methods were tested including logistic regression, support vector machines, decision trees, and random forests. The models are trained using non-prostate cancer patients to specifically address only the competing mortality risk.

Second, competing mortality risk is combined with prostate cancer risk factors and unsupervised learning algorithms are applied to create homogeneous groups. The experiments compared two popular unsupervised learning methods: k-means and hierarchical clustering. For comparison we created a set of 6 clusters based on an arbitrary high/low split of predicted survival and the 3 level D'amico prostate cancer risk score.

Evaluating clustering methods requires checking clustering consistency when multiple clusterings are applied to discover optimal number of clusters that consistently capture "patterns" in the patient population. In order to do so, each clustering was bootstrapped 100 times, and entire process cross-validated 5 times. The Jaccard similarity measure and cluster prediction strength [1] were used to find the optimal method and number of clusters. Homogeneity of groups also needs to be evaluated using measures independent from those used to create clusters. We assumed that similar patients who receive the same treatments should have similar outcomes. (This assumption ignores variance in treatment quality, but is sufficient for the presented work). Thus, at the last step, clusters are validated by measuring area under ROC curve (AUC) for logistic regression model that predicts true mortality based on cluster membership and treatment. The assumption is that the higher the AUC, the more similar are patients within a group.

Results

Data used in this study were from the SEER-Medicare database, complemented with claims for 5% random sample of Medicare beneficiaries. Main inclusion criteria for the study were: for non-cancer patients: continuous enrollment in Medicare Part A or Part B for at least a year prior to January 1 2004 and were at least 66 years old for the non-cancer patients; and for prostate cancer patients: at least 66 years old at age of diagnosis. A total of 119,028 non-cancer and 104,213 cancer patients were identified for potential inclusion in the study.

Based on 100 bootstrapped samples of cancer patients using predicted mortality and prostate cancer-specific characteristics the k-means and hierarchical clustering both indicated the data are best clustered into 5, 6 or 9 clusters (Jaccard index = 0.9, 0.8, and 0.85 respectively). The independent validation of clusterings indicated that the AUC for mortality model was the highest for 6 clusters obtained with the k-means algorithm. The results are: AUC(5 clusters)=0.83; AUC(6 clusters)=0.84; and AUC(9 clusters)=0.84. These numbers were significantly higher ($p=0.035$) than the AUC for a simple tabulation of predicted survival and D'amico scores into 6 clusters (0.80) and 9 clusters (0.82).

References

[1] Tibshirani, R. and Walther, G. (2005) Cluster Validation by Prediction Strength, *Journal of Computational and Graphical Statistics*, 14 511-528

SimProtocols – A Software Prototype for In Silico Comparison and Evaluation of Computer-based IV Insulin Infusion Protocols

Anthony F. Wong, MTech[1], Senthil K. Nachimuthu, MD, PhD[1], Peter J. Haug[1,2]
[1]University of Utah, UT, [2]Intermountain Healthcare, UT

Abstract

We are developing a prototype for in silico comparison and evaluation of computer-based IV insulin infusion protocols. It has three major components: patient module, interface module and comparator module. We used clinical data from ICU patients who were supported by eProtocol-insulin to simulate insulin-blood-glucose interaction in the patient module. The interface module manages the interaction between the computer-based insulin protocol and the simulator. The comparator module measures the performance of competing protocols by evaluating its outcome.

Introduction

Computer-based clinical protocols can standardize clinical decisions while adapting to contextual changes and individualizing patient care. Different computer-based protocols can generate different results and lead to different patient outcomes. An efficient method for comparing and assessing different computer-based clinical protocols, before committing them in clinical trials, would be valuable. We used IV insulin infusion protocols as an exemplar.

Methods

The prototype is based on a framework we developed for in silico comparison and evaluation of computer-based clinical protocols. The main function of the patient module is to simulate specific physiological process using data derived from real patients. We used clinical data from ICU patients who were supported by eProtocol-insulin[1], a heuristic rule-based protocol used for managing blood glucose. Internally, eProtocol-insulin assumes a linear rate of change in glucose in response to changes in IV insulin. Thus, the resulting clinical data is unique because it reflects the individual patient response (measured through the blood glucose) as a result of a controlled insulin treatment regime. We used these responses to simulate the next appropriate level of blood glucose once the patient module received the insulin dose recommendation from the competing computer-based IV insulin protocol.

The interface module manages the interaction between the working modules and various competing computer-based IV insulin protocols. It sends patient-specific data required by the computer-based IV insulin protocols in order to generate the necessary insulin dose recommendation. It then channels this result to the comparator module for further evaluation.

The comparator module compares the recommended insulin doses by competing computer-based protocols at different blood glucose ranges (low, on target, high)[2]. A favorability score is used to measure how well the competing computer-based protocol perform. We measure the performance using a set of rules we defined based on the recommended insulin dose. At low blood glucose (<80 mg/dL) and on target range (80-110mg/dL), we prefer lower insulin dose. At high blood glucose (>110 mg/dL), we prefer higher insulin dose.

Results & Discussion

Our clinical data included 408 patients with 18,984 eProtocol-insulin recommendations. We used this dataset to simulate new levels of blood glucose when regulated by insulin in the competing protocols. The projected change in blood glucose was calculated based on the aggregated change in blood glucose level for every given insulin dose in that specific patient. We used linear regression to predict the outcome of the new blood glucose value for every plausible recommended insulin dose. At this stage, we are developing a competing protocol to complete our simulation. We believe our proposed novel simulation framework using real patient data is a viable method of comparing different computer-based protocols.

References

1. Morris AH, Orme J, Truwit JD, Steingrub J, Grissom C, Lee KH, et al. A replicable method for blood glucose control in critically Ill patients. Crit Care Med. 2008 Jun;36(6):1787–95.
2. Wong A, Pielmeier U, Haug PJ, Morris AH. Evaluation and Comparison of Two Computerized IV Insulin-Treatment Protocols Using Patient Data from the ICU. AMIA Annual Symposium. 2013.

Engaging Patients with Advanced Directives Using an Experiential Information Visualization Approach

Janet Woollen, RN, MS[1], Suzanne Bakken, RN, PhD[1]
[1]Columbia University, New York, NY

Abstract

Despite the substantial benefits of advanced directives to both patients and care providers, they are often not completed due to lack of patient awareness. We propose designing an Experiential Information Visualization (1) to promote advanced directive awareness through vivid illustration of options, and (2) to inspire contemplation and conversation regarding patients' end-of-life journeys.

Background and Significance

An advanced directive (AD) is beneficial to patients and care providers in guiding end-of-life care; however, ADs are often not completed. A primary reason for this is lack of patient awareness[1]. Healthcare providers operate under great pressure in a system that does not provide incentives or training to discuss end-of-life care with patients and their families. It is unrealistic to expect that system to change any time soon, so it is left to patients and their families to navigate many of the decisions regarding end-of-life care. Patients want to be active participants in the end-of-life process; however, patients and families lack access to information on what to expect, what to look for, and what their options and rights are. They know the principles but they need ready access to pertinent information, presented in a way that facilitates understanding of the specifics.

Uninformed individuals confront numerous challenges when facing end-of-life trials: significant decisions being made under stressful conditions; unintended financial burden on patients, families and society; moral distress and conflict for healthcare providers; and perhaps most importantly, medical care that may be carried out contrary to an individual's wishes and values. Clinicians typically turn to family members for decision making when patients cannot advocate for themselves, but studies reveal that neither families nor clinicians accurately predict what patients want[2]. A lack of awareness and understanding of ADs must be addressed to remedy such problems and to set the stage for desired outcomes.

Approach

The purpose of this project is to introduce an experiential information visualization that would furnish immersive information to patients and families on end-of-life directives.

We propose designing an innovative tool to (1) promote AD awareness by engaging patients to learn about their options and (2) inspire contemplation and conversation regarding patients' end-of-life journey. An Experiential Information Visualization (EIV) may be able to communicate insights that are often communicated in words, but are much more powerfully communicated by example. An EIV could facilitate vivid understanding of options and inspire the beginning of often-difficult conversations between care providers, patients and loved ones. It may also save clinicians' time, as care providers may be able to spend less time explaining details of end-of-life care options.

Results

Study in progress.

Conclusion

It is important to ensure that patients have time for reflection and are informed of their options to die with dignity. Our goal is to investigate if an EIV can increase awareness and understanding of an AD and promote contemplation of the end-of-life journey.

References

1. Rao JK, Anderson LA, Lin F-C, Laux JP. Completion of advance directives among U.S. consumers. Am J Prev Med. 2014 Jan;46(1):65–70.
2. Shalowitz DI, Garrett-Mayer E, Wendler D. The accuracy of surrogate decision makers: a systematic review. Arch Intern Med. 2006 Mar 13;166(5):493–7.

Analysis of Patient Portal Use in Underserved Populations and Settings

Maria D. Wright, RN, BSN[1], Tammy Toscos, PhD[1], Ayten Turkcan, PhD[2], Brad N. Doebbeling, MD, MSc[3]
[1]Indiana University Purdue University, Fort Wayne, IN [2]Northeastern University, Boston, MA [3] Indiana University Purdue University, Indianapolis, IN

As health information technology evolves to meet the needs of healthcare providers and health systems seeking financial incentives through the HITECH Act, the needs of the underserved patient population should be considered. Underserved patients, including those who are uninsured or underinsured, comprise a patient population that has many challenges with access to care. A patient portal can improve access to health care by enhancing patient-provider communication, empowering patients, and supporting care between visits. Addressing obstacles that interfere with the uptake and use of patient portals is not only important for improving the patient experience, but can impact multiple quality dimensions including access, efficiency, satisfaction, and health outcomes.

Community Health Centers (CHCs) that provide care to the underserved are working to implement electronic health records and attest to meaningful use through multiple approaches, including the implementation of patient portals. The goal of our Patient-Centered Outcomes Research Institute (PCORI) funded research effort, *Improving Healthcare Systems for Access to Care and Efficiency by Underserved Patients*, is to redesign CHCs to deliver timely care to underserved patients with common health problems. As part of this project we conducted a review of literature that examines patient portal use in underserved patient populations. We are also examining patient perspectives on barriers, access to portals, and desired functions in a sample of seven CHCs from across Indiana.

Our review of literature explored peer-reviewed journal articles published from 2009 to 2014 with a focus on access to technology, perceptions and attitudes toward patient portals, and barriers to enrollment in underserved patient populations. This population has been found to have access to different forms of technology, search for health information online, and have an interest in using healthcare information supplied digitally [1-3]. Positive perceptions about the use of patient portals include themes of enhanced understanding, empowerment, a strengthened patient-provider relationship, and convenience [3-4]. Negative themes include privacy and security concerns, limited health literacy, usability, and fear of a change in the relationship with the healthcare provider [3-4]. Because disparities have been found in the enrollment and use of a patient portal, it is important to examine the root causes as they pertain to this patient population so that potential barriers can be addressed before implementation [4-7].

Our work with CHC patients, staff, and providers has uncovered several strategies to address the negative patient perceptions found in the literature review. Gradually adding functionality to a portal may allow privacy and security concerns to be addressed. Getting patients accustomed to using the portal for lower risk information first, e.g. scheduling and appointment reminders, allows for opportunity to provide education on the safeguards in place. There must be a concentrated effort to educate patients on how they can benefit from using the portal, e.g. for routine management of a chronic disease, and its use must be encouraged by the providers and staff to mitigate any concerns about changed relationships. In order to accomplish this, we suggest including providers and staff in the discussion of what functions the portal should possess. This may act to improve provider and staff acceptance of the technology and in turn stimulate use by the patients.

References

1. Sanders M, Winters P, Fortuna R, Mendoza M, Berliant M, Clark L, Fiscella K. Internet access and patient portal readiness among patients in a group of inner-city safety-net practices. J Ambul Care Manage 2013;36(3):251-259. doi:10.1097/JAC.0b013e31829702f9.
2. Schickedanz A, Huang D, Lopez A, Cheung E, Lyles C, Bodenheimer T, Sarkar U. Access, interest, and attitudes toward electronic communication for health care among patients in the medical safety net. J Gen Intern Med. 2013;28(7):914-920.
3. Dhanireddy S, Walker J, Reisch L, Oster N, Delbanco T, Elmore J. The urban underserved: attitudes towards gaining full access to electronic medical records. Health Expect 2012 Jun. doi:10.1111/j.1369-7625.2012.00799.x
4. Sarkar U, Karter A, Liu J, Adler N, Nguyen R, Lopez A, Schillinger D. The literacy divide: health literacy and the use of an internet-based patient portal in an integrated health system-Results from the diabetes study of northern California (DISTANCE). J Health Commun 2010;15S2:183-196. doi:10.1080/10810730.2010.499988
5. Goel M, Brown T, Williams A, Hasnain-Wynia R, Thompson J, Baker D. Disparities in enrollment and use of an electronic patient portal. J Gen Intern Med 2011;26(10): 1112-1116. doi:10.1007/s11606-011-1728-3.
6. López L, Green AR, Tan-McGrory A, King R, Betancourt JR. Bridging the digital divide in health care: the role of health information technology in addressing racial and ethnic disparities. Jt Comm J Qual Patient Saf. 2011 Oct;37(10):437-45.
7. Sarkar U, Karter A, Liu J, Adler N, Nguyen R, López A, Schillinger D. Social disparities in internet patient portal use in diabetes: evidence that the digital divide extends beyond access. J Am Med Inform Assoc. 2011;18(3):318-321. doi:10.1136/jamia.2010.006015.

Clustering Analysis of the Gene Chip Data for Two Types of Leukemia

Cai Wu, MS
The University of Texas MD Anderson cancer Center, Houston, TX

Abstract

This project distinguished leukemia subtypes acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) with clustering methods. The pre-processing, normalization and log-transform were involved in data preparation. Hierarchical and K-means algorithms were used for clustering analysis through Matlab built-in and modified functions. The two types of leukemia samples were clustered into different groups in hierarchical algorithm and K-means algorithms; the pre-preparation made datasets suitable for analysis.

Introduction

Leukemia is a type of cancer that affects developing blood cells in a patient's bone marrow. There are as many as 150 subtypes postulated to exist. So we need to find an effective way to distinguish between two important leukemia subtypes: acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). Array technologies have made it straightforward to monitor simultaneously the expression pattern of thousands of genes. The challenge now is to interpret such massive data sets. Genes (rows) can be clustered to identify groups of co-regulated genes and expression patterns; samples (columns) can be clustered to identify new classes of biological samples and detect experimental artifacts. This project uses pre-processing, normalization and log-transform for data preparation and clustering analysis to distinguish leukemia subtypes from the dataset.

Expression Data Preparation

The dataset was downloaded from http://www-genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi, which included 59 control genes, 7070 test genes. There were 38 patient samples, 27 were ALL and 11 were AML. In such large data set, not every gene acted on the same level with the samples. Thus, data preparation could be the effective way to treat the data before clustering analysis.

First all negative values were set to 1; then the mean values of control samples were calculated to normalize the dataset; if the sample value less than the mean of same sample values (mean of the column), the small expression value (< mean of the column) was converted to the mean; if all same gene expression values (one row) were less than the mean, the row was eliminated; then the dataset was Log-transformed for clustering analysis.

After the pre-processing, the gene numbers (row of the data set) became 1250. Since the first 27 samples were ALL type and the rest 11 were AML types, the sample correlation coefficients with the expression values were checked. Meanwhile, the 50 selected gene datasets from neighborhood analysis were also being calculated. Based on the correlation image, the 50 selected gene expression values were significantly correlated with sample types. The first 27 ALL type samples were all positively correlated with ALL samples and negatively correlated with the rest 11 samples (AML type). The same trend was also shown in AML samples, although the correlations between the sample types were not strong as the 50 selected samples.

Clustering Analysis

Hierarchical and K-mean algorithms were used for clustering analysis. In hierarchical clustering, the similarity or dissimilarity between every pair of objects in pre-processed dataset were found first using pdist function; then the objects were grouped into a binary hierarchical cluster tree using linkage function; finally the hierarchical tree was divided into clusters using cluster function and using dendrogram to plot the cluster tree. In K-mean clustering, K centers in the data set were randomly selected; sum of square of distance between each data point and nearest center was minimized by using the Euclidean distance to calculate the distance, and the distance between clusters was maximized, the mean of each cluster was calculated; each record to one of the K means was reassigned. The threshold as 0.9 to get the clustering group for hierarchical algorithm, and 3 as cluster numbers for k-means algorithm.

Both Matlab built-in function and modified K-mean function clustered two types of leukemia samples into two different groups using Matlab, the hierarchical algorithm gives more cluster groups than k-means; the pre-processing extracted and enhanced meaningful data characteristics and normalization accounted for systematic differences across datasets; log-transformed dataset was suitable for clustering analysis.

Extract and Analyze Useful Information from a Noised DNA Chip Image

Cai Wu[a,b], MS, Yiqing Chen[a], MS

[a]Department of Computer Science, University of Houston, Houston, TX; [b]The University of Texas MD Anderson Cancer Center, Houston, TX

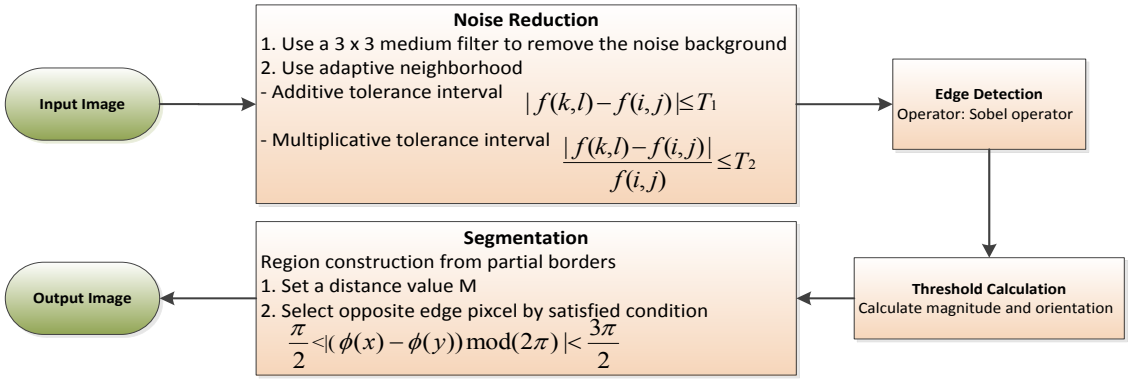
Abstract

This project analyzed a DNA chip image which contained DNA reaction spots and strong noise background, identified and extracted useful information, made a clear output image for further analysis. Median filter, Sobel operator and Hysteresis were used for edge detection, and region forming by checking opposite pixels was used to extract useful information. The processing reduced noise background and prevented edge blurring. The distance for region forming removed most of the noise background and kept the useful object information.

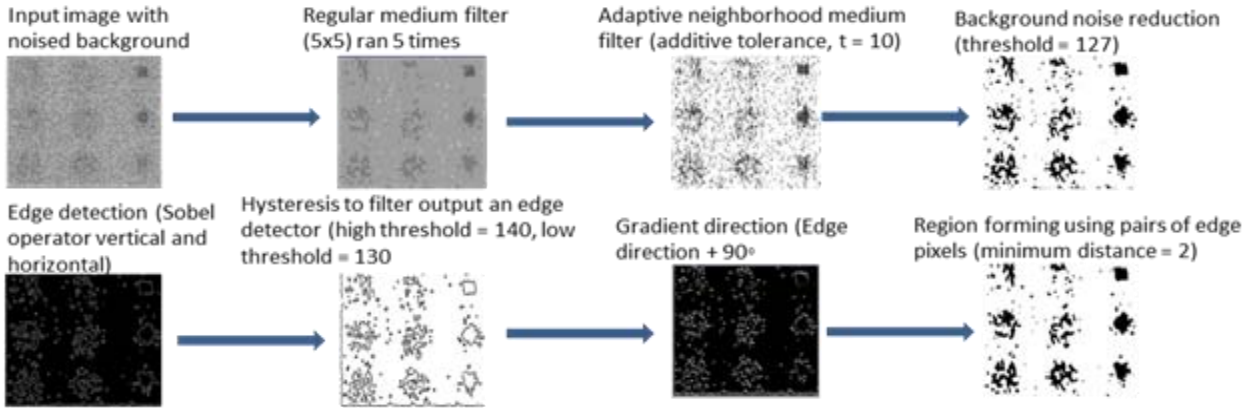
Introduction

Microarray is an important tool for gene expression analysis. However the noises introduced during the experiment affect the accuracy of the gene expression. Thus eliminating the noise effect is an important step for analysis. Image analysis includes grinding, spot recognition of the scanned image and removal noising background. In this project, median filter, Sobel operator and Hysteresis are used for edge detection, and region forming by checking opposite pixels is used to extract useful information in the microarray image processing procedure.

Input Image Processing



Results



Conclusion

Median filter, Sobel operator and Hysteresis reduced noise background and prevented edge blurring. The distance setting for region forming kept the useful object information.

Extending tranSMART for Meta-Analysis of Genomic Data Across Trials

Haiguo Wu, Ph.D[1], Elisabeth L. Scheufele, MD, MS[1,2], Dina Aronzon, MS[1], Matvey B. Palchuk, MD, MS[1,2]
[1]ConvergeHEALTH by Deloitte, Newton MA; [2]Harvard Medical School, Boston MA

Abstract

tranSMART is an open source knowledge management and high content data analytics platform. Data analyses were limited to a single clinical trial at a time. We constructed a data model and updated the application to load genomic data and run cohort analysis across multiple trials within tranSMART.

Introduction

tranSMART supports cohort discovery and hypothesis generation by combining clinical study data with genomic data. Initially, tranSMART was developed to run analysis within a single clinical trial. However, due to significant interest in analyzing the data from across multiple trials, we extended tranSMART to run meta-analysis across trials using clinical data. We are continuing to extend the meta-analysis functionality by including genomic data (e.g., gene expression data) and linking them to clinical data to enable more powerful analysis.

Methods

The data in tranSMART are stored within each trial and are mapped to separate ontology trees. For genomic data, there is the additional challenge of the data being normalized based on samples within a single study. The required extension is three fold: 1) support the genomic data node in custom across-trials ontology; 2) recalculate z-score; and 3) extend application layer for advanced data analysis.

The z-scores for each probe of each sample need to be recalculated and refreshed when a new study with genomic data is loaded. We extended the existing data model by adding a separate column to store the new z-score data for across-trial functionality. The ETL process was modified to handle these requirements, and application code base will be adapted to support the new data structure.

To test the new functionality, we used four parallel breast cancer studies, GSE32072, GSE22093, GSE23988 and GSE5462 (see Figure 1) with the raw gene expression data downloaded from the GEO repository (www.ncbi.nlm.nih.gov/geo/). These studies overlapped on the experiments performed - pre-treatment breast tissue tested with Affymetrix Human Genome U133A Array. We normalized the expression data with the RMA algorithm[1] for log values and calculated z-scores[2] for each study and across trials.

Implementation Example

The four studies had 21, 58, 61 and 103 patients in the genomic data node, respectively, which totalled to 243 patients. There are 22,283 probes in this array, which means that 5,414,769 records were updated. The z-score recalculation used about 3 minutes on a single-threaded server. This overhead only affects the initial loading but not the actual analysis. The benefit to run genomic data across trials surely outweighs the loading time.

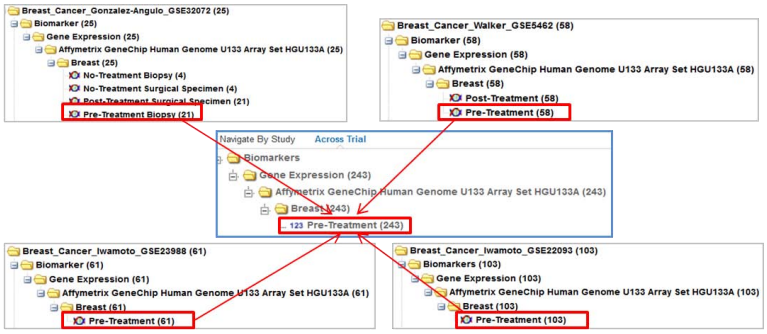


Figure 1: Mapping Genomic Data from Individual Trials to Meta-Analysis in tranSMART

Conclusion

We are actively working to extend tranSMART to perform meta-analysis of genomic data across trials, with good initial results.

References

1. Irizarry RA, et al. Comparison of Affymetrix GeneChip expression measures. Bioinformatics 2006;22: 789-794.
2. Cheadle C, et al. Analysis of microarray data using Z score transformation. J Mol Diagn. 2003;5: 73-81.

Challenges in Quantifying Narcotic Use from Drug Dispensing Records

Jianmin Wu, PhD, MS[1], Jon Duke, MD, MS[1,2], John T. Finnell, MD, MS[1,2]
Regenstrief Institute, Indianapolis, IN [1], Indiana University, Indianapolis, IN[2]

Introduction The epidemic of narcotic prescription drug abuse and its associated high cost and mortality have attracted national attention. Much research has been done to explore the distribution and determinants of the narcotic abuse. Narcotics prescriptions are usually ordered on a PRN basis (i.e., "as needed") with guidelines regarding the appropriate frequency of administration. Most researchers look at narcotic daily dose and days supply to determine a patient's risk for drug abuse. Currently, no standardized method exists to calculate these metrics from observational data. Inconsistency in narcotic measurement methods might lead to discrepancies in daily dose calculation and affect the stratification of patient risk for drug abuse. In this study, we will describe some of the challenges in quantifying narcotic prescriptions into morphine equivalents, the standard unit used for assessing patient risk and predicting mortality.

Methods Narcotic prescription data were collected from the Emergency Department of Wishard Health Services, from 2008 to 2010. Data elements used for calculating daily dose and supply days are: order_name, sig (free text describing administration instructions), quantity, script_dose (the amount of drug taken each time), frequency (time interval between doses), daily_dose, rx_dose (strength of the prescription per unit dispensed), and dose_unit (defining the dose in milligram or milliliters). The completeness of each data element was summarized as a proportion of the non-missing counts divided by the total prescription orders. "Maximum" rule was applied to dose calculation. For rx_dose PRN, e.g. '12.5-25 mL PO', the maximum volume '25ml' was selected as the dose taken by patient each time; for frequency PRN, e.g. 'PO Q4-6 hours', the time period between two doses was standardized as 4 hours; for any specified maximum dose, e.g. '1 tab PO Q4-6H prn pain (Do not take any more than 4 tablets per day)'; the maximum amount '4 tablets per day' was selected instead of the dose (6 tablets) calculated from structured data elements. Total morphine equivalents (Meq) for a single prescription was calculated using equation = quantity × rx_dose × dose_unit × conversion factor for Meq.. Supply days for each prescription were equal to the quantity divided by daily_dose or by the product of script_dose multiplied by frequency/24hours. Average daily morphine equivalent dose (MED) were determined by total Meq for a single prescription divided by supply days.

Results A total of 46140 patients generated 74266 narcotics prescription orders. The raw completeness of each data elements are: 100% (order_name), ~100% (sig), 100% (quantity), 1.5% (rx_dose), 85.9% (frequency), 22.3% (daily_dose), 22.3% script_dose, and 22.4% (dose_unit), respectively. Overall, about 20% of the prescriptions had sufficient data to calculate the average daily MED using the structured data elements. Around 80% of prescriptions required information from the free text 'sig' in order to calculate the daily MED. Only 0.1% of the prescriptions were could not be converted to daily MED due to missing or discrepant data. Some examples of these problematic prescriptions are listed in the table below.

Table. Examples of prescription orders with inadequate data to calculate morphine equivalent dose

Order_Name	Sig	Rx_dose	Quantity	Frequency	Daily_dose	Script_dose	Dose_unit
HYDROcodone & Actmnphn Elixir (7.5mg&500mg/15mL)|31383|M	5 mL PO Q6H PRN pain		***1 bottle***	Q6H	20	5	ML
HYDROcodone 5/Acetaminophen 500|12983|M	5 mg hydrocodone, ***1 tab PO q6h prn pain (Do not take any more than 2 per day)***		#15	Q6H			
HYDROcodone 5/Acetaminophen 325|36966|M	5 mg hydrocodone,***1 tab PO q6h prn pain*** (***Do not take any more than 1 per day per day***)		#10	Q6H			
OXYcodone|21543|M	***not more than 5 tab/day***	5mg	#30		5	5	TAB/D
Hydrocodone/Ibuprofen|32803|M	Take one tablet PO every 6 hours ***no more than 5 tablets per day***		#20	Q6H			
HYDROcodone 5/Acetaminophen 500|12983|M	***max 8 tabs/day***		#20		8	8	TAB/D
Tramadol|22482|M	***Take for headaches if necessary***		#10				
Tramadol|22482|M	25 mg PO QAM for 3 days duration then 25 mg PO BID for 3 days duration then 25 mg PO TID for 3 days duration then 25 mg PO QID for 3 days duration then 50 mg PO Q6H PRN pain (not more than 400 mg per day	50 MG	#20	***QAM|BID|TID|QID|Q6H***	25|50|75|100|200	25|25|25|25|50	MG|MG|MG|MG|MG
HYDROcodone 5/Acetaminophen 500|12983|M	5 mg hydrocodone, ***1 tab PO once daily before sleep q6h***		#10	QDAY,Q6H	9	5,1	MG,TAB

Conclusion A standardized method for measuring the daily dose of narcotics PRN prescriptions is important for secondary use of narcotic prescription data. Such standards will help improve the reliability of research identification from narcotics research.

Implementation of a Computer-Based Documentation System Improves Workflow Efficiency: A Case Report

Tzu-Yu D. Wu, MS[1], David J. Bradley[2], MD, Kai Zheng, PhD[1, 3]
[1]School of Information; [2]Division of Pediatric Cardiology; [3]School of Public Health, University of Michigan, Ann Arbor, MI

Abstract

Computer-based documentation (CBD) systems should promote workflow efficiency in addition to facilitating secondary use of data. In this case study, we examined the impact of implementing a CBD system designed to manage the performance, interpretation, and reporting of cardiac monitors at a referral academic medical center, and found a 45% reduction in the interpretation turn-around time by physicians. The observed improvements may be attributable to a more usable user interface and a higher level of accessibility of the system, which is being verified using qualitative methods.

Introduction

Computer-based documentation (CBD) systems have existed for over a decade. While their main design objective is to improve the quality and computability of clinical data captured in order to support both primary (i.e., patient care) and secondary use (e.g., quality assurance and research) purposes, they should ideally be designed to facilitate workflow to improve efficiency and to reduce user adoption barriers. In this abstract, we report a case where the introduction of new CBD system led to significantly improved workflow efficiency.

Method

Data were collected by a clinical team performing 24-hour ambulatory cardiac (Holter) monitoring on pediatric patients in a large teaching hospital. The team, consisting of three attending physicians, five technicians, and several nurses, has been using a legacy CBD system to record the progress of Holter cases for more than ten years. A Holter patient case has four milestone dates: the date when the monitor is placed on the patient, when the device is returned, when data are processed, and when data are reviewed and verified. In this study, we used the intervals between these milestone dates as surrogate measures of workflow efficiency. A shorter duration indicates better efficiency and more timely interventions. A new, web-based system, which provides a significantly streamlined user interface and a higher level of accessibility, was deployed in the hospital in October 2012. The analysis reported in this abstract compared three time periods: (1) The year beginning 24 months prior to deployment of the new application; (2) The 12 months prior to deployment; and (3) the 12 months following its deployment. The differences of time intervals were examined by ANOVA and TukeyHSD test with a 95% confidence interval.

Results

Table 1 shows the interval from preliminary Holter report creation to report verification for three time periods: Average Holter verification was reduced from 17.8 to 9.14 days ($p<0.001$) with implementation of the application. In contrast there was no significant difference between the two 12-month periods prior to application deployment ($p=0.45$). Figure 1 visualizes the results.

Table 1. Number of studies included

Stage	Period	# Studies	Avg. Dur. of Verification
D−24	Oct 10 ~ Sep 11	953	17.86
D−12	Oct 11 ~ Sep 12	1,176	16.64
D+12	Jan ~ Dec 13	1,315	9.14

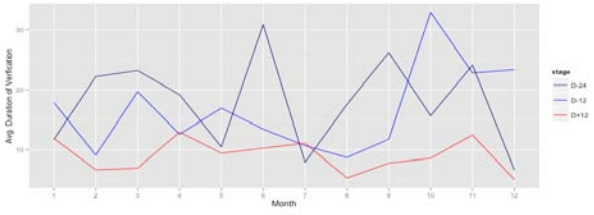


Figure 1: Avg. duration of verification per month

Conclusion

A properly designed CBD system improves workflow efficiency. This improvement has both clinical and financial implications. In this case, a shorter turnaround means that clinical reports can be returned to the ordering physicians more quickly to inform further decision-making processes. Faster return may also shorten the cycle of insurance reimbursement. Additional data are being analyzed and a follow-up qualitative study is ongoing to identify the specific aspects of the new CBD design that contribute to the observed improvements.

Development of a Unified Computable Problem-Medication Knowledge base

Yonghui Wu, Ph.D.[1], Adam Wright, Ph.D.[2], Hua Xu, Ph.D.[1], Allison B. McCoy, Ph.D.[3], Dean F. Sittig, Ph.D.[1]

[1]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA

[2]Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

[3]Department of Biostatistics and Bioinformatics, Tulane University School of Public Health and Tropical Medicine, New Orleans, LA, USA

Introduction: Medications and clinical problems are critical components of Electronic Health Records (EHR). A comprehensive knowledge base containing frequently used medications and their possible indications derived from clinical practice is very useful for clinical care and research. We developed a unified computable problem-medication knowledge base by integrating five knowledge bases derived from two EHR data sets and one patient claim data set.

Methods: We developed five problem-medication knowledge bases from two large scale EHR implementations and one large patient claim database, including the Brigham and Women's Hospital (BWH) EHR[1], the University of Texas Health Science center at Houston (UTH) EHR, a crowd sourcing (UTH_CS) knowledge base from the UTH EHR[2], a refined knowledge base from the UTH_CS with high reputation scores (UTH_REP) and a knowledge base developed from four years of patient claim data of Blue Cross Blue Shield of Texas (BCBSTX). We then developed an ensemble method to integrate the five knowledge bases into a unified computable problem-medication knowledge base. The medications from different resources were mapped to RxNorm using NLP and normalized to the ingredient level. Problems were normalized to root ICD codes. The confidence scores from different resources were integrated into a matrix to facilitate problem-medication knowledge extraction using various user-defined criteria.

Results: The unified knowledge base contains 128,928 problem-medication pairs among 2,118 normalized medications and 2,186 normalized problems. For the problem-medication pairs, the various confidence scores, including the chi-square score, the p-value, the Interest score, the crowdsourcing ratio score and the reputation score were incorporated into a matrix. Table 1 shows the number of problems and medications before and after use of the ensemble method for all the knowledge bases. To the best of our knowledge, this is the first time wherein a collective problem-medication knowledge base has been generated from heterogeneous data sets.

Table 1 Number of problems/medications/pairs before/after ensemble

	#Medications	**# Medications after ensemble**	**#Problems**	**#Problems after ensemble**	**# Pair**
BWH	1144	762	196	188	7593
UTH	2368	1394	563	501	16715
UTH_CS	2537	874	1575	1105	5719
UTH_REP	572	300	368	314	731
BCBSTX	7285	2577	1738	1202	106106

Conclusion: We developed a unified computable knowledge base using the ensemble method. The unified knowledge base can facilitate problem-medication knowledge extraction using different user defined criteria, potentially benefiting clinical care and research.

Acknowledgement: This study was supported in part by grant No. 10510592 for Patient-Centered Cognitive Support under the Strategic Health IT Advanced Research Projects Program (SHARP) and grant from the NLM R01LM010681.

References:

1. Wright A, Chen ES, Maloney FL. An automated technique for identifying associations between medications, laboratory results and problems. J Biomed Inform. 2010 Dec; 43(6):891-901.

2. McCoy AB, Wright A, Laxmisan A, Ottosen MJ, McCoy JA, Butten D, et al. Development and evaluation of a crowdsourcing methodology for knowledge base con-struction: identifying relationships between clinical prob-lems and medications. J Am Med Inform Assoc. 2012 Sep 1;19(5):713–8.

Investigating the Genetic Architecture of Pulmonary Arterial Hypertension Shared with Other Diseases.

Luke A. Yancy Jr[1,2], Atul J. Butte, MD, PhD[2]
[1]Biomedical Informatics Training Program, Stanford University School of Medicine, Stanford, CA; [2]Sytemems Medicine, Stanford University School of Medicine, Stanford, CA

Abstract

Pulmonary arterial hypertension (PAH) is a rare and fatal disease with a median survival of approximately 4 years for those afflicted. Many genome-wide association studies (GWASs) have attempted to determine the variants predictive of PAH, but often due to small sample sizes, studies have not led to highly predictive markers. However, GWASs have been successful in determining variants associated with many other diseases. We have an ongoing effort to curate these variants into a database called VARiants Informing MEDicine (VARIMED). VARIMED contains results from over 17,000 peer-reviewed genetic epidemiology studies (e.g. GWAS), covering over 460,000 variants associated with over 6,500 phenotypes. In this study, we leverage VARIMED to determine if PAH exhibits a shared genetic mechanism with other diseases. Finding other diseases that have a significant enrichment of PAH associated variants may provide insight into (novel) variants that play a significant role in PAH (and are otherwise not detectable in a PAH GWAS due to small sample size) or even novel therapies for PAH that can be "borrowed" from diseases with similar genetic architecture. Here, we perform a GWAS on exomes from PAH subjects, prioritize those variants, and use the top ranked variants to identify diseases in VARIMED with a similar genetic architecture. We found similarities between PAH and an unusual acquired kidney disease. This work can be extended to investigate the shared genetic architecture of other diseases (rare or otherwise understudied) by taking advantage of the large spectrum of diseases now with characterized genetic architecture.

What Tasks Are Clinical Teams Tracking in Daily Practice?
Usage of a Web-based Care Management Tool in Eight Veterans Affair Medical Centers

Jianji Yang, PhD, Judy McConnachie, MPH, Jonathan Sun, MS, Lisa Winterbottom, MD, MPH
Portland VA Medical Center, Portland, Oregon, USA

Abstract

The Care Management Tool is web-based application developed to support care coordination and task management in both Primary Care and Specialty Care teams within Veterans Health Administration. An analysis of the usage pattern indicates that it is used most frequently for prompting: a) telephone contacts with patients, b) scheduling future appointments, and c) lab test and consult follow-up. The analysis provides insight in the use of tasks for clinical reminders and can inform future development and deployment efforts.

Introduction

The Care Management Tool (CMT) was implemented in 8 Veterans Affair (VA) Medical Centers in mid-2012, initially in Primary Care (PC) and later introduced to Specialty Care (SC) services. CMT is a shared clinical task management platform and provides functionalities not available in most electronic health record systems such as task creation and tracking, role-based task cross-coverage and flexible reminders of tasks due (e.g. email reminders). Tasks are created to support care sourced in the medical record. For example, a PC provider orders a follow-up imaging study to be done in 6 months then schedules a task due in 6 months to confirm that the study was competed timely. When creating a task, the user can select from predefined lists of 'Task Type' (e.g. record review), as well as 'Task Reason' (e.g. education). Additionally, users can enter more details in the 'To Do Notes' field (Figure 1). In this work, the task data was analyzed to study the usage pattern by different clinical services, and to understand the types of clinical and administrative activities that are supported by the CMT . This analysis offers insight into users' needs and helps to inform future functionality development.

Methods

Task data from June, 2012 to Feb. 2014 was extracted from the database and analyzed in the following areas. First, descriptive statistics were calculated on number of tasks created by service, number of users by service, and the number of tasks per user by service. Second, the frequency of 'Task Type' and 'Task Reason' used were analyzed. Third, text in 'To Do Notes' was tokenized into single words. English stop words were removed and remaining terms were stemmed. The frequency of root terms was compared with 'Task Type' and 'Task Reason' results.

Results

Total number of tasks is 29,245, of which, PC, with the highest number of users (246), contributed the most at 14,482. Consult Management Service had 3 users, but created the most tasks/user (538). The most used Task Type and Task Reason were 'Telephone Contact' and 'Coordination of Care' respectively. In the 'To Do Note', the most used terms were: 'call', 'check', 'lab', 'bp', 'appt', 'consult', and 'schedule'.

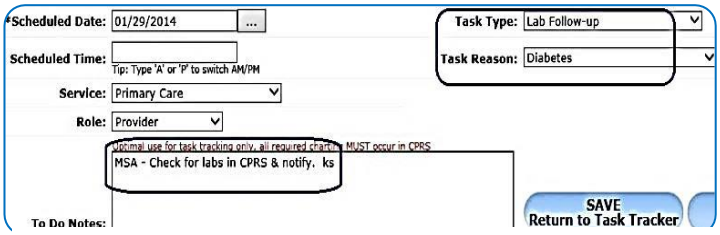


Figure 1. Partial screenshot of CMT showing task edit screen. Circled areas are where users enter the task attributes.

Discussion

As the service that provides the medical home to most VA patients, PC has the highest volume for CMT use. Heavy per staff usage appears to be in services that perform case management duties, e.g. Consult Management, Pain Management and Cancer Care Coordination teams, indicating that intense tracking of patient activities is a norm for these services. Additional workflow analysis may suggest future developments to better support multi-step or longitudinal care tracking in these services. The text analysis confirms the frequent use of "Telephone contact" and "Lab follow up", but also suggests scheduling appointments, checking on blood pressure and consult tracking are highly performed tasks that rely on a reminder system for completion.

Conclusion

Analysis of the two-year usage data for a web-based care management tool in eight VA facilities showed that the CMT supports clinical care by offering a common platform that is highly utilized to manage tasks for coordination of care. While PC has the highest volume of use, SC services also rely on the functionalities to support their work. The analysis of the task details provides insight into clinical staffs' use of the tool and informs future developments.

References

1. Yang J, McConnachie J, Sun J, Schreiner S, Winterbottom L. Modifying an Effective Eletronic Care Management Tool to Support Patient-Centered Team-Based Care Model in Portland VA Medical Center. *AMIA Annual Symposium Nov. 2012, Chicago, IL, USA.*

Design of Vendor-neutral Platform for Fast Prototype Model Verification and Deployment

Shiming Yang[1] PhD, Peter Hu[1] PhD, Yulei Wang[1] PhDc, Amechi Anazado[1] MD, Catriona Miller[2] PhD, Raymond Fang[2] MD, Stacy Shackelford[2] MD, Colin Mackenzie[1] MD

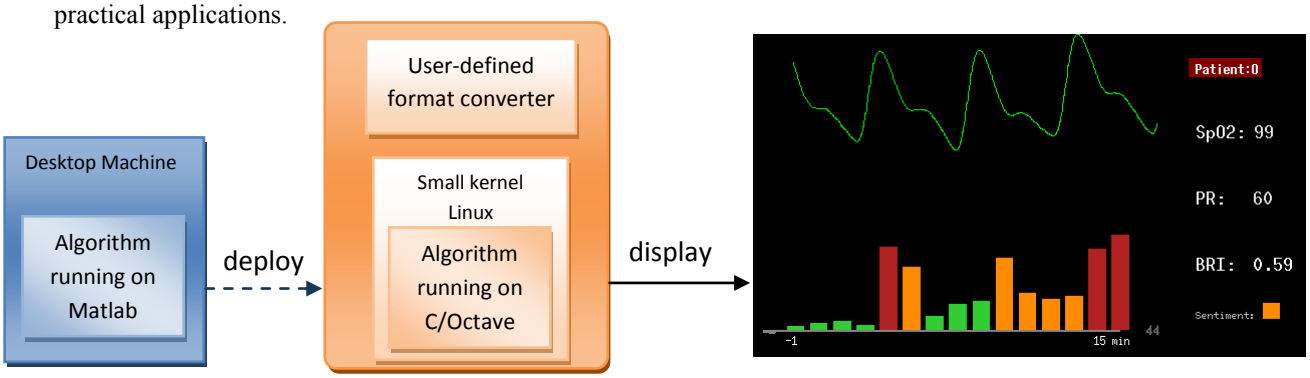
1. Department of Anesthesiology and Program in Trauma, University of Maryland School of Medicine. 2. US Air Force C-STARS Baltimore.

Introduction: High fidelity continuous vital signs waveform signal collection has great potential to assist emergency medical treatment, allowing real time visualization of the evolving physiological status of the patient and early detection of aberrant physiological events. In combination with vital signs feature selection and real-time processing, such information rich datastreams could potentially provide emergency decision-support on the need for life saving interventions, appropriate patient triage and resource acquisition or preparation for patients. Clinical deployment of algorithms requires both rapid prototyping and testing of algorithms' accuracy and robustness. Major obstacles to rapid prototyping are proprietary vital signs monitors with proprietary noise reduction algorithms, compression algorithms, and data storage formats and the commercial programming software environment, such as Matlab or LabView. The aim of this study is to describe the development of a vendor-neutral prototype system in the form of an inexpensive mobile device for the validation and testing of algorithm accuracy and robustness, testing and optimizing human factors concerns, and usability analysis.

Methods: We developed an algorithm to predict the probability of near-future blood transfusion or the Bleeding Risk Index (BRI) based on features of continuous photopletysmograph (PPG) 240Hz waveforms collected from pulse oximeters during trauma patient resuscitation. We designed a vendor-neutral platform to deploy our algorithm for near-real time use and testing. To minimize device start-up time and real time data processing and analysis by our BRI algorithm, a configurable, light-weight Linux operating system was utilized. The BRI algorithm was coded in native linux in a modular fashion, allowing future upgrades to decode proprietary datastreams or future algorithm optimization and development.

Interface design: To provide user friendly display and to maximize the information presented on a space-limited device, we divided the display into three areas. The top left displays the past 3 seconds raw PPG waveform. On the bottom, the past 15 minute BRI values are displayed minute by minute, color coded to corresponding to pre-defined 'critical' (RED: BRI≥0.5), 'neutral' (YELLOW: 0.2≤BRI<0.5), 'safe' (GREEN: BRI<0.2) threshold levels of the need for blood transfusion. The right most column, displays real time readings of oxygen saturation (SpO2), pulse rate (PR), BRI, and a sentiment score which aggregates past 15-min BRI as an overall indicator.

Results & Discussion: Proprietary medical devices and platforms do not allow for the implementation and testing of new algorithms. We developed a vendor-neutral platform deployed on small single-board computers for fast prototype algorithm validation, development, and user feedback. In preliminary tests, representative data files from 8 patients with a total of 2 million data points were processed immediately after collection with an average run time under 0.5 seconds for each BRI calculation, matching the efficiency of an a top of the line computer. The flexibility of streaming external device-independent data into the platform allows our algorithms to be quickly tested for many practical applications.



Using Narrative Storyboards to Inform Design Improvement of Hands-free Communication Devices

Yushi Yang, MS[1], Joy Rivera, PhD[1], Susan Bethel, MSN, RN[2]
[1]Clemson University, Clemson, SC; [2]Greenville Health Systems, Greenville, SC

Introduction: In recent years, there has been an increase in the implementation and use of Hands-free Communication Devices (HCD) in the clinical settings. While adding such devices to an already complex system can in some ways benefit healthcare professionals' (HCPs) communication, many questions arise related to their appropriate use and their fit within HCPs' workflow. Research using grounded theory found that nursing performance related to communication access is improved because HCD provide nurses with the flexibility to contact other HCPs anywhere and at any time[1]. Other research using quantitative data analysis showed that HCD increased the rate of responses and therefore increased efficiency of nurses' workflow[2]. However, there is little research focusing on the evaluation of HCD and their use within context to inform design improvement opportunities. Additionally, there is little evaluation of the design of HCD based on the results of field research conducted from a sociotechnical perspective[3]. The purpose of this study is to examine the end-user's interactions with HCD, to learn how they impact nurses' performance within the working environment, and to elicit design recommendations that will support a better work system.

Method: A total of 12 nurses were observed in two acute patient care units. Each nurse was observed for 2.5 hours totaling 30 hours of observations. The observation data were aggregated, coded and categorized into themes using the thematic analysis software NVivo 10. Three overarching themes were identified: interruptions, operational failures, and usability. Under each theme, narrative storyboards were created based on real-world data. Narrative storyboard is a method to describe the human-product interaction. It is informed by real-world field observations and the deliverable is a sequence of images, showing the location of the interaction and tasks individuals complete while interacting with the product[4]. By using this method, the most prominent problems related to HCD can be easily translated into a language that product designers understand.

Results: The final results, representing unintended consequences of HCD, were presented in a narrative storyboards format. For the interruption theme, the storyboards describe a nurse working on patient charting in a hallway while a call comes in, however, the nurse has limited mental resources for multitasking on both documentation and talking via the HCD. For the operational failure theme, the storyboards describe a nurse in the patient room with poor reception signal, which required the nurse to switch the call from the HCD to a telephone that is outside the patient room. For the usability theme, the storyboards describe a nurse who was in the patient room administering medication to her patient; she rejects an incoming call by saying "no" but her voice was not successfully recognized by the device. Each of the three storyboards includes a set of images with associated annotations.

Discussion and conclusions: From the storyboards, it is concluded that HCD can increase nurses' workload and escalate the complexity of the medication administration workflow during certain contexts. Therefore, designers need to consider the management of mental resource availability, which can be achieved through a program that is able to proactively reject an incoming call based on an automatic assessment of the nurses' availability. Also, the voice quality and voice recognition needs to be improved. High-level system design recommendations from the perspective of sociotechnical perspective were also proposed based on the storyboards, for example, the integration of Radio-frequency Identification (RFID) or other locating technologies with HCD would let the device recognize nurses' location. When it is identified that the nurse is close to the patient when administering medications or doing other direct patient care tasks like repositioning the patient, it would not allow incoming calls that are not emergency.

References

1. Richardson J, Ash J. The effects of hands-free communication device systems: communication changes in hospital organizations. J Am Med Inform Assoc. 2010 Jan-Feb; 17(1): 91–98.
2. Kuruzovich J, Angst CM, Faraj S, Agarwal R. Wireless communication role in patient response time: a study of vocera integration with a nurse call system. Comput Inform Nurs. 2008; 26(3):159–66.
3. Hendrick HW, Kleiner BM. Macroergonomics: An Introduction to Work System Design.Human Factors and Ergonomics Society; 1999
4. Greenberg S, Carpendale S, Marquardt N, Buxton B. The Narrative Storyboard: Telling a story about use a context over time. Interactions. 2012 Jan-Feb; 19(1): 64-69.

Visualization of Publication Timelines using 4K Monitors

Dean Yergens[1,2,*], Evan Minty[1], Christopher J Doig[1]

1.University of Calgary, Canada, 2.Healthcare Simulations Inc,, Canada

Abstract

Introduction

High resolution monitors (4k) are beginning to enter the marketplace. These 4K monitors have the ability to represent 4 to 6 times the standard resolution of a typical monitor. With this increase in resolution the ability for displaying increasingly complex visualizations is possible. One such application area is the visualization of literature review references due to the massive amounts of information that is often included. We present a web-based application developed for displaying academic publications based upon a national public health cross sectional survey using these high resolution displays.

Methods

A literature review was previously conducted to examine how the Canadian Community Health Survey (CCHS) was presented in the literature during the years 2002 to 2012. The results (data) of this literature review were then exported into a custom developed web-based application for displaying the literature information. The application consists of several visualization techniques including a Timeline approach for displaying temporal data, calendar view for displaying frequency of publications and standard charting techniques (bar, line, pie) for displaying quantitative information. The java-script visualization frameworks D3.js and MIT SMILE widgets were used in the development. A 39" 3840x2160 high resolution monitor was then used to display the web-based application.

Results

677 references from the CCHS literature review were displayed in the application. A Timeline visualization which represents the detailed individual information about the references was displayed by title with the timeline interface having the ability to scroll across the multiple publications. Supplemental information on the reference is available through clicking on the reference. This supplemental information included the authors, journal, abstract and a link to the PDF article. The references were color coded based upon which year of the CCHS survey was used in the analysis for the paper. Aggregated information including publication year, top journals and frequency of publications is also displayed at the bottom of the interface.

Conclusion

High resolution monitors present the next evolution in the visualization of health informatics information. Our web-based system presents one application where these higher resolution monitors may be of benefit. Future research will include evaluating information understanding using the 4K interface compared to standard size based monitors and mobile devices.

A New Corpus for Structured Microbiology Results

Wen-wai Yim[1], Xavier Engle[2], Heather L Evans[3], Meliha Yetisgen, PhD[1,4]
[1]Biomedical and Health Informatics, [2]School of Medicine, [3]Department of Surgery, [4]Department of Linguistics, University of Washington, Seattle, WA

Abstract

Microbiology results data is an important part of clinical diagnosis, disease treatment, and outbreak detection. However, the free-text form of microbiology reports limits their incorporation into real-time surveillance applications. Our task is the extraction of microbiology information from free-text into a structured form for secondary use. To create an extraction system, it is important to define a template that captures microbiology results, such as bacterial microorganisms, microorganism concentration, and drug susceptibilities. In this abstract, we describe a corpus annotated using our microbiology results template. Our corpus is comprised of microbiology culture and gram stain reports, where we extract results from a variety of microbiology reporting free-text.

Introduction

Microbiology results provide clinicians with the relevant information to identify sources of bacterial infection, determine between differential diagnoses, and adjust antibiotic treatment. However, these results are represented in dense, domain-specific free-text statement. In order to leverage such information in downstream surveillance applications, such as our ongoing ventilator-associated pneumonia project, we thus needed to develop an extraction system for microbiology results. To accomplish this, we defined a microbiology results template with help from a clinical expert to capture relevant infection information and used this template to annotate microbiology reports. Our annotated corpus included 1442 microbiology reports.

Annotation

In our template, microbiology results in our corpus were represented as a collection of the microbiology results entities: (1) *organism*—a microorganism found in a culture (e.g., bacteria, flora, fungus, yeast), (2) *organism quantity*—a measurement of the amount of the organisms found in a culture (e.g., >10,000 col/ml, one colony, no, isolated), (3) *rating*—a qualitative measurement of the amount of organisms found in a culture (e.g., 1+,2+,3+,4+), (4) *drug*—a drug that was tested on an organism (e.g., penicillin), (5) *drug resistance*—a susceptibility of an organism to the drug (e.g., susceptible, intermediate, resistant, no CLSI interpretive criteria), (6) *MIC*—minimum inhibitory concentration of an antimicrobial that inhibited the growth of a microorganism after overnight incubation (e.g., 2.0 µcg/ml), and several other entities. A template can have several of the same entities, while some fields may be empty. Since microbiology reports may describe multiple cultures, multiple templates are needed. To determine which entities belong together, entities in the same template are joined by relations. Figure 1 shows two examples of microbiology reports annotated with entities and relations. Two annotators, one medical student and one graduate student, annotated microbiology results for 100 records. The inter-rater agreement showed an entity-level f1-measure of 0.982 (macro) and 0.964 (micro), a relation-level f1-measure of 0.937, and a template-level f1-score of 0.833. Using the brat rapid annotation tool, one annotator annotated the rest of the 1442 report corpus with 3720 entities and 1196 templates. We plan to release the corpus through our research lab's website (http://depts.washington.edu/bionlp/index.html).

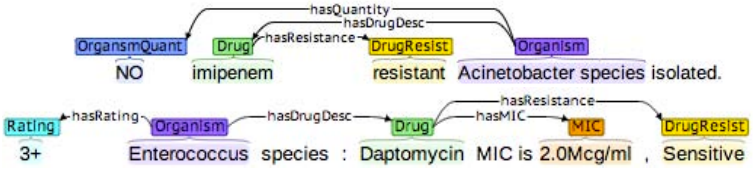


Figure 1. Example microbiology reports with entity and relation annotations.

Future Work

We used this corpus to build an information extraction system based on natural language processing and machine learning to extract culture information from microbiology reports. Our current entity, relation, and template extraction f1-score performances are 0.889, 0.795, and 0.606 respectively. The details of the extraction tool will be presented in another paper. Our ultimate goal is to incorporate the extracted microbiology results as features in an overall statistical real-time surveillance. One application is in our ongoing ventilator-associated pneumonia project.

Acknowledgements

AHRQ UW CER Career Development Award K12 HS019482, Microsoft Research Connections, University of Washington Research Royalty Fund, Institute of Translational Health Sciences UL1TR000423.

Application of Data Mining Techniques to Predict Physical Activity

Sunmoo Yoon, RN, PhD[1,2], Suzanne Bakken, RN, PhD[1,2]
[1]School of Nursing, [2]Dept. of Biomedical Informatics, Columbia University, NYC, NY

Abstract
We applied data mining techniques to a community based behavioral dataset to develop prediction models for gaining insights for a social media based intervention for an urban Hispanic population. Main predictors for active transport were look up internet for seek health information on Internet and perceived stress. The main predictors for sitting time were age, ability to make time and having a place for exercise. Data mining methods were useful to build physical activity prediction models prior to designing self-management interventions.

Introduction
Despite proven health effects, promoting physical activity has been challenging in population health. 'Promoting walking to school or work (active transport)' is the most recent community based guideline by CDC. And HealthyPeople 2020 emphasizes 'reducing screen time' as a community-based strategic plan. The Washington Heights/Inwood Informatics Infrastructure for Comparative Effectiveness Research (WICER) project aims to build an infrastructure to understand health behaviors to improve the health among an urban underserved population. Seventeen percent of those who completed the WCIER survey use social media. Given this, social media has a potential to be an interventional platform for promoting physical activity. To examine this potential, this study aimed to develop prediction models for physical activity by applying data mining techniques to the WICER dataset.

Methods
We extracted data on the sample participated behavioral survey (n=6,579) from the WICER RDW using the RedCap. The unique sample (n=5,868) was identified using manual review and automated abnormality detection mining techniques such as temporal activity sequence detection. 127 behavior variables were iteratively selected from 945 variables by a domain expert (SY). 114 variables were further selected after applying M99.0 algorithm to remove useless attributes. 51 unique variables were selected using *CFS attribute evaluator*[1], to select variables that were strongly related to the physical activity. 10 variables were ranked according to its degree of correlation to physical activity after manual removal of variables of multicollinearity. Physical activity was operationalized as an outcome variable of 1) engaging in active transport – walk to school or work or do errands (yes or no) , and 2) screen time for TV or video watching (≤3 hours a day, > 3 hours day).We applied several classification algorithms (J48, RandomForest, MultilayerPerceptron, AdaboostM1), to iteratively generate models with random 10-fold validation. We selected final models based on predictive ability and clinical meaningfulness of variables.

Results
Perceived stress (confidence score of handling personal problem, 0-4 scale, higher; more confident) and use the Internet to seek health information were the main predictors of active transport (accuracy 65%, AUC=.63). Those with confidence scores >2 were more likely to engage active transport. Among those with confidence scores <3, using the internet for seek health information was associated with increased likelihood of engaging in active transport (Figure 1). The prediction model for sitting behavior had higher accuracy (accuracy 68%, AUC=.69%). Age, ability to make time and have a place for physical activity were the main predictors for screen time. Participants less than 59 year old were more likely to have shorter TV or video time (Figure 1).

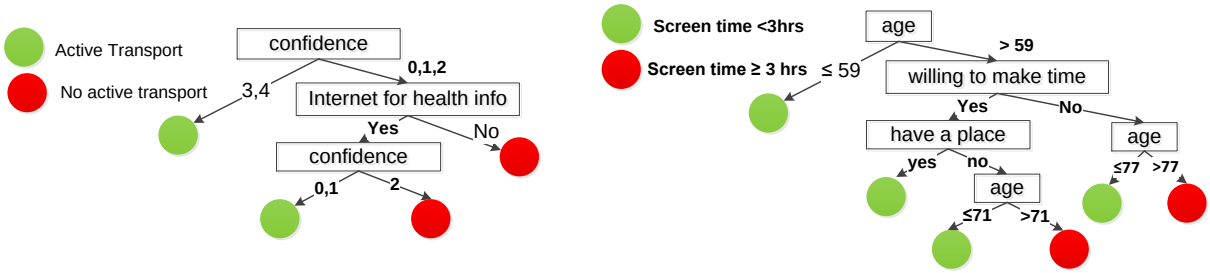


Figure 1. Prediction model for physical activity; active transport (left) and screen time (right)

Conclusion
Data mining methods were useful for domain experts to build prediction models for physical activity.

Acknowledgement: This study was funded by WICER4U R01 HS022961 (PI: Bakken)

References

1.Hall M., Frank E., Holmes G. et al. *The weka data mining software: An update; sigkdd explorations*. 2009

M-health for Individuals with Dexterity Impairments: The Needs & Challenges

Daihua X. Yu, MS, Bambang Parmanto, PhD, Brad E. Dicianno, MD, Valerie J. Watzlaf, PhD, Katherine D. Seelman, PhD
University of Pittsburgh, Pittsburgh, PA

Introduction

Mobile health (mHealth) using a Smartphone is a cost-effective solution to improve health through supporting self-care tasks. However, before people with dexterity impairments who experience difficulties in manipulating small buttons, icons or control scan harness the potential of mHealth, the accessibility of mHealth apps must be examined. We have developed a mHealth system for supporting self-care and adherence to self-care regimens, called iMHere (interactive Mobile Health and Rehabilitation). The overall goal of this study is to explore and to identify the accessibility needs and preferences of individuals with dexterity impairments when they use the iMHere system.

Methods

Two iMHere apps were used in this study: 1)MyMeds app for managing medications; 2)Skincare app for managing wounds and other skin issues. Nine subjects were asked to use these apps for approximately one week in a field trial. A lab test implementing the following tasks was conducted after the field trial: 1)Scheduling a new medication alert, which includes searching and finding the correct medication and setting up a medication schedule; 2)Modifying a medication reminder; 3)Scheduling a skin checkup alert; 4)Responding to a skincare reminder that includes taking a picture and reporting issues. The time for a subject to complete each task, the number of possible errors a subject made, and the number of errors a subject was able to self-correct were recorded and analyzed. The Telehealth Usability Questionnaire (TUQ) data was collected during the interviews so that subjects could rate their experiences and give feedback on usability factors (e.g. usefulness, ease of use, effectiveness, reliability, and satisfaction). Framed questions were also asked and responses were recorded by the researcher during the in-depth interviews.

Results & Discussion

Nine subjects with various levels of dexterity abilities were included in the study. Ages ranged from 18-55 years, including four women and five men. Eight had spina bifida (SB) and one had a spinal cord injury (SCI). Based on Pegboard instrument scores, subjects were classified into three groups: 1)Subjects with mild dexterity impairments; 2)Subjects with moderate dexterity impairments; 3)Subjects with severe dexterity impairments. A significant difference in error ratio was found among the three groups: $F(2, 33)=3.604$, $p=0.038$, using ANOVA. Bonferroni's t-test revealed that group 3 subjects had significantly higher error ratios than subjects in group 1 ($p=0.045$). A statistically negative correlation was identified between subjects' dexterity levels and their error ratios by using a Pearson product-moment correlation, $r=-0.434$, $n=36$, $p= 0.004$. This means subjects with a higher degree of dexterity impairments may experience more problems in task completion. Approximately 51% of errors were self-corrected without any help, but other errors called for resolution from a researcher. The average TUQ score was 5.94 (out of 7 points). Although subjects were overall satisfied with the iMHere system and stated they would like to use it in the future (average=6.39), the sections for "ease of use & learnability," "interface quality" and "reliability" received lower scores (5.56, 5.67 & 5.56 respectively).

Conclusion

The current design of the iMHere system might not be suitable for users with dexterity impairments. Due to the diversity of subjects dexterity impairments, their needs and preferences differ one from another. The most common suggestions from subjects to improve the accessibility of iMHere system are as follows: 1)Simplify and reduce the complexity of these apps; 2)Provide longer training time; 3)Provide feedback within the apps themselves to users as to whether they are doing the task correctly. In addition, more than half of the subjects commented on the button size. They would prefer to have larger buttons. Two of them indicated that they might be more comfortable with dark text on a white background. Some of them would like to try different picture backgrounds to make the apps more personalized. Features identified by subjects in this study will help developers and designers enhance the usability and accessibility in future studies. Finally, personalized design may be the key to approaching thse challenges in improving accessibility.

PinTopics: A Tool for Visualizing Topic Models Using Multiple Redundantly Coded Word Clouds

Zhiguo Yu, MS[1], Todd R. Johnson, PhD[1]
[1]The University of Texas School of Biomedical Informatics at Houston, Houston, TX

Abstract

Topic models are powerful tools for automatically characterizing and categorizing documents and their contents, but there is very little work on user interfaces that support human use of topic models for understanding and navigating sets of documents. In this poster, we present a tool that allows a person to easily and interactively use topic models to navigate, explore and understand large sets of PubMed citations by presenting topics as word clouds that use redundant visual features to make the gist of each topic "pop out" of the display.

Introduction

Topic Models[1] are a class of statistical machine learning algorithms that when given a set of natural language documents, extract the semantic themes (topics) from the set of documents, and also describe the topics for each document, and the semantic similarity of topics and documents. Although researchers continue to develop better and faster topic model algorithms, there is very little work on user interfaces that support human use of topic models for understanding topics and navigating sets of documents. Topic Browser[2], developed by Chaney and Blei, has shown great potential to summarize the corpus and reveal the relationships between topics and documents. However, each topic in Topic Browser is displayed by a list of ranked words (Figure 1a), which makes it a challenge for users to see what the different topics are about (Figure 1c). Our tool, PinTopics, displays each topic by using a word cloud that uses redundant visual codes (font size, intensity, and color) to encode word frequency. These redundant codes create a visual saliency map that can guide the observer's attention, making the general gist of each topic "pop out" of the visualization.

Figure 1 (a) and (b) compares the Topic Browser and PinTopics representation of a single topic. Figure 1 (c) and (d) compares the Topic Browser and PinTopics representation of multiple topics. The Topic Browser does not use visual salience to convey word frequency, making it difficult to get an immediate sense of what each topic is about. In contrast, PinTopics' use of redundant visual codes makes it easier and faster to scan topics and comprehend their meanings. We are presently conducting usability tests to compare PinTopics and the Topic Browser on a number of tasks using PubMed query results. We are also exploring algorithms for creating topic models in real time to support highly interactive query refinement.

(a), list-of-word (b), word-cloud (c), list-of-word topics overview (d), word-cloud topics overview

Figure 1. (a) and (b) are the same topic: 'Clinical Natural Language Processing'. (c) and (d) show multiple topics. We used LDA-C to generate 60 topics from 7095 citations (titles and abstracts) returned by the PubMed query 'biomedical informatics'.

References

1. Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of Machine Learning Research 3 (2003): 993-1022.
2. Chaney, Allison June-Barlow, and David M. Blei. "Visualizing Topic Models." ICWSM. 2012.

Informational Support Exchanges in Online Health Communities

Shupei Yuan, MA[1]; Jina Huh, PhD[1]
[1]Michigan State University, East Lansing, MI

Abstract
Informational support provides critical help to those visiting online health communities. In our small pilot study with an online diabetes community, community members received informational support the most, and the main contents of informational support consisted of clinical information and patient expertise. In our main study, we examine how community members respond to conversation initiators who ask for these two kinds of information sources and, depending on the response, how the conversation threads evolve accordingly.

Introduction
Studies have found that online social support can bring positive influence for various diseases[1,2]. Heaney and Israel[3] argued that social support consists of instrumental, informational, appraisal and emotional support. Among these social support types, informational support can play a critical role in fulfilling community members' information needs while visiting online health communities[4]. When a community member is asking for informational support, whether the members will receive the source of information they ask for is still a question. Our research questions are: What types of informational support are being asked and received in online health communities? How do community members provide the information that the conversation initiators ask for?

Pilot study
To answer the research questions, we first conducted a pilot study to gain insights on developing codes for examining informational support exchanges in online health communities. We chose the WebMD online diabetes community as our starting dataset due to its active, diverse support activities among community members[2]. We randomly selected 250 posts and qualitatively examined, using open coding[5], centering on our research questions.

Our preliminary results showed 56.1% of the informational support provided in the replies contained clinical information and 47.7% of the threads were patient expertise (not mutually exclusive). 81.3% of the conversation initiators were looking for clinical information and 96.1% of them received clinical information. Although members received needed clinical information, needed patient expertise was not necessarily received. Our main study then is to further examine the possible reasons for this phenomenon.

Main Study Design and Conclusion
We use a larger sample to further understand how clinical information and patient expertise exchanged. Conversation initiators' replies can reflect their engagement and attitude toward the threads and the information received. To analyze the conversation initiator's follow-up with the replies, we focus on conversation threads with only one unique replier. We randomly selected 20% data that fit (277 cases) as our study sample.

With the data collected, we will code and record the following items: The type of informational support asked; the type of informational support offered by the first replier; the engagement of the initiator. The concept "engagement" in this study includes whether initiators expressed appreciation, query, or other feedback, etc.

We will further conduct simple linear regression tests to help us understand the relationship between initiators' engagement with the type of information they received. Moreover, moderation analysis can help us find the factors that moderate this relationship, such as the length of the conversation.

Our work helps to further develop online social environments in which patients receive enriched social support specifically around receiving peer patients' support unlike other clinical environments. The findings from this study can provide insights to how we can help online forum users exchange appropriate information. Our study also helps the community organizers provide better solutions in facilitating informational support exchange.

Reference

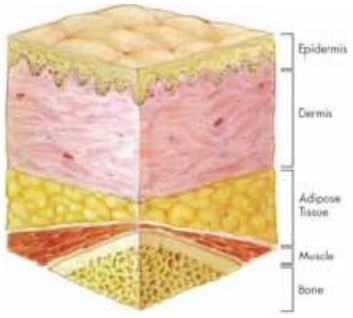
1. Meier A, Lyons EJ, Rimer BK, Frydman G, Forlenza M. How cancer survivors provide support on cancer-related Internet mailing lists. Journal of Medical Internet Research. 2007; 9(2).
2. Huh J, Mcdonald DW, Hartzler A, Pratt W. Patient Moderator Interaction in Online Health Communities. in Proc. Am. Med. Informatics Assoc. 2013; 627–636.
3. Heaney CA, Israel BA. Social networks and social support. Health behavior and health education 2002.
4. Frost, JH, Michael PM. Social uses of personal health information within PatientsLikeMe, an online patient community: what can happen when patients have access to one another's data. JMIR. 2008; 10(3).
5. Strauss AL, Corbin J. Basics of qualitative research. Sage Newbury Park, CA:1990.

Human-Centered Design in Wound Care Guidelines

Rafeek A. Yusuf, MBBS[1], Amy Franklin, PhD[1]
[1]The University of Texas Health Science Center at Houston, School of Biomedical Informatics, Houston, TX

Abstract
Current clinical guidelines, including those for wound care, are often generated following medical content rather than principles of user centered design. Although guidelines may be shared with patients and caregivers, appropriate modifications for non-clinical users may be absent. Using the TURF framework for usability in human-centered design, we explore methods for creating useful and useable guidelines in the home setting.

Introduction
Although clinical practice guidelines are defined as "systematically developed statements for practitioner and patient decisions about appropriate health care for specific clinical circumstances"[1], it appears current guidelines have laid more emphasis on the practitioner than the patient. For example, most health education materials, including guidelines, are written at the 12th grade level without consideration of the average American's 8th grade reading ability[2]. We propose that usability methods often employed to assess, develop, and redesign medical technology such as medical devices and electronic health records systems can be extended to the domain of guideline materials. In our case study, we focus on wound care guidelines. Approximately 6.5 million people in the U.S. live with chronic wounds[3] and a third of these individuals require home based care[4] provided by family and other caregivers[5]. In this project, we explore homecare guidelines for pressure ulcers and highlight the need for (re)design following usability principles. Through expert review, we demonstrate areas of modification necessary given the health literacy level, reading ability and visual acuity[6] of the caregiving population using these guidelines.



Methods
A heuristic evaluation following Zhang's[7] principles was conducted by the authors on a published wound care guideline[8] directed towards nursing assistants and caregivers. Adherence to design principles such as reducing memory demands and match between the system and the real world was assessed.

Results
291 violations of 11 types were noted across the guideline. The average severity rating of these problems was 3.1 (i.e. major violation). For example, images depicting different classifications of wounds are stylized three-dimensional cross-sections not reflecting the actual view (point) of a caregiver (see diagram).

Conclusion
Usability in human-centered design can be applied to redesign useable guidelines for healthcare providers, patients and caregivers particularly in the home setting. Ongoing efforts include the redesign of this guideline to more strongly adhere to usability principles and to better serve the home care provider.

References

1. Institute of Medicine. Clinical practice guidelines: directions for a new program. (M. J. Field, & K. N. Lohr, Eds.) National Academy Press. 1990.
2. Billek-Sawhney B, Reicherte E, Yatta B, Duranko S. Health literacy: physical therapists' perspectives. The Internet J Allied Health Sci Pract. 2012;10(2):1-6.
3. Singer AJ, Clark RA. Cutaneous wound healing. N Engl J Med. 1999;341(10):738–746.
4. Appleby SL. Role of the wound ostomy continence nurse in the home care setting: a patient case study. Home Healthc Nurse. 2011;29(3):169-177.
5. Jones SL, Hadjistavropoulos HD, Janzen JA, Hadjistavropoulos T. The relation of pain and caregiver burden in informal older adult caregivers. Pain Med. 2011;12(1):51-58.
6. Williams MV, Parker RM, Baker DW, et al. Inadequate functional health literacy among patients at two public hospitals. JAMA. 1995;274(21):1677-1682.
7. Zhang J, Walji MF. TURF: toward a unified framework of EHR usability. J Biomed Inform. 2011;44(6):1056-1567.
8. Continuum Health Partners. Preventing and caring for pressure ulcers for nursing assistants and family caregivers. Continuum Health Partners, New York, 2010:1-9.

Placeholder Registration Addresses in a Regional Health Information Organization

John Zech[1], MA, Jason S. Shapiro[1], MD, MA, Gregg Husk[2], MD, Thomas Moore[3], MPA, Gilad J. Kuperman[4], MD, PhD

[1]Icahn School of Medicine at Mount Sinai, New York, NY, [2]Beth Israel Medical Center, New York, NY, [3]Healthix, Inc., New York, NY, [4]New York-Presbyterian Hospital, New York, NY

Abstract: *To match patient records across sites for health information exchange, regional health information organizations (RHIOs) rely on a master patient index (MPI). An MPI is responsible for matching demographic information for each unique individual, including home address, across multiple healthcare provider organizations. This field is commonly mandatory, and some registration staff may use nonsensical placeholder values that do not represent legitimate addresses to complete this field (e.g., "X") when legitimate address data is unavailable. This may cause the MPI to incorrectly separate records belonging to the same patient. We examine some of the most common placeholder addresses used in a functioning New York City RHIO and describe how the use of these addresses may affect record matching.*

Introduction and Background: A Master Patient Index (MPI) assigns patient records at different healthcare facilities to an internal MPI identifier using probabilistic matching algorithms.[1,2] Each MPI identifier is intended to represent a unique patient. The address data a patient supplies at registration is frequently used as part of this matching algorithm: if a patient's records at different sites have different address information associated with them, the MPI may not link them, causing it to appear as if a unique individual is actually more than one patient. This may happen when registration staff at different sites use different *placeholder* addresses when registering a single patient.

Methods: We analyzed patient registration data from Healthix, a New York City based regional health information organization (RHIO) that has enabled health information exchange. We linked registration records in the database using an exact match on first name, last name, and date of birth. We identified all address street lines with which 100 or more patients had registered. The number 100 was chosen as we believed it was very unlikely that a single legitimate residence would have such a large number of corresponding registrations. We determined which address street lines were *placeholder* values via manual review. We calculated the total number of patients who had been registered with a *placeholder* address.

Results: There were 587 address street lines with which 100 or more patients had registered. 84 of the top 587 address street lines (14.3%) were *placeholder* addresses. Examples include "1", "NEED", "…", "UNKNOWN", and "XX". There were 161,479 patients (2.1% of all patients) who had registered with one of these *placeholder* addresses.

Discussion and Conclusion: We determined that *placeholder* addresses were used for over 160,000 patients in Healthix data. These addresses may cause records not to be successfully linked by an MPI's probabilistic matching algorithm. With awareness of this issue, RHIOs can implement strategies to mitigate its impact on the performance of their matching algorithm. Realizing that *placeholder* addresses do not represent legitimate address information, a RHIO could treat *placeholder* addresses as missing values if their algorithm can accommodate missing data. Alternatively, a RHIO could reduce the weight their algorithm gives to address agreement when one address is a known *placeholder* address. Ideally, if human review is available, RHIOs could prioritize records with *placeholder* addresses for human review.

References

1. Fellegi IP, Sunter AB. A Theory For Record Linkage. *J. Am. Stat. Assoc.* 1969;64(328):1183–1210.

2. Grannis SJ, Overhage JM, McDonald CJ. Analysis of identifier performance using a deterministic linkage algorithm. *Proc. AMIA Symp.* 2002:305–9.

A Quest for HIE Success - Agency Theory and Technology Mutual Adaptation Framework

ABSTRACT: This work proposes a framework that explains the factors that hinder successful implementations of HIE solutions and offers potential solutions to overcome them. The framework approach is based on two theoretical lenses: technology implementation mutual adaptation model and agency theory. We argue that during HIE implementation, there are misalignments between available HIE technologies and the actual HIE stakeholder environment, which leads to a loss of productivity and the inability to reap the benefits promised by HIE. These misalignments are often compounded by the divergence of interests and ineffective cooperation of different stakeholders during HIE implementation and adoption. We propose a framework to explain what causes these misalignments in HIE implementation and adoption and how they can be mitigated for successful HIE implementations.

METHODOLOGY: In this study we adopt the technology implementation mutual adaptation model between user and environment proposed by Leonard-Barton [1]. According to Leonard-Barton, the initial implementation of technical innovations is best viewed as a process of mutual adaptation- the reinvention of the technology and the simultaneous adaptation of the organization. The main tenants of the theory is: 1) adaptation processes are necessary because a technology almost never perfectly fits the user environment, 2) misalignments can be corrected by either altering the technology to fit the environment or vice versa, or both and 3) misalignments result in productivity losses, which spurs the adaptations. Three types of misalignment are identified: 1) technical-which is the misalignment of technical specifications; 2) delivery systems- which are the organization's infrastructure, and 3) value- which is the job performance criteria. Misalignments leads to adaptive responses, represented as cycles in Figure 1, which can cause the redesign of technology (to fit the environment) or organizational change (to fit the design). Cycles can be (figuratively) large or small, where large represents changes that are more fundamental (e.g. the replacement of an old technology with a new one) and small cycles represent changes are less substantial (e.g. fine-tuning). This model was developed to explain the dynamics of technology implementation in very large companies. We, nonetheless, find that the model is very appropriate for the research of Health Information Exchange (HIE), and we make necessary modifications of the model to reflect the more macro, inter-organizational relationship based HIE research. We propose a modified framework. The two basic structures of our proposed framework, technology and environment, are slightly altered. More specifically, technological solutions (for HIE) replace technology because the latter is what the end users of HIE actually utilize and consequently, are concerned about. These solutions can be HIT vendors' HIE products, such as Epic Connections ®, or government solutions, such as Florida's Patient Look-Up (PLU), or some healthcare providers' privately owned, in-house developed HIE project. We replace user environment with all the different HIE stakeholders: healthcare providers (either the aggregate level—hospitals or individual level—medical professionals), payers (either the aggregate level—insurance companies or governments or individual level—patients), patients, software vendors, and the government. Multiple studies, and evidence from current HIE implementations, reflect that misalignments for HIE are common. On one hand, many HIE initiatives were terminated because of the lack of success. On the other hand, we suspect that the convergence of technology and environment are inevitable, but not necessarily in the short run. In any on-going HIE project, while misalignments exist, adaptive responses from both technology and environment are present. Technologically, large cycle adaptations are typified by the change of the technology (and its capabilities) offered by vendors, and/or the termination and re-initiation of HIE projects. We also observe complex dynamics among different stakeholders where different parties try to make adjustments in how they work and interact with each other in order to make HIE a success.

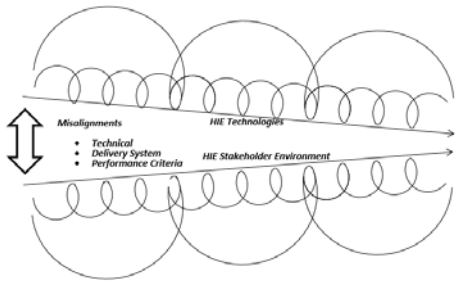


Figure 1- HIE Technology Mutual Adaptation Model adopted from Leonard-Barton (1988)

We argue that the stakeholder environment adaptation is the largest obstacle that hiders the development of HIE, which can be readily explained by the agency theory [2]. Agency theory posits that is any agency relationship there are two parties: a principal and an agent. The principle relegates work responsibilities to an agent but because of the asymmetry of information, the principle does not have total knowledge about what the agent is doing. Because of the divergence of interests between the principle and the agent, the principle must moderate the work of the agent, thus incurring agency costs. Because HIE's success is largely based on the participation of different stakeholders, and different stakeholders have different interests, the agency cost is often so high as to render HIE projects ineffective. For example, although HIE in theory can improve the quality of patient care, the improved patient care often comes as a loss for the HIE participating hospitals, for reasons such as reduced lab revenue due to the reduction of duplicated tests or reduced hospitalization due to a healthier population. Thus we argue that the alignment of the interests of different HIE stakeholders is of the tantamount importance, which may be achieved by formalizing technical standards for HIE and providing the HIE participants with both financial and policy support.

CONCLUSIONS: Using agency theory, technology implementation mutual adaptation model and IT Success model, we identifies factors that hinder the development of HIE and offer suggestions for solutions for future effective HIE implementations, which can be instrumental for both policy makers and practitioners alike.

REFERENCES: [1] D. Leonard-Barton, "Implementation as mutual adaptation of technology and organization," Res. Policy, vol. 17, no. 5, pp. 251–267, 1988.

[2] M. C. Jensen and W. H. Meckling, "Theory of the firm: managerial behavior, agency costs and ownership structure."

A Motivation Framework for Knowledge Translation in China

Yinsheng Zhang[1], Haomin Li, PhD[1*], Huilong Duan, PhD[1]
[1] Zhejiang University, Hangzhou, China

Abstract

Globally, there exists a great gulf between medical knowledge and clinical practice. Translating knowledge into actionable clinical decision support application has become the biggest challenge faced by evidence-based medicine. A comprehensive motivation framework was designed to support the entire knowledge translation life cycle, including knowledge acquisition, management, and application. The framework has been implemented in a 3000-bed Chinese hospital, and typical clinical decision support applications were developed based on the framework.

A Motivation Framework for Knowledge Translation

Knowledge translation has now become the biggest challenge faced by evidence based medicine[1]. This research aims to build a comprehensive motivation framework to support the entire knowledge translation life cycle, which includes not only acquisition and dissemination, but also "ethically sound application of knowledge" in healthcare system[2]. The framework uses a comprehensive knowledge base based on a unified ontology to manage various knowledge, such as drug use rules, diagnostic rules, clinical protocols, etc. Knowledge can be acquired by computer aided knowledge discovery techniques, or manually edited through a knowledge authoring web portal (demo at http://www.cktp.org:8006). Based on the knowledge base, fundamental knowledge-driven services such as inference engine and NLP-based data acquisition are provided as an infrastructure for developing clinical decision support applications. This framework was implemented in a 3000-bed Chinese hospital (DaYi Hospital, ShanXi Province, China), and typical CDS applications have been developed with the framework. This pioneer effort is expected to eventually promote knowledge translation in more healthcare institutes in China.

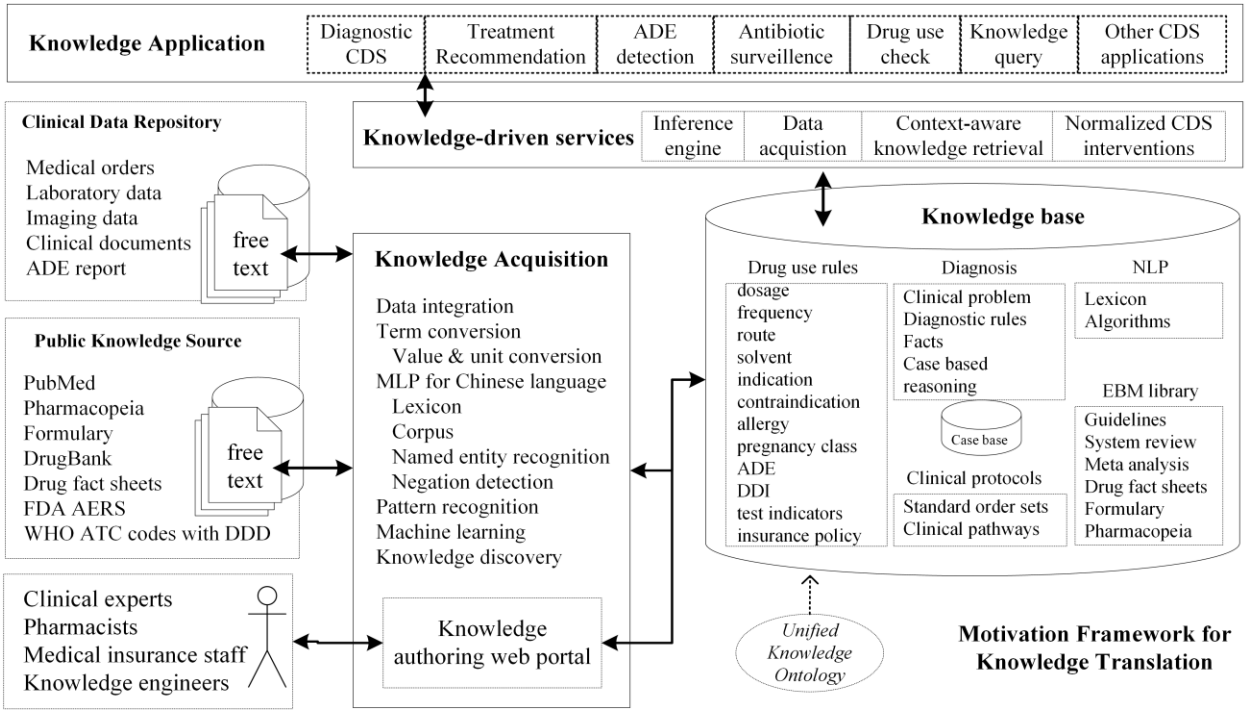


Figure 1. A motivation framework for knowledge translation.

References

1. Guyatt G, Cook D, Haynes B. Evidence based medicine has come a long way. British Medical Journal 2004;329:990.
2. About knowledge translation & commercialization. (Accessed 2014, at http://www.cihr-irsc.gc.ca/e/29418.html.)

An Extensible Integration Framework for CDS Applications

Xiang Zheng, BS[1], Haomin Li, PhD[1*], Yinsheng Zhang, BS[1], Huilong Duan, PhD[1]
[1]Zhejiang University, Hangzhou, China

Abstract

The lack of mechanism to easily integrate CDS (Clinical Decision Support) applications into clinical workflow has become the major obstacle for the adoption and utilization of CDS applications. We design and develop an extensible integration framework to effectively manage and targeted deliver CDS applications to appropriate scenarios.

Introduction

Reducing medical errors and improving health care quality via CDS applications is one of the most important goals of medical informatics. While in practice, CDS application are frequently blamed for not being well integrated into the workflow, a major cause of which is lack of an extensible mechanism for integrating various CDS tools into CIS (Clinical Information System). Because of the deficiency of workflow integration and data interoperability, clinicians have to do burdensome and repetitive data entry work. Furthermore, the interventions produced by CDS tools are not directly actionable for user. Meanwhile, though various computer-aided tools such as calculators, evaluators and planners are helpful in clinical practice, effective mechanism to make them spread is absent.

To address the above problem, this study aims to provide an extensible framework for the dissemination, update and management of CDS applications. The mechanism is able to targeted delivery CDS tool to where it is needed, and achieve interoperability between CDS and hosted information systems.

Method

A registry web platform was designed and developed to manage CDS applications and associated these CDS tools to specific clinical problems and treatments. Based on this information, an integration agent module embedded in CIS ~~(clinical information system)~~ could smartly push the CDS tools to the right scenarios based on the specific patient information. The interoperability between CDS applications and clinical information system was achieved through a two-level interoperability protocol.

The first level interoperability was used to integrate CDS applications which are legacy and independent third-part tools. These kinds of CDS are unable and usually need not to directly exchange data with host CIS. Through the registration and association process CIS will know how and when to invoke this application.

The second level achieves data interoperation between CIS and CDS applications. This is supported by a plug-in mechanism. Data interoperability makes it real that not only clinicians have no need to input data repeatedly but also the results produced can be translated directly into actionable items in CIS. We designed a plug-in interface protocol to achieve this. In this protocol, basic information such as patient ID and user ID is sent to CDS tool so that it can get more information from CIS via a public Web Service, which provides comprehensive data query functions for CDS applications. A normalized data format of various interventions which includes adding clinical problem, placing orders and sending messages or alert is designed. The CIS is ready to accept CDS output in compliance with this format.

Result

A prototype has been developed and implemented in a 3000-beded hospital in China. Several protocol-complaint CDS tools have been developed and some other third-party ones have been collected to verify this framework, including DAS-28 Evaluator, customized micro-pump helper and BMI widget. These tools can be invoked to provide information based on advanced research to help physicians to make decisions, or make it easier at least. Moreover, the results of tools can be put into CIS in forms of clinical problems or medical orders, which are actionable. For instance, DAS-28 is a quantitative measure used to monitor the treatment of RA (Rheumatoid Arthritis). In our applications, this tool can be invoked from clinical problems like rheumatoid arthritis.Thus physicians can use this method to evaluate the activity of RA and make a better decision.

iGenetics: An Individualized Genetic Test Recommendation System Based on EHRs

Qian Zhu, PhD [a], Hongfang Liu, PhD [a], Christopher G. Chute, MD, DrPH [a], Matthew Ferber, PhD[b]

[a] Department of Health Sciences Research, Mayo Clinic, Rochester, MN

[b] Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN

Introduction

In individualized medicine, genetic screening and other tests give doctors more evidence for tailoring treatments to patients, which can potentially improve health care and save money. With the recent advances in genetic technology, genetic testing is available at more than 500 laboratories in the United States for more than 2,000 rare and common conditions and can help health professionals determine or predict patients' genetic conditions. However, physicians have not actively incorporated these tests into their regular clinical practices due to insufficient medical educational resources and clinical evidence as well as genetic testing guidelines. Therefore, we introduce an individualized genetic testing recommendation, called iGenetics, which leverages a high volume of patient information and well-evaluated testing guidance to provide a comprehensive genetic testing knowledge base (GTKB) and suggest appropriate genetic tests that might be positive for individual patients by exploring emerging semantic web technologies (SWT) and advanced informatics approaches.

System Description

iGenetics is to recommend appropriate genetic tests to individual patients. Consequently, two major fundamental components have been developed and described as below, data support module and test recommendation module.

A) Data support module: Genetic Testing Knowledge Base (GTKB) [1], a centralized genetic testing repository contains genetic testing information extracted from Genetic Testing Registry (GTR) including general test information, publically authoritative genetic testing guidelines (GTG), and patient medical information from Electronic Health Records (EHRs). More specifically, common clinical characteristics for each genetic test has been identified and deposited into the GTKB to support genetic test recommendation by iGenetics.

B) Test recommendation module: "Clinical characteristics identification" component aims to identify relevant clinical characteristics for individual patients via clinical characteristics mapping between the GTKB and SemMedDB[2], which provides drug-gene-disease associations. "Genetic test recommendation" component is to generate prediction models by leveraging machine learning, network analysis and SWT for test recommendation. Figure 1 shows the architecture of this recommendation module.

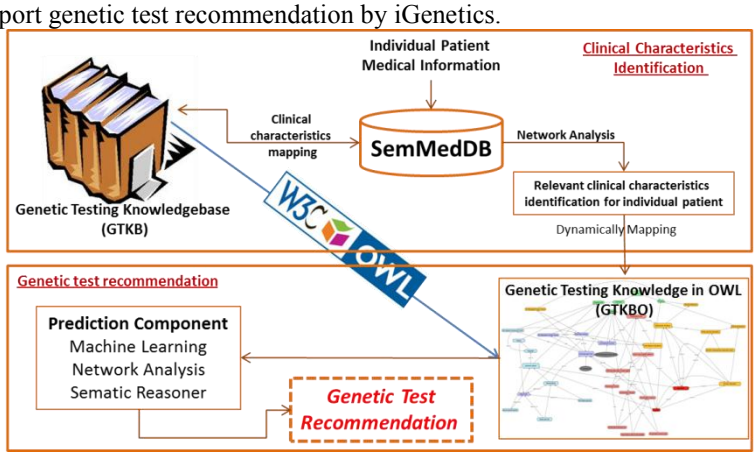


Figure 1. Test recommendation module for iGenetics

Discussion

We developed an intelligent genetic test recommendation system that not only provides a comprehensive genetic test knowledgebase serving as a genetic testing education resource, but also recommends appropriate genetic tests might be positive for patients on the basis of clinical evidence mined from the EHRs. iGenetics can encourage and assist physicians in incorporating genetic tests in their regular clinical practices to deliver high quality of health care.

Acknowledgement

This work was supported by the Pharmacogenomic Research Network (NIH/NIGMS-U19 GM61388) and partially by R01GM102282-01.

Reference

1. Zhu, Q., et al., *Genetic Testing Knowledge Base (GTKB) towards Individualized Genetic Test Recommendation – An Experimental Study.* submitted to AMIA 2014 Annual Symposium.
2. Kilicoglu, H., et al., *SemMedDB: a PubMed-scale repository of biomedical semantic predications.* Bioinformatics, 2012. **28**(23): p. 3158-3160.

Transforming NHANES Database to the OMOP Common Data Model

Vivienne J. Zhu, Rupa Makadia, Amy Matcho, Martijin Schuemie, Patrick B. Ryan
Janssen Research & Development, LLC, Titusville, NJ

Abstract: *We explored the feasibility of transforming survey-based data to the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM). The National Health and Nutrition Examination Survey (NHANES) data were transformed to five major OMOP CDM tables: PERSON, OBSERVATION_PERIOD, DRUG_EXPOSURE, CONDITION_OCCURRENCE, and OBSERVATION table. NHANES patient level data from 761 tables (7 cycles from 1999 to 2012) for 71,916 distinct participants were mapped to the OMOP CDM. NHANES source data were maximally mapped to the OMOP standard vocabularies using different strategies.*

Background: The OMOP Common Data Model (CDM) facilitates efficient, comparable and systematic large-scale drug safety analysis across disparate databases. However, most CDM transformations have been using claim-based databases, converting a survey-based data to the OMOP database has not been explored yet.

Methods: National Health and Nutrition Examination Survey (NHANES) includes demographic, socioeconomic, dietary and health-related questions. The examination component consists of medical, dental, physiological measurements, and laboratory tests. The CDC has been publishing these data every two years (cycle) since 1999. In this NHANES CDM transformation, patient level data were mapped to five major OMOP CDM tables: PERSON, OBSERVATION_PERIOD, DRUG_EXPOSURE, CONDITION_OCCURRENCE, and OBSERVATION table (Figure 1). Since no specific dates for survey events were provided in NHANES, we imputed that the observation period started at the first day of survey cycle (1st of January) and ended at the last day of survey cycle (31st of December). NHANES weightings (WTINT2YR, WTMEC2YR) for each participant were mapped to the CDM OBSERVATION table. Different mapping strategies have been applied to source codes to the OMOP vocabulary mapping: fuzzy string matching for NHANES drugs (previously parsed to single ingredients) to RxNorm ingredients, weighted string matching for NHANES questionnaires, and human expert matching (manually reviewing questionnaires for certain disease) for NHANES conditions.

Results: NHANES patient level data from 761 tables (7 cycles from 1999 to 2012) for 71,916 distinct participants were mapped to five major OMOP CDM tables. A total of 21 NHANES tables with pooled sample or subset of sample were not mapped to the CDM. A total of 1,967(95.7%) NHANES single ingredients were matching to the RxNorm ingredient concepts, 40 NAHANES conditions were mapped to SNOMED concepts, and 761 (9.7%) NHANES questionnaires were mapped to the LOINC concepts.

Conclusion: It is feasible to transform the NHANES data to the OMOP common data model. Although survey data usually vary on the data structure and have no standard code to record patient data, our study maximally maintains information of the NHANES raw data and effectively transforms most import information to the OMOP CDM. To evaluate the accuracy and robustness of the NHANES CDM, we will conduct a replicate study using the NHANES CDM compared with NHANES raw data in the near future.

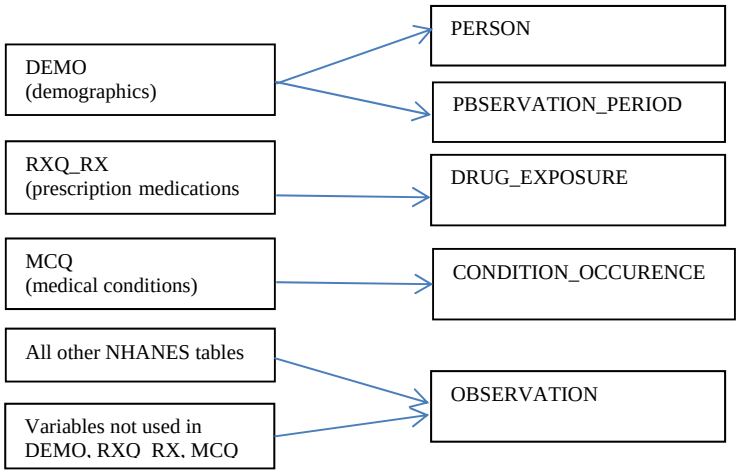


Figure 1. NHANES and OMOP Tables Mapping

Patient-Centered Decision Support for Pediatric Asthma Signs and Symptoms: Development of a Web-Based User Interface for Parents

Maryam Zolnoori,, Katherine Schilling, Josette F. Jones
m.zolnoor@iupui.edu

Introduction

Asthma is a complex, chronic disease with a heterogeneous distribution across in all regions of the World [1]. Epidemiological data for asthma, particularly in developing countries, show a considerable disconnect between asthma diagnosis, versus the actual prevalence of the disease [2].

In recent years, a variety of computer-based applications have been developed to assess and manage this chronic disease [6]. One of the negatives inherent in many existing, computerized applications is their minimal reliance on and inclusion of the gold standard, evidence-based research on asthma, particularly within pediatric populations [7]. Additionally, existing tools often fail to adequately represent medical knowledge to parents in text and formats that are easily read or understood.

Purpose

This poster describes the design and development of a patient-centered, web-based application for improving patient families' understanding of pediatric asthma signs and symptoms.

Methodology

Evidence-based literature, in combination with clinical decision support systems (web-based CDSS) approaches were leveraged in the development of the asthma application user interface.

Contextual inquiry techniques also informed the design, based on input from five nurses (for clinical representation) and parents (for familial understanding of asthma signs and symptoms). The interface was tested through heuristic evaluation, a method that identifies usability problems in the user interface design.

Results

Participants (nurses and parents) evaluated the application on three metrics including effectiveness, efficiency, and satisfaction. Additionally, the Questionnaire for User Satisfaction (QUIS) measured parents' satisfaction with the web-based tool for helping them understand their children's asthma signs and symptoms.

Initial results demonstrate that the application positively influenced patient engagement and patient empowerment, providing parents with a better understanding of the causes of pediatric asthma, and signs and symptoms with which they should discuss with their physicians for an official diagnosis. In addition, research may help the authors to understand the types of customization that could be considered in electronic personal health records when chief complaints include lung symptoms (coughing, wheezing, chest tightness, and shortness of breath) indicative of asthma. Detailed findings will be reported.

Conclusion

Improving health disparities in developing countries, particularly in asthma diagnosis and treatment depends, in part, on the development of quality, patient-centric, evidence-based applications.

Patient engagement is an important determinate of the success of health applications, particularly those that are designed to provide evidence-based information for understanding the signs and symptoms of pediatric asthma and other diseases. Patient engagement depends on several factors including user-interface design, content display and readability, and others. Ultimately, improving our understanding of the impact of system features and design elements on patient satisfaction and application usage will improve our abilities to develop products that empower parents to make informed choices on behalf of their children's health.

References

1) Bassam H Mahboub, Suleiman Al-Hammadi, Mohamed Rafique, Nabil Sulaiman, Ruby Pawankar, Abdulla I Al Redha, and Atul C Mehta, 'Population Prevalence of Asthma and Its Determinants Based on European Community Respiratory Health Survey in the United Arab Emirates', *BMC Pulmonary Medicine*, 12 (2012), 4.
2) CKW Lai, Richard Beasley, Julian Crane, Sunia Foliaki, Jayant Shah, and Stephan Weiland, 'Global Variation in the Prevalence and Severity of Asthma Symptoms: Phase Three of the International Study of Asthma and Allergies in Childhood (Isaac)', *Thorax*, 64 (2009), 476-83.
3) Maryam Zolnoori, Mohammad Hossein Fazel Zarandi, and Mostafa Moin, 'Application of Intelligent Systems in Asthma Disease: Designing a Fuzzy Rule-Based System for Evaluating Level of Asthma Exacerbation', *Journal of medical systems* (2012), 1-13.
4) Karin Yeatts, Carl Shy, Mark Sotir, Stan Music, and Casey Herget, 'Health Consequences for Children with Undiagnosed Asthma-Like Symptoms', *Archives of pediatrics & adolescent medicine*, 157 (2003), 540.

Online Patient Center: Expanding Patient Portals by Integrating Patient-Generated Data Directly into a Primary Care Provider's EHR Workflow

Jon-David Ethington, MHS, PA-C[1], Jianlin Shi[1], Scott Nelson, PharmD[2,1]
[1]Biomedical Informatics Department, University of Utah, Salt Lake City, UT;
[2]Department of Veterans Affairs, Salt Lake City, UT

Abstract

Patient portals are in need of new tools to increase patient engagement and improve communication between patients and providers. An Online Patient Center (OPC) is a repurposed, multi-tethered Personal Health Record (PHR) with connectivity to act as a single portal integrated with EHRs from multiple healthcare data providing organizations. The OPC contains an integrated app that guides patients in reporting clinically relevant data in a format preferred by providers, and integrates that data into the specified provider's EHR workflow. The OPC also allows for connected health apps, currently used by PHRs, which educate and promote health self-management. Beyond increasing patient engagement, an OPC provides closed-loop feedback and a reward system, which further encourages patient participation. An OPC utilizes the ONC's "Three A's" initiative to provide patients with better access to their healthcare data, an avenue to take action with that data, and tools to shift provider-patient attitudes towards collaborative efforts.

Background

Patient portals are patient-oriented E-health tools that aim to engage patients by providing access to their medical data and a method to electronically communicate with their health care providers. However, Patient portals have suffered from low adoption and are in need of new methods to engage patients [1] (Common barriers are explained with Appendix A in the supplement). Patient engagement has recently been identified as an area needing increased attention by the Office of the National Coordinator (ONC). In its primary role to improve health through health information technology (HIT), and to support the adoption of HIT, the ONC has developed a "Three A's" strategy to increase patient engagement. This strategy aims to increase **A**ccess of health information to patients, enable patients to take **A**ction with that information, and to shift **A**ttitudes so that patients and providers think and act as partners in managing health and health care [2]. More specifically, current patient portals do not effectively promote patient engagement due to a lack of integration in provider workflows and the need for using institution-specific portals. First, electronic messaging through patient portals is not well integrated into provider workflows (Figure 1b) [3], especially when patients want to take action by sharing data from externally obtained health care encounters (i.e. specialist visits, radiology reports, etc.) with their provider. In a review of patient portals, Ammenworth et al. [4] found that the use of patient portals resulted in decreased office visits, an increase of messages sent to providers, and no change in health outcomes. In outpatient settings where providers are reimbursed for services provided in-office, this translates to adverse effects on workflow with no additional reimbursement and can further hinder attitudes of the patient-provider relationship. Additionally, the risk of miscommunication with patient portal messages increases because in order to view patient messages, providers commonly need to open a tab or mailbox that is outside from their normal EHR software. In many cases, providers can even grant proxy access to other staff members so that they don't even check for or see messages. This discourages patient use of portals.

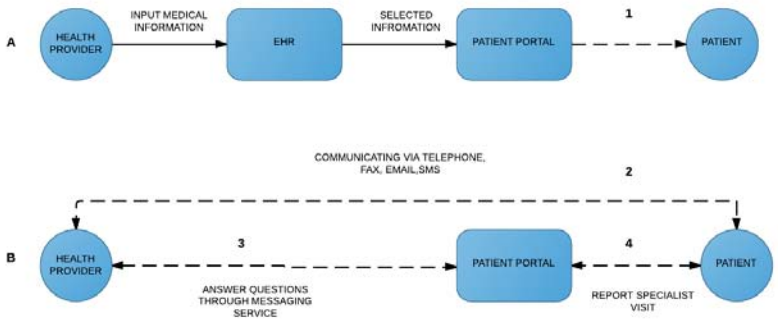


Figure 1. Current information flow when using a patient portal. Dotted lines indicate areas where barriers exist and are further explained in the supplement Appendix A

The second area that is a challenge for patient portals is that one portal typically links a patient to only one healthcare organization (Figure 1a). This creates an access problem for patients who receive care from more than one organization because they would have to create a new user account for each system they use in order to access their data. This scenario is very common, even in areas where only a few health care organizations serve a large portion of the population [5]. For patient portals to increase patient engagement and address the ONC's "Three A's" strategy, they need to integrate patient messaging into provider workflows, and provide a way for patients to access their healthcare data from multiple organizations in one location.

Approaches

Current alternative strategies to integrate patient messaging of specialist visits and other healthcare encounters into provider workflows include further implementation of the Direct platform [6], achieving system interoperability, and expanding current patient portals to integrate with provider EHRs. The Direct platform is a method to securely transmit personal health information to intended recipients, and transmit other document types. The Direct project however only sends and receives data. Any integration of messages into EHR workflows would be dependent on each EHR with which the system interfaces. Achieving system interoperability would be another solution, and would reduce the need for patients to report summaries of healthcare encounters to their providers because they would already be viewable in the provider's EHR. However, despite current efforts, there is much progress to be made in this approach. Another approach includes expanding current patient portals to allow patients to communicate specialist data directly into the provider's EHR workflow, but this approach may be difficult to convince EHR vendors to commit resources to build independently. Additionally, any data entered by patients through this method would then become part of the organization's data silo, and would not be owned by the patient. Furthermore, even if patients could communicate directly to providers' workflows, access would still require multiple logins, and data silos would still exist for patients who receive care from multiple healthcare organization.

An existing platform capable of providing more forms of communication between patients and providers than a patient portal is a personal health record (PHR). Despite the relatively low utilization of PHRs, it has been suggested that by integrating the PHR with the EHR, adoption would increase and patient-provider attitudes could be shifted in a manner that could improve health[7]. PHRs can be tethered to multiple EHRs to allow for direct connectivity between the PHR and EHR, without the need for patients to manually load their data. This could allow patient messages to be integrated into an EHR workflow, and would provide a way for patients to access their healthcare data from multiple organizations in one location. Furthermore, patients have more ownership of data in a PHR compared to a typical patient portal. New communication methods utilizing existing tools, such as those available from PHRs are needed to augment patient portals in new ways that foster patient engagement and provide better information exchange between patients and providers. This new application of existing technologies will be referred to as an Online Patient Center (OPC) hereafter.

An ideal arena to implement an OPC is in the primary care setting. Primary care providers (PCP) and PHRs have similar fundamental roles of "advising and supporting patients in education and health self-management." [8]. Patients would have the opportunity to have multiple interactions with such a system in this setting due to the frequent encounters that occur in a longitudinal fashion. It is also common for a patient to have multiple healthcare encounters in-between primary care visits, such as for specialist visits, emergency department encounters, or even hospital admissions. An OPC could provide much-needed, timely summaries of these encounters to PCPs at the point of care in order to provide high-quality healthcare. Furthermore, PCPs have generally positive views about integrating this type of data into the EHR as long as it does not adversely impact their workflows[3]. The primary objective of the solution presented here is to describe how an OPC can transform the nature of patient engagement by allowing a patient to electronically report information received during a specialist visit directly into his/her PCP's EHR workflow. This will be done in a way that meets Meaningful Use criteria, reduces the burden on the provider workflow, and empowers patients to play a more central, productive, and active role in their own healthcare.

Solution description

An OPC improves patient engagement by providing patients with access to their healthcare data, giving them the ability to take action with their health information, and shifting attitudes of patients and providers towards a more collaborative effort. An OPC is a repurposed, multi-tethered Personal Health Record (PHR) with connectivity to act as a single patient portal integrated with EHRs from multiple healthcare data providing organizations. Apart from

being a multi-organizational health portal, an OPC leverages PHR tethering connectivity to share clinically relevant patient-generated data directly into a specified provider's EHR workflow. For example, creating an email message from scratch may be difficult for patients. The OPC provides patients with a template that prompts them to enter the information that will be most valuable to their healthcare provider and then puts that information directly into the providers EHR workflow- not their message box. The OPC also adds an incentive points reward program, similar to credit card reward programs, to the list of patient health apps, currently used by PHRs, which can educate and promote health self-management. Because an OPC uses existing technologies, there are general requirements of PHRs and EHRs to implement an OPC in a primary care setting (Table 1). The specific functions of an OPC are listed in Table 2. A use case is found in Appendix B of the supplement and a link to a video demo is here: http://youtu.be/9E5H6PXoLIc

Table 1. *PHR, EHR, and healthcare organizational requirements for implementation of an OPC*

The use of an existing, secure PHR with either current or planned capabilities to tether with multiple EHRs	A healthcare organization currently using a qualifying EHR interested in expanding their current Patient Portal and meeting additional Meaningful Use requirements
An EHR vendor with connectivity capabilities to the selected PHR	All organizations should be willing to make adaptations to allow core functionality including EHR-OPC integration, user interface adaptations, etc.

Table 2. *OPC components and functions*

Integration of a patient-generated summary of a specialist visit into a primary care provider's EHR workflow
Component/Function: With guidance from providers and patient representatives, create a short form that will guide the patient in creating a clinically relevant visit summary of his/her specialist visit. • **Description:** Compared to having a blank email, this form will make it easier for a patient to describe their visit. Only the most critical pieces of information (type of provider seen, what they told the patient, were any tests done, etc.) will be transmitted to the provider, ensuring that any time spent reviewing the data will help move the clinic forward.
Component/Function: Provide the OPC module/app that creates a new section in the PHR called "Visit Summaries." This module allows the patient to fill out the specialist visit summary form, and saves it as "patient-generated" data. • **Description:** Provides a location in the OPC for secure patient to provider messaging of specialist visit data.
Component/Function: Once the patient completes the visit summary form, allow patient to specify which provider(s) will have access to the data. • **Description:** This puts patients in control of which providers see this data. Additionally, providers will not receive extraneous messages.
Component/Function: Link "patient-generated" data to EHRs as a "Patient-Generated Encounter." • Create screen space or an icon in the EHR workflow where a link to the patient-generated data can be displayed (i.e. Integrated into the existing chronological list of encounters already common to EHRs). • **Description:** Providers do not have to navigate away from their EHR software and to their messaging service to see the specialist visit summary.
Component/Function: When this information is accessed by a provider, send a notification to the patient explaining that the data they entered has been viewed. • **Description:** This closed loop feedback can positively shift attitudes of patients and providers towards collaboration because patients will know that their action of completing the specialist visit summary form were viewed as valuable information by their provider. As providers use this data it can additionally demonstrate that their patients are capable of accurately recalling medical encounters [9].
Component/Function: Reward points to the patient each time a summary visit is entered and each time that information is viewed by a provider. Provide a health-app marketplace in the OPC similar to credit card reward points where points can be redeemed for health products, discounts, etc. • **Description:** Uses a familiar, successful method to encourage and reward patient involvement.
Leverage PHR tethering to provide a single access point for patients to view data from multiple organizations
Component/Function: Provide a combined display containing all linked data shared by available EHRs in one, longitudinal format. • **Description:** One login for patients and one location to view data.

Healthcare organizations using OPCs have additional potential to meet additional Meaningful Use criteria. The specific functions of an OPC (Table 2) can help an organization attest to meeting Meaningful Use in the health outcomes policy priorities of engaging patients and families, and improving care coordination [10]. Depending on the function, one or both of these areas may be met in any of the Meaningful Use stages, one through three.

Implementation and Dissemination

Implementation of an OPC can be approached following a common Systems Development Lifecycle Model [11], consisting of five main phases: plan, analyze, design, implement, and maintain. Throughout all phases, tasks will be developed with the end goal of engaging the main two groups of end users- patients and providers. Key stakeholders will need to be identified in the first phase including PHR and EHR vendors, a participating healthcare organization, and provider and patient representatives. During the planning phase, a query to identify and secure core businesses that will participate in the points reward program will be initiated. While core components and functions (Table 2) are developed, stakeholders will also develop marketing strategies that will educate both patients and providers about the benefits of the system (Table 2) as well as how to enroll. During the design phase, the OPC functions will be integrated with the Meaningful Use system attestation module. Instructional video tutorials will be developed and imbedded into the OPC for viewing by patients. Additional instructional videos will be made available to providers and healthcare staff on their workstation desktops. Prior to system roll-out, core OPC functions will be deployed and tested and a baseline evaluation will be conducted to be used as part of the overall evaluation approach.

Dissemination of an OPC for use on a large scale would be facilitated by its foundation in existing PHR and EHR vendors. Once the core components and functions have been successfully implemented between the chosen PHR and any one particular EHR vendor, only a willing healthcare organization using that particular EHR vendor would be needed to implement another OPC. For every single EHR vendor that participates, it then becomes possible for any or all of their clients to participate. This scalable dissemination method could provide additional motivation for EHR vendors to participate beyond the benefits already discussed in this design.

Evaluation

A pre-post design will be used to evaluate the implemented functions, along with qualitative, descriptive, and human-factors research. Evaluations will be conducted through system audits and surveys administered to patients and providers at baseline (pre-implementation), 1, 3, and 6 months (post-implementation) to assess the OPC. Topics of interest include Meaningful Use attestations, usability of OPC interface using the Systems Usability Scale[12], satisfaction with specialist visit summary form, patient satisfaction with the point reward system, perceived engagement of patients, barriers to enrollment, provider satisfaction with availability and quality of information, impact on provider workflows, overall system adoption, and other perceived or realized benefits (patient outcomes). In order to minimize selection bias, surveys will be available electronically, via mail, and via in-person temporary staff members stationed in waiting areas and clinical areas. User engagement will be evaluated by the point reward system. Through measuring the system use as well as rate of system use, we can identify how both patients and providers are using the OPC.

Discussion

While an OPC uses existing PHR and EHR technologies, it is not just an appendage application to those systems. An OPC takes the best components of those systems, adds fully integrated electronic messaging, and provides these services to the patient in one online center. Because the OPC is reliant on PHR and EHR technologies, there is some inherent risk with this model, as key assumptions must be made. First, this project assumes that EHR vendors will be willing to collaborate with stakeholders and modify their user interfaces. Mitigation strategies for this potentially serious risk include achieving additional areas of Meaningful Use that the vendor may not otherwise meet on their own. By focusing on satisfaction of patients and providers, this may put added pressure on EHR vendors to participate. Secondly, the project assumes that patients will participate by logging into the OPC and completing the specialist visit forms. By providing a system focused on the ONC's "Three A's," we submit that patients will feel empowered with this opportunity to take charge of their own health information. Furthermore, patients will continue to use the system as they see their providers using that same information and they receive closed loop feedback. For those patients who need a different kind of motivation to use the system, the point reward system, which is similar to credit card rewards, may encourage participation. As the reward system develops, points earned by patients could be

used towards sponsor services or products (i.e. pharmacy discounts, wellness coaching, exercise equipment, gym memberships, etc.), or for reducing insurance premiums in a manner similar to many employee wellness initiatives.

By creating integrated EHR communication between providers and patients, as with this transformative approach, new opportunities of patient-provider communication and collaboration will appear. Rather than requiring patients to login to the OPC to report a specialist visit, the OPC could potentially use natural language processing to discover and process the specialist visit information from the patient via their smartphone or social media feed (see supplement Appendix C). Another example is when patients view radiology reports through the OPC and don't understand a particular section, the patient could highlight that line or paragraph, and when the provider views that same report, they will see what the patient highlighted. The provider would then know to address that during the next visit. Yet another possibility would be to allow providers to create links that appear on certain lab criteria (Vitamin B levels for example) explaining his/her opinion about what the optimal level for that test is. This provides the patient with more personalized guidance above and beyond the laboratory reference ranges commonly provided.

Conclusion

To emphasize the ONC's "Three A's" initiative, patients need better access to their healthcare data, an avenue to take action with that data, and our healthcare system needs tools to help shift provider-patient attitudes towards collaborative efforts. Primary care settings can especially benefit from increased patient engagement and are an ideal area to implement the OPC. A successful implementation of this solution will demonstrate that patients can understand the key aspects of their healthcare encounters and that they are motivated to share that understanding with their primary care providers in a manner that decreases the provider's burden. Further study will be needed to see if this new level of engagement can help lead to better health outcomes.

References

1. Bell, Heath. Portals Hold Promise for Patient Engagement but Challenges Remain. *iHealthBeat*http://www.ihealthbeat.org/perspectives/2012/portals-hold-promise-for-patient-engagement-but-challenges-remain (accessed 10 Jun2014).
2. Ricciardi L, Mostashari F, Murphy J, *et al.* A National Action Plan To Support Consumer Engagement Via E-Health. *Health Aff (Millwood)* 2013;**32**:376–84. doi:10.1377/hlthaff.2012.1216
3. Davidson E, Simpson CR, Demiris G, *et al.* Integrating Telehealth Care-Generated Data With the Family Practice Electronic Medical Record: Qualitative Exploration of the Views of Primary Care Staff. *Interact J Med Res* 2013;**2**:e29. doi:10.2196/ijmr.2820
4. Ammenwerth E, Schnell-Inderst P, Hoerbst A. The impact of electronic patient portals on patient care: a systematic review of controlled trials. *J Med Internet Res* 2012;**14**:e162. doi:10.2196/jmir.2238
5. Root, PhD J, Hoffman, MD M. National role of HIE's in healthcare's future: biomedical informatics challenges. 2014.http://healthsciences.utah.edu/videolibrary/player.php?id=0_j7czgitc
6. The Direct Project. http://directproject.org/ (accessed 31 Jul2014).
7. Kerns JW, Krist AH, Longo DR, *et al.* How patients want to engage with their personal health record: a qualitative study. *BMJ Open* 2013;**3**:e002931–e002931. doi:10.1136/bmjopen-2013-002931
8. Archer N, Fevrier-Thomas U, Lokker C, *et al.* Personal health records: a scoping review. *J Am Med Inform Assoc* 2011;**18**:515–22. doi:10.1136/amiajnl-2011-000105
9. Brown JB, Adams ME. Patients as reliable reporters of medical care process. Recall of ambulatory encounter events. *Med Care* 1992;**30**:400–11.
10. Health IT Policy Council Recommendations to National Coordinator for Defining Meaningful Use Final-August 2009. 2009.http://www.healthit.gov/facas/health-it-policy-committee/health-it-policy-committee-recommendations-national-coordinator-health-it (accessed 13 Jul2014).
11. Hunt EC, Sproat SB, Kitzmiller RR. *The Nursing Informatics Implementation Guide*. New York, NY: : Springer New York 2004. http://dx.doi.org/10.1007/978-1-4757-4343-2 (accessed 1 Aug2014).
12. Jordan PW, Thomas B, McClelland IL, *et al. Usability Evaluation In Industry*. CRC Press 1996.

On the SamePage: Supporting Communication and Informed Decision Making Through a Surgical Portal Extension

Rebecca J Hazen, Amanda Lazar
University of Washington, Department of Biomedical and Health Informatics, Seattle, WA

Abstract
Motivated by an experience of a close family friend, we present a patient portal solution designed to aid patients in making decisions about gynecological surgeries and procedures. Gynecologic conditions often have a stigma attached to them, and are not openly talked about. This can make finding and discussing information about them challenging. Surgical repairs for these conditions have long-lasting implications on many facets of a woman's life. After identifying challenges with current communication and information-sharing mechanisms in the patient-clinician interaction, we isolated a design question and created a series of personas and scenarios to better illustrate the users, uses, and implications of the proposed solution. Our portal solution, SamePage, includes avenues for detailed information about conditions and surgical options, utilizing diagrams and information tailored to the patient's specific anatomy and diagnosis, shared question lists, and venues for continuous communication with healthcare professionals. Next steps include iterative user testing and evaluation.

Introduction
Health Information Technology (Health IT) is changing how patients and clinicians communicate, exchange information, and interact with health data and each other. Innovations in biomedicine, computer science/information technology, and human computer interaction research, as well as federal initiatives continue to motivate the use of technology in healthcare. Use of Health IT has largely been seen inside the walls of the hospital in the form of electronic prescribing and laboratory systems, and the use of the electronic medical record in clinical practice and billing activities. With the implementation of Meaningful Use, Health IT is expanding beyond hospital walls to provide access to health information and services to patients at home and outside of the clinical environment. Patient portals allow patients to view and interact with their health data, and carry out routine administrative tasks such as requesting appointments, refilling prescriptions and sending secure messages. Despite their potential to provide the patient with access to information and services, they are often underutilized[1]. Getting patients enrolled is a significant challenge, but once achieved, repeated use is becoming more common[1, 2]. Barriers to enrollment and long-term use include lack of information or motivation, distrust and negative perceptions of the portal technology and its usefulness, concerns about timely responses, and prior negative experiences, especially in regards to the secure messaging functionality, and to a much lesser extent 'connectivity obstacles' [2, 3]. The nature of these reasons suggest that the challenge is social rather than technical, and that the ways in which they are introduced and used in clinical interactions is important to uptake and long-term use. While patient portals have great potential for patient-provider communication and transfer of information, current underutilization exposes an opportunity for redesign, both of the portal technology, and the clinical workflow in which it is used. By taking a new approach to these portals, we create an environment supportive of engaged and informed communication and decision-making for patients and their healthcare providers.

Background and Challenge
For this design challenge, we have chosen to focus on improving communication between patients and their healthcare providers to support patients as equal, informed partners in the decision-making process. We accomplish this by stepping beyond the features typically embedded in these portals and providing patients with resources to support elevated participation. Our proposed solution is designed for women who are candidates for gynecological and urological surgeries and need assistance in understanding and making decisions about their surgical options. We began our design process with a persona and scenario based on a real-world encounter. We then identified problems and challenges associated with the encounter, and developed an updated scenario and persona to illustrate the proposed solution and its role in the patient encounter.

Brief Scenario: Our interest in this challenge arose from a real-world encounter experienced by a close family friend of one of the design team members. The scenario, presented in depth in the supplemental materials, involved an encounter with a gynecologic surgeon for a diagnosis and consultation for a pelvic organ (bladder) prolapse repair. The patient, Karen, was presented with several surgical options during the appointment, and was then sent on her

way with an informational pamphlet and a web address. Karen left the encounter feeling dissatisfied and confused; she felt as though the options presented did not reflect her understanding of the diagnosis. Karen and her doctor were not seeing eye-to-eye, especially on the issues of how urgent the procedure was and whether or not a hysterectomy would be necessary. Wanting to start a family soon, this was unacceptable. The resources provided by the surgeon left her with many unanswered questions. She tried searching online on her own, but was unsure of the details of her diagnosis and could only find a cryptic, one line description in her online portal. She considered contacting the surgeon again, but felt that without having a better understanding going into the appointment, the unsatisfactory encounter would likely repeat itself. Feeling dismayed, she put surgery on hold until she felt that she and her doctor could better communicate and come to a mutually acceptable decision.

Pelvic organ (bladder, rectum, or uterus) prolapse is a common condition that occurs in women as a result of weakening of the muscles and ligaments supporting the pelvic organs. Typical symptoms include incontinence, bulging or pressure sensation, painful intercourse, and spotting or bleeding[4, 5]. Over 200,000 women undergo surgery to repair the condition each year, with as many as 30% of patients experiencing recurrence after the initial surgical repair[5]. Age and childbirth increase the risk of developing a prolapse[5]. Due to trends in aging, it is estimated that by 2050, the prevalence of pelvic organ prolapse in US women will reach 4.9 million[6]. The actual prevalence of pelvic organ prolapse remains unclear, as definitions and classifications vary, and it is unknown how many women do not seek medical attention[5]. It is likely that many additional women would benefit from treatment, but due to reasons such as stigma, embarrassment, or lack of knowledge and resources, they do not seek treatment and instead opt to live with potentially painful or embarrassing symptoms. As illustrated in the scenario above, decisions surrounding these types of surgeries are not always simple and straight-forward. Many procedures are irreversible and can have lasting effects on overall health, including childbearing abilities. It is essential that the patient understands her options, and is able to work with her surgeon to reach a mutually acceptable plan moving forward. In order to reach this point, the patient must understand her diagnosis and treatment options, and must be able to trust and communicate effectively with her clinician. This is often difficult to achieve in a single consultation as patients may not know the extent of their diagnosis and treatment options prior to their appointment. Moving communications outside of the clinical appointment adds an additional challenge to the information sharing and decision making process. Patient portals have the potential to facilitate ongoing communication between patients like Karen and clinicians outside of the clinic by presenting patients like Karen with resources and information needed to make informed decisions.

DESCRIPTION OF THE PROPOSED SOLUTION

Reflecting on the experience presented in this scenario, it is clear that the patient did not feel as though she knew enough about her options to be able to make an informed decision. She faced challenges in understanding the information presented, accessing applicable information, and in interpreting her diagnosis and treatment options. Challenges in communication were also identified as barriers to moving forward with the decision making and treatment process. In an attempt to resolve some of these challenges and reduce barriers between the patient and her clinician, we developed the SamePage Patient Portal Solution, as illustrated in the form of an updated scenario.

Updated scenario: Elaine is a 56 year old secretary living with her husband of 36 years and two daughters in Los Angeles. After experiencing increasingly frequent urinary incontinence and pain during sexual intercourse over the past few years, Elaine decided to see her doctor for a referral to a specialist. Before Elaine's appointment with the surgeon, registration staff introduced her to SamePage and made sure her account was set up so she could access a pre-appointment question list. During the appointment, Elaine and her doctor discuss treatment options and preferences. Instead of handing Elaine informational handouts and web links, the surgeon logs into the provider-side of SamePage through the EHR and hands Elaine a tablet to log into the patient-side. Elaine's diagnosis and health information populate in the portal via an HL7 interface between the EHR and the portal system. Her doctor quickly and easily adds information and personalized diagrams tailored to her individual anatomy using basic checkboxes and lists as they talk. They discuss the fact that Elaine is finished having children, and has placed a high value on pain-free sexual intercourse. These values and preferences are reflected in the personalized information included in Elaine's account. Together they review the information, using diagrams, images, and information features within the portal to better explain and understand each other. As there is much to consider, Elaine decides that she would like to review the information further before making a decision. At the end of the appointment, the surgeon dictates an exam summary which is automatically transcribed, approved, and transmitted to SamePage for Elaine to view.

At home, Elaine logs into her secure patient portal and reviews what she and her surgeon discussed in the office. Here, she reads more about her condition and what it means for her overall health. She reviews her recent vitals, medications and lab results made available by her doctors. In SamePage, she clicks through the surgical options discussed and reads an in-depth explanation of how each is performed, details of the recovery process, and potential complications. Personalized interactive diagrams and simulations are displayed alongside these explanations to help her better understand each approach. As she reads through the relevant medical information, she highlights the things she would like to remember to discuss with her surgeon. She discovers that she can also attach notes to the highlighted sentences and add them to a question list for her next appointment. The question list helps Elaine and her surgeon prepare for the appointment, and directs conversation towards what Elaine is most interested in and concerned about. She can also send secure messages ahead of time to request additional information. The Help menu explains features and functions and answers basic questions about the portal technology. Elaine also visits the Settings menu where she can grant limited or full access to a family member or health delegate.

After reviewing the information available on SamePage, Elaine requests a follow-up consultation through the portal. She has the option of an in-person appointment, or a video conference appointment carried out through the portal. Real-time video conferenced appointments take full advantage of the portal features and telemedicine technologies, allowing the patient and the provider to simultaneously view materials, and address any questions the patient may have from the privacy of their own home. Feeling confident about how well SamePage has worked so far, Elaine chooses to avoid LA traffic and try something new and schedules a video conference appointment. Elaine's surgeon allows option of audio recording the appointment and storing it for future reference, a feature Karen is impressed by. During the appointment, Elaine is able to communicate her concerns to the surgeon and feels confident that they are on the same page. Together they come up with a decision on how to proceed. At the end of the appointment, the surgeon dictates a summary which then becomes available on the SamePage and in her medical record. She then submits a request to the scheduler to be contacted to set up a surgery date and her surgeon activate the appropriate prep and follow-up documents, as well as a module for collecting patient-reported outcomes and symptom updates.

DISCUSSION OF ALTERNATIVE SOLUTIONS CONSIDERED

In the process of designing SamePage, we considered several approaches. One alternative was the use of mobile health technologies in supporting and expanding patient portal functionalities related to communication. Mobile health applications have been widely used as a means of accessing, distributing, and collecting health related information in the US and worldwide. Successful examples of mobile health technologies can be seen across the literature for activities including promoting self-management in chronic conditions including diabetes, asthma, and hypertension, and medication adherence in HIV/AIDS[10]. A second, approach considered was shifting patient portal technologies to the level of the Regional Health Information Organizations (RHIO) or Health Information Exchange (HIE). RHIOs and HIEs have been established across the country in order to allow for the exchange of health information, and to provide clinician access to patient information obtained at other connected healthcare entities. We considered the potential of using these exchanges as an access point for providing enhanced communication and information services to patients as a part of our exploration, potentially providing fewer points of central implementation and more widespread use.

DISCUSSION OF STRENGTHS AND WEAKNESSES AS COMPARED TO ALTERNATIVE SOLUTIONS

In identifying a design solution to pursue to address these challenges, we considered the strengths and weaknesses of each potential idea before selecting the option that we believe best supports the patients we are designing for, as well as the clinicians and healthcare organizations where we imagine this solution being implemented.

Comparative strengths: Our proposed solution is ideal for patients needing additional time and information before making a major medical decision. Patients have access to relevant health information when they want it and can learn on their own time, and in the privacy of their homes. Our solution allows patients to interact with personalized health information and carry out consultations remotely using video communications or in person. This provides flexibility for patients who are unable to make repeat office visits due to distance or scheduling constraints, serving to reduce a major barrier to access. It also allows patients to interact with personalized information and diagrams to assist them in understanding of their own diagnosis. Question lists created beforehand allow the patient and clinician to prepare for and guide discussion throughout the appointment. By using the portal solution in office as well as at home, we also increase enrollment and utilization of these technologies. In-clinic interactions help the patient to see the value and benefits of the portal, and understand the ways in which each person will interact with the system.

By using the existing infrastructure and extending traditional portal features, we reduce barriers to entry for patients, clinicians, and organizations. Patients and clinicians will be faced with fewer changes, and organizations will not be forced to make significant investments of time, money, and other resources to implement technologies duplicating those already in place. A major strength of SamePage is that once the initial set-up is complete, creating individualized patient material is very simple and efficient. SamePage features and functionalities not only support the patient at home but provide an interactive tool for us in the clinic as well. The novelty in our approach involves restructuring expectations of the technology, and changing how it is used in the clinical environment and at home to better serve communication and information needs of patients.

Mobile health: While mobile technology has provided unprecedented access to health information and new methods of patient-provider interaction, it also raises privacy concerns. The sensitive nature of the condition, volume of information, and personalized diagrams would likely be inappropriate for accessing in public locations or environments where a mobile device is more convenient than a PC. Additionally, the information would ideally be presented on a large screen so patients could simultaneously interact with information and images displayed.

RHIO/HIE: The advantage of working with an HIE/RHIO is that data collected by various organizations is available at a centralized access point, meaning fewer log-ins and systems to learn. However, developing even simple access functionality is difficult, and requires compromise and narrowing of scope, as experienced and reported by Ancker et al[7]. The many technical, logistical, political and financial challenges involved in this approach make it an impractical option for our focus of improving communication mechanisms and providing information to engage and inform patients. The use of standards and the demands of transforming and translating information at the interface level would need to be addressed, as well as concerns about privacy and data ownership. Moving portal activities away from the individual healthcare organization adds additional steps to the clinician workflow and introduces new potential failure points. Many healthcare organizations are still struggling to connect with HIEs, or do not yet have exchanges to connect to, limiting the ability of organizations and patients to access these features. Another challenge is that these systems are built in different ways; some exchanges are centralized, storing all data in a single location where it can be accessed, while others are federated, remaining in their home institution's database and provided to the requesting user through a call to that database. This poses many challenges developing portal applications to access this information, and questions to feasibility of such functionality in the future.

Comparative weaknesses: Our portal has several acknowledged limitations. The current solution only addresses a single group of conditions for a specific patient population. Even with increasing numbers of patients being affected by pelvic organ prolapse, it is unclear whether this will lead to widespread change in how portal technologies are developed and incorporated into the clinical workflow and decision making process. However, we believe that portal solution could easily be adapted for other conditions and scenarios where treatments options require careful consideration and has heavy implications such as joint replacement and bariatric surgeries. Another limitation of our solution is that it does not inherently support long-term use. While repeat surgery for corrections/failures is not a rare event, repeat use of the main consultation feature is not assumed. Instead, we approach this as an initiating event that will serve to build trust in the technology, and will lead to long-term use of other included and potential future features. Because we are proposing incorporating SamePage into the existing portal infrastructure, it would be reasonable to believe that existing functionality would still exist, and that it would be relatively straight forward to develop and include other modules for self-management activities in the future.

Plan for Dissemination and Implementation

Prior to dissemination and implementation, additional evaluation and user testing will be necessary to ensure usability issues are resolved and that the needs of the patient and provider are met. This is especially important as SamePage has thus far been designed around the needs and experience of a single subject, which poses a large risk towards generalizability. We advise further iterative design and evaluation involving all stakeholder groups.

We created this solution with the idea of creating as few barriers to dissemination and implementation as possible. Acknowledging that resources have already been put into developing and implementing EHR technologies and the low likelihood of uptake for a system that would replace one already in production, we created SamePage to be integrated within existing infrastructures. This allows current portal features to remain in place, and permits single sign-in access to these features alongside those we have proposed. Although the portal is re-designed and the workflow restructured, transition to the new use case should be straightforward. The best approach to dissemination involves partnering with existing EHR vendors to incorporate the new functionality and features into their build, and

providing them to healthcare organizations as a part of a planned upgrade. This would allow time for testing and training on new features in advance, and for adapting and adjusting to the needs, values, and workflows for clinicians. By making SamePage available through popular vendor systems, wide spread dissemination will be more rapidly achieved. Our plan will take into account factors from Roger's Diffusion of Innovation[8]. Organizations, clinicians, and patients need to believe that the proposed solution is an improvement on existing technologies, see how it will benefit them, and understand any associated consequences. The portal must align with user and stakeholder beliefs, values and perceived needs, and must not be overly complex or inflexible to adaptations. For the organization as a whole, trialability and observability can be addressed by rolling out the portal features to a smaller group at first and allowing other clinicians to test and train on the new features as they wish, receiving regular updates from clinicians who are currently using the system as a part of their clinical workflow[8]. This works well for our solution, as it would initially be implemented with a small group of specialized surgeons to start who will interact with other potential future users and stakeholders including primary care and referring providers.

EVALUTION PLAN

Prior to implementation, several evaluations will be carried out, with future iterations to follow throughout the implementation and usage periods including an expert evaluation of the patient and provider interfaces based on the criteria set forth in Neilsen's heuristics[9]. Once implemented, we are interested in how the portal experience affects the clinic visit. We will evaluate the technology and any corresponding interactions to determine whether these interactions are optimized by the intervention, and whether patients feel engaged as informed decision makers after interacting the clinician and the portal system. This can be investigated through surveys and interviews with patients and clinicians, comparing their comfort and understanding of their diagnosis and options following the initial consultation and again following the second meeting. Qualitative approaches to ascertaining such information can serve to inform design evolution, and determine whether communication and engagement goals are met. We will also analyze is the impact of providing interactive, comprehensive patient-specific information through the portal and will look at quantitative data such as the number of questions asked during appointments to determine whether this is influenced by the implementation of SamePage. The ultimate goal is to have patients feel informed and confident in their ability to communicate and make decisions alongside their healthcare providers. Evaluation and iterative improvement will continue throughout the life of SamePage to continuously ensure that this goal is met.

References

1. Neuner J, Fedders M, Caravella M, Bradford L, Schapira M. Meaningful Use and the patient portal patient enrollment, use, and satisfaction with patient portals at a later-adopting center. American Journal of Medical Quality. 2014 Feb 21.
2. Goel MS, Brown TL, Williams A, Cooper AJ, Hasnain-Wynia R, Baker DW. Focus on personal health records: Patient reported barriers to enrolling in a patient portal. J Am Med Inform Assoc. 2011;18: i8-i12.
3. Wade-Vuturo AE, Mayberry LS, Osborn CY. Secure messaging and diabetes management: experiences and perspectives of patient portal users. J Am Med Inform Assoc. 2013 May-Jun; 20(3): 519-525.
4. Pelvic organ prolapse – symptoms [Internet]. 2012 Oct. Accessed 2015 July 28. Available: http://www.webmd.com/urinary-incontinence-oab/tc/pelvic-organ-prolapse-symptoms
5. Rogers RG, Fashokun TB. An overview of the epidemiology, risk factors, clinical manifestations, and management of pelvic organ prolapse [Internet]. 2014 May 14. Accessed 2014 July 24. Available: http://www.uptodate.com/contents/an-overview-of-the-epidemiology-risk-factors-clinical-manifestations-and-management-of-pelvic-organ-prolapse-in-women
6. Wu JM, Hundley AF, Fulton RG, Myers ER. Forecasting the prevalence of pelvic floor disorders in U.S. Women: 2010 to 2050. Obstetrics & Gynecology. 2009 Dec;114(6):1278–83.
7. Ancker JS, Miller MC, Patel V, Kaushal R, with the HITEC Investigators. Sociotechnical challenges to developing for patient access to health information exchange data. J Am Med Inform **Assoc. 2014;21:664-670**
8. Berwick DM. Dissemination innovations in health care. J Amer Medical Association. 2003; 289(15): 1969-1975.
9. Affairs AS for P. Heuristic evaluations and expert reviews [Internet]. 2013 [cited 2014 July 21]. Available: http://www.usability.gov/how-to-and-tools/methods/heuristic-evaluation.html
10. Klasnja P, Pratt W. Healthcare in the pocket: mapping the space of mobile-phone health interventions. Journal of Biomedical Informatics. 2012 Feb; 45(1): 184-198.
11. Kallander K, Tibenderana JK, Akogheneta OJ, et al. Mobile health (mHealth) approaches and lesion for increased performance and retention of community health workers in low- and middle-income countries: a review. Journal of Medial Internet Research. 2013 Jan; 15 (1): e17.

Sintesi: Making Health Information Meaningful

Silis Y. Jiang[1], Jennifer E. Prey, MA, MS[1], Jamie S. Hirsch, MD, MSB[1], Andy Chiang[1]
[1]Columbia University, New York, NY

Abstract

Our project, Sintesi, which is Italian for synthesis, aims to bring patient data together in a meaningful, personalized way. Beyond being a patient portal with only basic data, sintesi will bring customized visualizations to users based on their health conditions and literacy levels. We believe that to fully engage patients, a 'one size fits all' solution is not adequate. It is only with specific tailoring that we can make information become more than just words and numbers on a page.

Introduction

Almost every person in the world will, at one point, be considered a patient. For many of us, we were literally born into it. We are present at every doctor's appointment, procedure, and episode, making us the "sole subject matter expert"[1] on ourselves. Therefore, it is important to leverage the knowledge we, as patients, have and to be included in managing our health and making healthcare decisions. Historically, healthcare has been a particularly paternalistic environment, where patients simply followed the directions their clinicians gave them. More recently, with the 'blockbuster drug' of patient engagement, we are seeing the tide begin to turn, and patients becoming involved.[2] This involvement creates an "increased burden on patients to understand health-related information to make fully informed choices about their medical care."[3]

While the need to put consumers 'in charge' of their own healthcare is beginning to be acknowledged, how to do so remains uncertain. The current challenge in information delivery is that data are rarely contained in a single system, making it difficult for patients to consolidate their health information, and severely constraining effective continuity of care. Additionally, information provided to patients is often not formatted in a 'patient-friendly' manner, but rather is simply repurposed from electronic health records without patient-centered annotation or formatting. Meaningful use requirements have encouraged the proliferation of patient portals, but the legislation is very vague on the details, including requisite data and format. Many of today's patient portals lack information which is actually useful to patients. For over 15 years, studies have shown that patients value custom-tailored information specific to their personal situation and data.[4] Even with today's technological advances, this type of tailored information is still very difficult to find.

The value of providing useful health information to improve health communication and engagement is clear. Engaged patients tend to have higher levels of patient satisfaction [4–7] as well as increased understanding of their care and motivation to adhere to the treatment plan.[4,5,7] Studies have shown that provision of information can have positive outcomes. For example, three-quarters of the OpenNotes project study participants reported taking better care of themselves as a result of having more information.[8]

Our goals were to explore what information patients would like to see, and how they should see it, and then to design a system which could facilitate patients' understanding of their health and improve communication with their care providers.

Design Methodology

To inform the design of our system, this project was rooted in a user-centered design model. We conducted ten semi-structured interviews with English-speaking adults. The participants were sourced from the research team's personal relationships with friends and family members. Conversations were audio-recorded and notes from the interviews were aggregated, synthesized, and discussed with team members. Participants ranged in age from 24 to 93, with a median age of 31 (average of 44). Four males and six females were included. We also conducted a brief review of the literature to identify the needs of patient populations which we did not address in our interview demographics (e.g. low literacy and low income populations).

Based upon participant interviews and the literature review, we identified user needs and used them as the basis for an iterative design method. A design solution was conceived, mocked up, reviewed, and then iterated based on team feedback.

Proposed Solution

Based on the results of our interviews as well as a review of the literature, we believe that to facilitate patient engagement and patient-clinician communication, it is very important to provide patients with access to their own data. Patients consistently commented that gaining access to information was difficult, and often required multiple phone calls to follow-up. They also commented that finding and understanding appropriate information was difficult. Therefore, we believe it is important that information be presented in a way which is appropriate to each particular patient. This is consistent with findings in the literature; Alpay, et. al. commented that it has "become a prime importance to develop Information and Communication Technology (ICT) based tools that can provide tailored information."[9]

Tang et al. identified key basic attributes for sharing information with patients including that it is "concise, clear, and illustrated with graphics if appropriate."[4] They found patients wanted a permanent record which they could keep and refer to after a visit so they could access the information later, especially when talking about their care with family and friends. Additionally, with the varying levels of literacy in patient populations, we believe that data visualization customization that is patient-specific, and appropriate, is key to increased comprehension.[10,11]

Sintesi is a customized patient portal in which the presentation of information to the patient is dependent upon his/her characteristics, namely, the patient's level of health literacy and health conditions. These criteria will help determine which data should be most easily available to the patient, and in what format the data should be visualized. The full set of mockups of *Sintesi* is available in Appendix A of the supplementary materials.

A key feature to creating useful visualizations of data involves having a robust data source. We believe having a central repository where patients are able to aggregate their data from multiple sources is important to allow for greater continuity of care and for end-to-end data analysis. We envision a model similar to the online personal finance software, Mint.[12] Mint is an aggregator of financial information. The user provides his/her login information to each of his/her financial institutions, and then Mint is able to pull account information from each of these sources. All of the user's data are then in one place and are more easily analyzed for trends and budgeting. While others have failed at being a health data aggregator (e.g. Google Health[13]), we believe the healthcare technology environment has changed to make this idea more feasible. For example, one industry start-up, MDCapsule is following this model.[14] Additionally, with the expansion of projects like the Veteran Administration's Blue Button, interoperability is slowly becoming a reality.

In *Sintesi*, upon first login, literacy will be evaluated through a two-part, 11-question survey. The first part will assess health literacy using the three-question Brief Health Literacy Screen.[15,16] The second part will assess numeracy using the Subjective Numeracy Scale (SNS).[17] The SNS tests numeracy level without having to ask actual math questions. It can be conducted in less time and with fewer burdens on the participants than traditional numeracy tests. The results of these questions will be used to place each patient into one of three health literacy levels: Basic, Intermediate, or Proficient, as adapted from the National Assessment of Health Literacy levels.[18] Each of these levels will have different representations of information. Design of the graphics themselves was informed by the work of the Washington Heights/Inwood Informatics Infrastructure for Community-Centered Comparative Effectiveness Research (WICER) patient visualization project[19] and the HARVEST project[20] which developed visualizations for clinicians. Design of the medication display was informed by Kripalani et al.[21] While patients will be automatically assigned to a particular level of health literacy, and thereby shown visualizations from a certain group, s/he may switch to the other forms of visualizations as desired.

The patient's health conditions will be obtained from the patient's problem list; if there is more than one item in the problem list, multiple visualizations will be shown. The varying metrics which are important per condition type will be retrieved from a reference list which is currently in development (Appendix B) and will default to general values of weight and blood pressure if a disease or condition-specific variable is not available.

Additionally, while we believe this portal will primarily be accessed by patients while they are away from the doctor's office, we believe it could be a useful tool to facilitate communication during the doctor's visit itself. A new workflow could be one in which the patient is able to login to the portal while sitting in the doctor's office. The doctor and patient could discuss which metric(s) are important for the patient to keep track of, and look at the trend of those measurements together. They could then decide on a treatment plan, and even add a goal on to the portal. This way, there is a visual representation to facilitate communication and the patient will have a record of what they are working towards.

Additional features to be included on the portal are facilitating secure messaging with providers and assistance in understanding clinician notes. Secure messaging is a feature which is part of Meaningful Use Stage 2 and as such, most institutional portals will be able to do this. Our portal will display all members who were involved in the patient's care, and allow the patient to send them a direct message or obtain their phone number. As more and more information is being shared with patients, including notes, it can be difficult for patients to understand the 'doctor-speak' in which the notes are written. A fairly simplistic way to help make the notes more patient-friendly is to partially 'translate' these notes. This could involve mapping acronyms to spelled-out words and including a feature like InfoButton to help explain terms which might not be familiar to the patient.[22]

In order to increase utilization and gather more patient-generated data, there will be a mobile component to the website (not shown in mockups). Using a mobile application or text messaging (as decided by the user), the system will occasionally check-in with the patient. Use of text messaging to engage with patients is a growing trend and is especially effective for low-income populations.[23] Trigger events for these check-ins could include when a patient begins a new medication, the mobile system could prompt the patient to provide how they are feeling and ask if they are having any new side-effects. A mobile application could also facilitate collection of patient-generated data from personal devices (e.g. fitness trackers or glucose monitors). Patient-generated data are becoming more common and it is suggested that these data can useful to physicians.[24,25]

Alternative Solutions

Our team considered a variety of solutions to address how to increase patient engagement and patient-clinician communication. Some ideas were completely outside of a portal and others were varying design characteristics of the portal which we settled on.

One alternative we considered was to simply provide a better system for direct phone contact between patients and providers. Some of our older interview participants described their desire to be in greater phone contact with their providers. As the older generations are less likely to be computer-savvy, we believe that our enhanced portal solution is unlikely to provide much benefit to them. This alternative would provide a direct line to a physician for patients who are above a certain age or describe themselves as non-computer users. By providing this increased communication access, patients could have their questions answered more immediately and it is possible that they would more appropriately come into the office. In order for a system like this to be feasible, a reimbursement structure for this interaction would need to be put in place. This type of solution could also potentially put a heavy burden on clinicians and overwhelm their already busy schedules. We chose not to go with a phone-only solution as we believe that by providing patients with more direct access to information, they will be able to more fully understand their health situation and be more engaged in their care, rather than primarily relying on responses from clinicians. Additionally, as the population ages, more and more elderly patients will be proficient with electronic devices.

In our research we found many new healthcare startups that are designing mobile-only solutions. While we believe that having a mobile component is important to increase interaction and data collection, we decided that with data as complex and vast as healthcare data, it really necessitates online, web browser use and the ability to see more than can be displayed on a 3.5" screen.

An important design choice, about which we had lengthy discussions, was whether to create a solution which was tethered or untethered to a particular personal health record (PHR). This debate has existed as long as patient portals have. While some studies suggest that a tethered solution seems to be the only successful option,[26] we believe with the changing healthcare environment and increased interoperability, the idea of a central repository which is patient-controlled, and institution-agnostic, can be a reality. If we had chosen to go the tethered route, another approach to integrate our design on a portal-by-portal basis, could have been to develop a plug-in which would sit on top of each institution's PHR and would be able to interact with the data at that institution directly. While this could have allowed for easier access to individual sites' data, we believe there is significant value in cross-institution information visualization and thus went with the stand-alone design option.

One additional solution we considered, but did not implement, was to focus on facilitating decision-making for patients. The ability for patients to convey priorities to their physicians when discussing treatment options is crucial, and there are currently few ways to help clinicians obtain these priorities from patients. By using tools such as dynamic decision trees, which could take patient preferences (e.g. quality of life, pill burden, etc.) into consideration, recommended treatment options could be updated so patients receive the best treatment for them. We hope this could be a feature in a future implementation of our portal.

We believe the solution we have chosen is one which addresses the important problem of providing patients with their information in ways which are comprehensive, understandable, and can facilitate communication with their providers. We also believe that this framework allows for future features to be integrated. Additional improvements we believe should be considered are working to further optimize the verbiage and visualizations based on patients' cultures and additional demographics, and integrating other common portal services (e.g. appointment scheduling). Overall, we believe this solution to provide valuable information and be flexible enough to allow for future growth.

Implementation and Dissemination

To further develop this solution into something which is fully functional and available, more work would need to be conducted to iterate the design. We would recommend conducting usability tests with various types of users including those from all three literacy levels and with different health conditions. Feedback from the usability tests should be integrated into new designs until the product is ready for deployment. Frameworks from work already accomplished by others such as the EnTICE software design platform could be leveraged during development.[27] Once recommendations from the usability tests are integrated, it's time for a staged roll-out in which certain patients are targeted to sign-up and use the system prior to a full go-live. Any feedback from early adopters can then be incorporated into the design. Once these updates have been made, we recommend doing a full release.

Gaining users for the system could be initially difficult. Many people have portals they are currently using and would therefore need to see the benefit of a new portal in order to switch. We believe that by marketing our solution as a 'Mint.com' for healthcare, it could entice users to take the time to join a new portal, input their credentials to link to other existing portals, and find the enhanced capabilities we provide. The addition of 'gamification' features and social media in the future could also encourage users to return to the site repeatedly to provide additional data points to supplement those provided by their healthcare providers. Challenges to implementation will include interfacing with the institutional portals and guaranteeing security of data. Additionally, a data-sharing model needs to be developed, particularly concerning appropriate access for parents of children and adolescents.

Evaluation

Multiple rounds of evaluation should be completed on this project. As discussed in the implementation plan, usability testing should be completed prior to rolling out the system. After the system is operational, we would recommend evaluating the portal through comprehension testing to ensure the information is being presented in an understandable and satisfying way. Thereafter, we would like to measure changes in engagement based on portal usage. This could be done either through finding matched controls to people who are already participating in portal use, or by finding a sample to randomize to having access to the portal or not. After a certain amount of time, perhaps a year, we could then analyze changes in patient activation using the Patient Activation Measure[28] and analyze other data such as the amount of health services used, the number of communications conducted with their care providers, and changes in health status.

Conclusion

In today's culture, with "Have it your way" slogans at places like Burger King, the idea of 'one size fits all' no longer applies. Especially in an arena as important as healthcare, information should be shared in a way which is customized for the person viewing it. Provision of information allows for more sophisticated communication and expression of personal preferences and desires. It is only with real understanding by both the patient and the provider in which we will be able to really achieve improved health.

References

1. Goetz T. The Decision Tree: How to Make Better Choices and Take Control of Your Health. Rodale; 2011. 339 p.
2. Dentzer S. Rx For The "Blockbuster Drug" Of Patient Engagement. Health Aff (Millwood). 2013 Feb 4;32(2):202–202.
3. Reyna VF, Nelson WL, Han PK, Dieckmann NF. How numeracy influences risk comprehension and medical decision making. Psychol Bull. 2009;135(6):943–73.
4. Tang PC, Newcomb C. Informing Patients: A Guide for Providing Patient Health Information. J Am Med Inform Assoc. 1998 Nov 1;5(6):563–70.
5. Tang PC, Lansky D. The Missing Link: Bridging The Patient-Provider Health Information Gap. Health Aff (Millwood). 2005 Sep 1;24(5):1290–5.

6. Verlinde E, De Laender N, De Maesschalck S, Deveugele M, Willems S. The social gradient in doctor-patient communication. Int J Equity Health. 2012;11(1):12.
7. Street RL, Millay B. Analyzing Patient Participation in Medical Encounters. J Health Commun. 2001 Jan;13(1):61–73.
8. Trossman S. OpenNotes initiative aims to improve patient-clinician communication, care. Am Nurse. 2013 Oct;45(5):10.
9. Alpay, Laurence, Verhoef, John, Xie, Bo, Te'eni, Dov, Zwetsloot-Schonk, J.H.M. Current Challenge in Consumer Health Informatics: Bridging the Gap between Access to Information and Information Understanding. Biomed Inform Insights. 2009 Jan 1;2(1):1–10.
10. Eichner J, Dullabh P. Accessible Health Information Technology (IT) for Populations with Limited Literacy: A Guide for Developers and Purchasers of Health IT [Internet]. Rockville, MD: Agency for Healthcare Research and Quality: National Opinion Research Center for the National Resource Center for Health IT; 2007 Oct. Report No.: No. 08-0010-EF. Available from: http://healthit.ahrq.gov/sites/default/files/docs/page/LiteracyGuide_0.pdf
11. Galesic M, Gigerenzer G, Straubinger N. Natural Frequencies Help Older Adults and People With Low Numeracy to Evaluate Medical Screening Tests. Med Decis Making [Internet]. 2009 Jan 6 [cited 2014 Jul 22]; Available from: http://mdm.sagepub.com/content/early/2009/01/06/0272989X08329463
12. What is Mint – Your Financial Life All In One Place | Mint.com [Internet]. [cited 2014 Jul 31]. Available from: https://www.mint.com/what-is-mint/
13. Google Health Discontinued [Internet]. 2011 [cited 2013 Oct 1]. Available from: http://www.google.com/intl/en_us/health/about/
14. If Google Health Failed, Why Will Your Health Portal Company Succeed? [Internet]. Forbes. 2014 [cited 2014 Jun 22]. Available from: http://www.forbes.com/sites/robertszczerba/2014/06/12/if-google-health-failed-why-will-your-health-portal-company-succeed/
15. Chew L, Bradley, KA, Boyko, EJ. Brief questions to identify patients with inadequate health literacy. Health Lond Engl 1997. 2004;11:12.
16. Wallston KA, Cawthon C, McNaughton CD, Rothman RL, Osborn CY, Kripalani S. Psychometric Properties of the Brief Health Literacy Screen in Clinical Practice. J Gen Intern Med. 2014 Jan 1;29(1):119–26.
17. Fagerlin A, Zikmund-Fisher BJ, Ubel PA, Jankovic A, Derry HA, Smith DM. Measuring Numeracy without a Math Test: Development of the Subjective Numeracy Scale. Med Decis Making. 2007 Sep 1;27(5):672–80.
18. Kutner M, Greenburg E, Jin Y, Paulsen C. The Health Literacy of America's Adults: Results from the 2003 National Assessment of Adult Literacy. NCES 2006-483. Natl Cent Educ Stat. 2006;
19. Arcia A, Bales ME, Brown W, Co MC, Gilmore M, Lee YJ, et al. Method for the Development of Data Visualizations for Community Members with Varying Levels of Health Literacy. AMIA Annu Symp Proc. 2013 Nov 16;2013:51–60.
20. Hirsch J. A Task-Based Evaluation of HARVEST, a Clinical Data Aggregation and Summarization Tool. Poster presented at: National Library of Medicine Informatics Training Conference; 2014 Jun; Pittsburgh, PA.
21. Kripalani S, Robertson R, Love-Ghaffari MH, Henderson LE, Praska J, Strawder A, et al. Development of an illustrated medication schedule as a low-literacy patient education tool. Patient Educ Couns. 2007 Jun;66(3):368–77.
22. Cimino J, Del Fiol G. Infobuttons and point of care access to knowledge. Clin Decis Support Road Ahead. 2007;345–72.
23. Househ M. The role of short messaging service in supporting the delivery of healthcare: An umbrella systematic review. Health Informatics J. 2014 Jul 18;1460458214540908.
24. Shapiro M, Johnston D, Wald J, Mon D. Patient-Generated Health Data. White Pap Prep Off Policy Plan Off Natl Coord Health Inf Technol Res Triangle Park NC RTI Int [Internet]. 2012; Available from: http://www.healthit.gov/sites/default/files/rti_pghd_whitepaper_april_2012.pdf
25. Huba N, Zhang Y. Designing Patient-Centered Personal Health Records (PHRs): Health Care Professionals' Perspective on Patient-Generated Data. J Med Syst. 2012 Dec 1;36(6):3893–905.
26. Tang PC, Lee TH. Your Doctor's Office or the Internet? Two Paths to Personal Health Records. N Engl J Med. 2009 Mar 26;360(13):1276–8.
27. Arcia A, Velez M, Bakken S. Style Guide: An Interdisciplinary Communication Tool to Support the Process of Generating Tailored Infographics From Electronic Health Data Using EnTICE3. Stakehold Symp [Internet]. 2014 Jun 7; Available from: http://repository.academyhealth.org/symposia/june2014/panels/6
28. Greene J, Hibbard JH. Why does patient activation matter? An examination of the relationships between patient activation and health-related outcomes. J Gen Intern Med. 2012;27(5):520–6.

Beyond Patient Portals: Engaging Patients with their Healthcare Providers: "HealthUp": An Active Patient Portal Beyond Sickness

Maher Khelifi, Pharm.D, Ahmad Aljadaan Ms, Micheal Anando Seng, BA
University of Washington, Seattle, WA

I. Abstract

A challenge in the effort to engage patient's with medical providers is the low utilization of patient portals. One aspect that contributes to this is that portal tools are currently more relevant for periods when the patient is in ill health. In this paper we outline a design for a portal that encourages patient engagement with the portal in periods of good health through active patient participation - reporting health data and challenge identification. To achieve patient participation our design includes features of gamification and interaction of the tool with those of patients' families and friends.

II. Introduction

Patient portals are online digital healthcare tools that provide patients with secure access to their health information and allow protected methods for communication and information sharing.(1) These portals are intended to extend the effect of health care outside of medical visits and health incidences by providing the appropriate tools to the patient for health management.(2) *Meaningful use terms* have helped to raise the number of health settings provided by such tools. Unfortunately, in spite of their potential benefits, some studies have shown that patient portals have generally low enrollment rates(3)(4).

III. Specific challenge defined

According to the Health National Statistics Report(5), patients visit physicians' offices an average of four times per year in USA. Outside these visits the patient manages their health by themselves. Regrettably, adherence to physicians' recommendations involving lifestyle changes, such as exercise, frequently pose significant difficulties for patients. Outside medical visits and when health has improved there is a tendency for patients to grow less compliant with recommended lifestyle changes. Our challenge is to transform the patient portal by extending its role, of providing health information and communication tools, to a platform that the patient can use to identify challenges and manage his health beyond sickness conditions. We aim to make the portal more interactive with the user and his environment to expand the adoption of these portals and to ensure a successful patient engagement. To achieve this, we need the portal to:

- Help the patient manage relevant information about his health conditions.
- Personalize health challenges based on patient's profile and health information.
- Create an engagement process to optimize the utilization of the patient platform.

Our goal is to create portals for patients to use in both ill and good health conditions by integrating the portals into patients life. The fact that the patient portal currently only provides reports from the patient's period of ill health, limits the use of the portal to sickness conditions. Patient portals should have use also when the patient is feeling better, to increase the chances of having an engaged patient who is equipped to manage his own health in the case of sickness. We hope to change patients behavior to encourage him check his health information more often, make healthier lifestyle choices, and manage his health condition. Our idea is encapsulated in the following quote: "The patient's role is changing from a patronized patient to an informed patient and further to a responsible, autonomous and competent partner in his or her own care".(6)

IV. Description of the proposed solution and design process:

1. The proposed solution:

We propose, as a solution; a patient portal accessible from a website and phone application.
This portal will help the patient to a) manage his health information, b) personalize his challenges, and c) engage with gamification platform.

a) Manage health information

The platform will provide the patient's medical records and proffer a space for collecting real-time health information. It will additionally display this information based on patient priorities and health challenges.

i. **Collect Health information**:

The health information will not be only entered by the provider, but also by the patient. The user can add information to his profile but it will not be saved in the medical record in order to not detract from its quality. Information captured outside of the medical visit will be saved in a different section.

Information sources are:

- Health providers: health providers can enter information to the medical record. Then the medical record will be synchronized with the portal.
- Automatic reporting from technology sensors: technology used by the patient can help him update his health information. The platform will facilitate automatic reporting by devices used by the patient. Smartphones, for example, can be a very valuable aid; they can report sleep hours, heart rate, and activity. The automatic reporting will help to reduce the burden on the patient and keep the platform running.
- The patient: patient can update his own information manually. This is especially useful for information not obtainable by technology.

ii. Display of health information

The platform will highlight the categories that the patient should attend to. The health provider will be able to help in identifying what health variables the patient should monitor. For these highlighted categories, the patient will be able to see the last recorded value. For example, a diabetes patient could have his glucose level highlighted and then he will be able to see the last reported value.

Having (close to) real-time information will provide patients with a clear thorough image of their health. It will aid in the objective and continual evaluation of the patient's health.

b) Personalize health challenges

After collecting health information, the application will help the patient identify and personalize his health challenges. Health challenges and tasks will be primarily identified by health providers during medical visits. Since the application is collecting lifestyle information about the patient, it can help the provider identify the barriers stopping the patient from living a healthy life. The application will propose a course of actions, or "health-plan", that the patient can follow to achieve his challenges. For example, if the patient's challenge is to manage diabetes, the health-plan will include diabetes diet, activities tasks (walk twice a week), and health monitoring (measure your blood sugar daily).

Challenges can also be updated based on new health information updated by the patient. For example if the patient's reported blood pressure is high, the application will invite the patient to add some tasks to lower his blood pressure.

c) Apply Gamification to close the patient engagement loop

After collecting health information and identifying challenges, we need to encourage the patient to use our platform. We found that gamification can be a good asset for that. According to gamification experts, when used properly, it can be deployed to achieve a wide range of desired outcomes such as fostering engagement or improving motivation.(7) Gamification is defined as "the application of game elements and theories to non-game contexts with the intention of modifying behaviors, increasing fidelity or motivating and engaging users"(8) Gamification design frameworks, found in the internet and literature, state that the application should create activity cycles to keep the stream of motivation going.(9) We used the engagement loop mentioned in gamification literature. We applied the engagement loop framework to our platform.

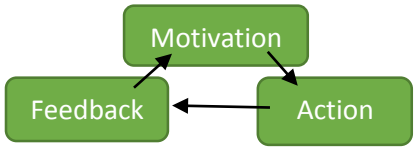


Figure1. Gamification Engagement loop(9)

i. Motivation

Self-determination theory (SDT) is one of the most well-known theories in motivational psychology. It explains human beings' innate psychological need for personal development and wellbeing and the impact on individual's motivation.(10)(9) This theory defines three innate needs: Competence (experience a sense of ability), Relatedness (social interaction, relationships, connectivity), and Autonomy (give the player the freedom to make their own choices). These needs are not affected by gender or culture, they are universal.(11) This theory delineates also the origins of motivation. According to SDT these include:

Extrinsic motivation: external stimuli such as rewards, threats, financial compensation.

Intrinsic motivation: experience being successful, social connections, satisfying work.

Extrinsic rewards such as monetary incentives are not enough to maintain behavioral changes or engagement. According to Zichermann,"The best systems of motivational design speak to intrinsic motivation of the user while providing extrinsic rewards that they value(9)(12).

Features added to the portal:

- Progress bar: The patient can evaluate the progress of his work.
- Patient network: the patient can have his family members and his selected friends connected to his portal. Users can share activities and post achievements (if they want to) .Social facilitation or the fact that people perform better in group can help motivating users.(13)(14)
- People support: High-five System. If the user share a post, his network can support him by giving him a high-five

ii. Action

After building up motivation, the user is ready to perform actions. The application will coach the user in organizing his tasks. It will help the user add tasks to his calendar, based on his life schedule. This technology will also remind the user to achieve his tasks. Instead of simple reminders, we will use geolocation and time factors to personalize alerts and reminders.

Features added to the portal:

- **Add task:** the patient can add tasks to his timeline, additionally he can invite people to participate in his tasks. Friends can join in real life (e.g. go for runs simultaneously) or they can just motivate each other virtually through the application.
- **Smart-reminders:** Smart phones can recognize when the user comes back home (medication reminders) and when he's within 5 minutes to Walgreens ("You are close to Walgreens: do you want to check your blood pressure?")

iii. Feedback

Achievement stars: the patient will receive a star every time he accomplish a task. These stars can be changed to reduction coupons for Fitness centers, healthy restaurants.

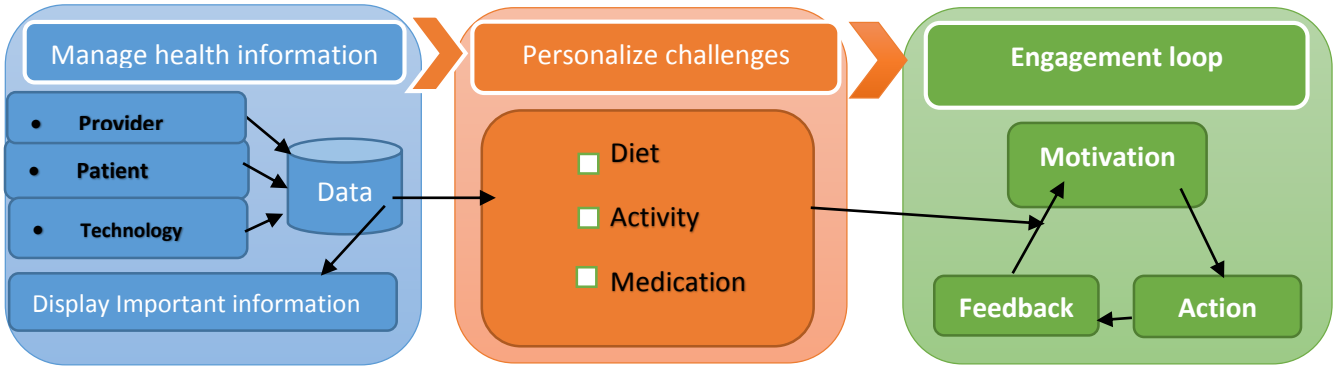


Figure2: Working process of the proposed solution

2. The process of the design:

During our design process, we had to go through different steps that can be summarized in the following list:

- Brainstorming different solutions for the three parts of the application (managing information, identifying challenges, and the gamification dynamics).
- First mockup of the platform.
- Scenarios and personas for different potential users (chronic patient, healthy user, patient network). This work helped us to optimize the features and functionalities used in the design and optimize the workflow of the portal.
- Realize a new mockup. (Described in the appendix)

V. ALTERNATIVES CONSIDERED

The research on patient engagement in health portals is dense and multidimensional due to the different problems that they currently have. We started our thinking to focus on one problem which is the existing gap between the information provided in the health portal and the patient. The information provided in the health portals are dense and speaks in medical terms that is hard for the normal patient to interpret. For that, we came up with the "Down-to-earth education". An app that can transform medical images, medical information- found in patient portals- to be more understandable by the patient. The patient can use this app to record physician notes, his own notes to improve his knowledge and skills in managing his own health.

Then, we started thinking about the problem that patient portal have as a one to one portals (patient to providers). We wanted to introduce the patient network (family, friends) as part of the patient portals. The network can assist patient in accomplishing their task or motivate them if needed. Therefore "The Angel Guardians" idea came to life. A network for patient-patient and patient-family where patient can look for help or provide help in a network.

VI. TRENGTHS AND WEAKNESSES

The proposed design focuses on 4 features that will help to increase patient engagement in the portal.

Table1: Strengths and weaknesses

Features	Strengths	Weaknesses
Dynamic Information Exchange	- Continuity of information where patient can send their real time information to the provider. - Facilitate the patient-provider meeting in which appointment times are increasingly truncated.	- Interoperability: Bridging our platform to existing Electronic Health Records (EHR) used in different hospitals - Information reported by the patient might be inaccurate. - Accessibility limited to patients with internet access and smartphones.
Personalize Challenges	- Adaptive Learning from patient-reported data to tailor tasks based on behavior. - Patients will be more motivated to complete the challenge knowing that it is designed for them personally. - Strengthens the relationship between the physician and the patient. - Develops skills and knowledge of the patient in maintaining health.	- There is potential that patients may provide inaccurate information about their behavior which will lead to ill-suited challenges. - May cause patients to overestimate their ability to manage health.
Patient Social Network	- Social facilitation: people perform better in groups(13) - Using the power of influence to galvanize behavior change(15)	- Privacy challenges. - Some tasks can be too personal to be shared with the social network.
Gamification	- Sense of control provided by the game spirit. - Follow the progress. -Motivation. - Engagement loop.	-The line between game and gamification is so delicate. There is a risk that the portal wouldn't be taken seriously. - It is important to find a balnce between cooperation and competetion in the gamification spirit. - Patient can overestimate their health condition (having a lot of reward stars doesn't necessarily mean don't return to the doctor)

VII. PROPOSED IMPLEMENTATION AND DISSEMINATION PLAN

We propose to realize our implementation and dissemination plan over 3 steps:

1. Pre-implementation:

Before starting the implementation, we need to study the existing systems used by hospitals to be sure that our application can integrate with the existing processes. It is important to understand the work flow of health providers. We intend that our solution will save time and improve communication between patients and providers. We need to identify our stakeholders and guarantee that we have them onboard to adopt this new solution. Then, we need to bridge our portal to the current EHR system that most hospitals are using (such as EPIC and Cerner).

2. Implementation and dissemination

The most two important stakeholders of our platform are: patient and health providers.

a) Health providers

It is vital to confirm that health providers are ready to adopt our solution. To optimize the adoption process we can develop learning tutorials to aid the provider in using the platform. Providers will have an adapted screen to check on their patient and to communicate with them by sending personalized or group texts. Additionally the portal should be easy to use and will hopefully reduce the burden on providers of managing patients outside of medical institutions.

b)Patients

Adoption is easier for people prepared to change. We should identify user groups that are least resistant to change and most willing to try a new tool to manage their health. The network aspect will help us to disseminate the application after recruiting the first group. We believe people with healthy life style are our target and from them we will build our network and diffuse it to others.

3. Follow-up

Provide feedback mechanism for users and providers to improve the application functionalities.

VIII. PROPOSED EVALUATION PLAN.

In order to determine that our developed tool is having the intended impact, we must have a comprehensive evaluation plan that assesses our solution. To measure the portal effectiveness we will need to focus on provider and patient satisfaction, usability testing, and portal compliances with other EHR system.

1. Provider and Patient Satisfaction.

Though we would keep our two populations (providers and patients) separate throughout the data collection period; they would essentially each be exposed to similar protocols during the evaluation process. First, a group of subjects would be asked to complete a pre-intervention survey. The survey would contain a mix of Likert-scale, True/false, and short answer questions to assess aspects of the portal relating to the following outcomes: usability, accessibility, design and usefulness. For patients, the survey would be oriented towards the gamification concept of the portal and the effect of social network has on their use of the portal.

2. Usability Testing

Approximately 15-20 subjects would participate in a focus group. The focus group would attempt to assess the needs of each profiled group of users of the portal. We would provide computer consoles on which providers and patients could use an interactive mock-up of the portal. Then we would elicit comments as they use it which examine the outcomes above.

After the pre-survey and focus groups, we would ask providers to use the portal in the actual health record for two weeks. Information would be available to us from usage logs and interactive feedback features. In addition, at the end of the two week period we would ask the providers to complete a post-survey, thus giving us pre/post study data.

3. Compliance with other EHR Systems

We will be working with teams from two of the major EHR providers (Cerner and Epic) to make sure our portal is entirely compliant with their EHR system. We chose Cerner and Epic because combined they have about 75% of the market share of the Electronic Health Record in large US hospitals. We will be conducting a conformability testing with those teams to make sure that the portal meets their EHR standards.

4. **Quantitative indicators**

We will identity a list of quality indicators like number of users, log-in times per day, interaction between network members, rewards shop. These indicators can be used to further develop and improve the portal tool.

IX. Future Direction:

The portal will be built in a rapid prototyping environment where we will be adding features and rules as we keep developing. Gamifying patient portal and adopting an online social network raises privacy concerns. We hope this portal will create rules that can protect the patient privacy while fostering a cooperative environment. To ensure sustainability of the portal we hope to develop an evolving process in our portal. One possible way that we considered is to develop ranking system, however it is important to be sensitive to the patient state of heath.

REFERENCES

1. Caroline Lubick Goldzweig, MD, MSHS; Greg Orshansky, MD; Neil M. Paige, MD, MSHS; Ali Alexander Towfigh, MD; David A. Haggstrom, MD, MAS; Isomi Miake-Lye, BA; Jessica M. Beroes, BS; and Paul G.

Shekelle, MD P. Electronic Patient Portals : Evidence on Health Outcomes , Satisfaction , Efficiency , and Attitudes. 2013;

2. What is a patient portal? | FAQs | Providers & Professionals | HealthIT.gov [Internet]. [cited 2014 Jul 15]. Available from: http://www.healthit.gov/providers-professionals/faqs/what-patient-portal

3. Goel MS, Brown TL, Williams A, Cooper AJ, Hasnain-Wynia R, Baker DW. Patient reported barriers to enrolling in a patient portal. J. Am. Med. Inform. Assoc. [Internet]. 2011 Dec [cited 2014 Jul 23];18 Suppl 1:i8–12. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3241181&tool=pmcentrez&rendertype=abstract

4. Foundation CH. Measuring the Impact of Patient Portals : What the Literature Tells Us. 2011;(May).

5. Schappert SM, Rechtsteiner E a. Ambulatory medical care utilization estimates for 2006. Natl. Health Stat. Report. [Internet]. 2008 Aug 6;(8):1–29. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18958997

6. Ammenwerth E, Schnell-Inderst P, Hoerbst A. The impact of electronic patient portals on patient care: a systematic review of controlled trials. J. Med. Internet Res. [Internet]. 2012 Jan [cited 2014 Jul 28];14(6):e162. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3510722&tool=pmcentrez&rendertype=abstract

7. Lee JJ, College T, Ph D, Hammer E, Interdisciplinary M. Gamification in Education : What , How , Why Bother ? What : Definitions and Uses. 2011;15:1–5.

8. Deterding S, Sicart M, Nacke L, O'Hara K, Dixon D. Gamification. using game-design elements in non-gaming contexts. Proc. 2011 Annu. Conf. Ext. Abstr. Hum. factors Comput. Syst. - CHI EA '11 [Internet]. New York, New York, USA: ACM Press; 2011;2425. Available from: http://portal.acm.org/citation.cfm?doid=1979742.1979575

9. Paz BM De. GAMIFICATION : A tool to improve Sustainability Efforts. 2013;

10. Deci EL, Ryan RM. Self-determination theory: A macrotheory of human motivation, development, and health. Can. Psychol. Can. 2008;49:182–5.

11. Deci EL, Ryan RM. The " What " and " Why " of Goal Pursuits : of Behavior Human Needs and the Self-Determination. Psychol. Inq. 2000;11(4):227–68.

12. Intrinsic and Extrinsic Motivation in Gamification | Gamification Co [Internet]. [cited 2014 Jul 1]. Available from: http://www.gamification.co/2011/10/27/intrinsic-and-extrinsic-motivation-in-gamification/

13. Strauss B. Social facilitation in motor tasks: a review of research and theory. Psychol. Sport Exerc. [Internet]. 2002 Jul;3(3):237–56. Available from: http://linkinghub.elsevier.com/retrieve/pii/S146902920100019X

14. Carver CS, Scheier MF. The self-attention-induced feedback loop and social facilitation. J. Exp. Soc. Psychol. [Internet]. 1981 Nov;17(6):545–68. Available from: http://linkinghub.elsevier.com/retrieve/pii/0022103181900391

15. Spatharou A. Changing patient behavior : the next frontier in healthcare value. 2012;

UHealth for Your Health: Enhancing Utilization of Patient Portals and Its Experience

Mandana Salimi, Adriana Stanley, Mehdi Rais, MD, MS, CPHIMS, Vickie Nguyen, MA, MS
School of Biomedical Informatics,
The University of Texas Health Science Center at Houston, Houston, TX

Abstract

With the increasing use of health information technologies and patient-centric focused care and involvement in healthcare practices, collaborative and cooperative systems, in the shape of patient portals, will be vital for maintaining ones health. In its current state, patient portals require an additional push to increase its use. Here we provide a potential solution to the patient portal conundrum by utilizing novel features and employing aspects of Computer Supported Cooperative Work, Persuasive Technology, and User-Centered Design in the creation of our mobile device patient portal application, UHealth. We then examine the strengths and weaknesses of our solution and present a potential plan for implementing, disseminating, and evaluating this application through the use of a scenario-based evaluation method and user satisfaction survey.

Introduction

Patient portals are a collaborative and cooperative system that aims to provide streamlined communication, collaboration, and cooperation between patients and their providers[1]. When available, patient portals have given the consumer an opportunity to engage and interact with their providers and their own data in novel ways. Such interactions have the opportunity to contribute to the well-being of the patient and the respective relationship with the medical team as it offers a myriad of abilities such as emailing questions to providers, requesting refills online, reviewing medical records, reviewing lab and imaging results, paying bills, and scheduling appointments[2]. Despite all of these benefits, many patients and clinicians are having problems with consistently using their patient portal. A 2013 study by the Cleveland Clinic, whom had 240,000 patients enrolled into their respective patient portal, found that only five percent of those patients actually used the patient portal[3].

In examining this problem of patient compliance, several reasons could account for such poor use. 1) Many portals have a modular structure with specific functionalities turned on and off, so many patients are left with experiences that lack content to interact with and subsequent cognitive involvement is voided. 2) Patients with multiple providers potentially have several portals to follow their own health. This leads to confusion in functionalities and capabilities to interact with respective providers in different systems. 3) The portal may lack the ability to download the record for a patient's own use or to share with family, friends, or other members of the care team. 4) Patient portals might lack many of the common functions a patient would seek outside of the provider's offices (*e.g.*, scheduling, prescription refills, or feedback). 5) Usability could be an afterthought in the patient portal design. 6) The portal is not available in the ways which patients are engaging data in their regular lives with mobile platforms and through many different data sources.

With so many opportunities for failure in engaging the patient portal, it is no wonder why patient adoption and use is such a challenge. To help foster the benefits of these systems, this research team is focused upon making the portal accessible to patients when they want it, where they want it, how they want this information delivered, with the right information, and in a user-friendly format. To drive this development, this team's aim will be to develop a solution for a patient portal that will be accessible from any mobile, tablet, laptop, or desktop device via respective browsers or cross-platform applications that will incorporate principles from some of the following areas: Computer Supported Cooperative Work (CSCW), Persuasive Technology, and an eye towards User-Centered Design (UCD).

Computer Supported Cooperative Work

Computer Supported Cooperative Work is the coordination and collaboration of information, in this case health information, with others at any moment in time facilitated by a number of technologies, asynchronously or synchronously, and its social, psychological, and organizational effects[4]. For the design of our solution, we have kept in mind some of the tenets of CSCW by viewing the patient, patient's family, and providers as a group coordinating and collaborating health information through the means of a patient portal or health information technology. Therefore, all users must have an awareness and shared knowledge of what one another has to offer

within the coordination of the patient care process to promote better management of health and positive changes in health behaviors.

Persuasive Technology

Persuasive Technologies are systems designed to help reinforce, change, or shape behaviors without coercion or deception[5]. A reinforced behavior occurs when a pre-existing behavior is further reinforced by the technology causing further resistance to change. Persuasive technology that changes behaviors influences the user's initial response to a situation, task, or issue[5]. Behaviors that are shaped by the persuasive technology result from creating a behavior that was not previously existent[5]. These technologies utilize a computer-human or computer-mediated persuasion[5]. To create a successfully persuasive technology, it is important to understand the issues or needs, the context of the use of the technology, and how the technology's features can promote these changes. We have kept these elements in mind as we have created our solution using computer-human persuasion. We are hoping to shape a user's patient portal behavior and change current health behaviors to healthier behaviors.

User-Centered Design

User-Centered Design is an iterative design process that takes into account the users' requirements or needs when designing a system. The user's role (*e.g.*, patient, patient's family, providers), the number of users (*e.g.*, individual or group), the user's needs (*e.g.*, medication adherence, information guidance, motivation), the user's tasks (*e.g.*, maintain health goals, create appointments, update information), and the visualization of these requirements (*e.g.*, immediate access to items most used, like information near like information) are analyzed before and during the design process in order to provide a usable and viable product for the end users[6]. We have kept these factors in mind while creating our solution to create a solution that provides the most positive impact.

Mobile Technologies

Data and information are exchanged in mass quantities every millisecond, and more and more people have access to knowledge than in any other time in history thanks to current advances in technology. The Pew Internet and American Life Project, a department of the Pew Research Center, conducted a 2009 study examining the demographics of internet users and found that teens and millennials are the widest adopters to the internet landscape as they have grown up with the internet and view it as a part of life, a necessity[7]. For those born before the millennial generation, they have slowly adopted these technologies into their work life and personal life. A large part of this adoption comes with ease of use. Ease of use is one of the most important factors that contributes to technology use. The ease of faster connections increases the percentage of people and frequency of technology use. Sharing data and information has increased both awareness and knowledge in the global population at an ever expanding rate, creating a new global culture linked by technology.

With the ubiquity of mobile phone use, new consumer expectations increased. Mobile phones are the new accessory to life and the need for knowledge and information on the go grew from a luxury to a necessity. In 2013, nearly two-thirds of adults with cell phones use it to go online[8]. Mobile Applications (apps) were the newest accessory to cell phones as millions were created to entertain the user. The Facebook app lured online users to take their networking with them and games like Angry Birds and Candy Crush kept the users glued to their phones. Yet, not all software companies set out to create the next best entertainment app. Once the influence of mobile apps was clearly seen as a way to reach this new technological audience, financial, educational and business companies took notice.

Now, desktop and wired internet connections are phasing out of popularity and mobile and wireless access has taken over. In 2009, 75% of teenagers and 83% of American adults owned a cell phone[7]. The Pew study concluded that a person's technology use is correlated to how a person chooses to connect with others and access information[7]. This brings to light the circumstances that are necessary for optimum adoption of cell phones and information exchange. Ease of information exchange can influence people to use their already ubiquitous cell phones to connect to the world. Our proposed solution uses an already created opening for healthcare and informatics to create useful and novel ways to exchange vital information to help patients better manage their disease or improve their health.

Our Proposed Solution

Patient portals have traditionally been coupled with the electronic health record (EHR) system that the provider might be currently using. There have also been standalone personal health records that can be accessed through the Internet and created by the user. Increasingly, a mobile app for managing health is now the new trend. For this project, we have chosen to utilize a patient portal app that is accessible in all types of technological mediums with a focus of its mobile app format for our mockup demonstration. Besides the usual patient and provider information management features available in patient portals, (*e.g.*, paying bills, creating appointments, and managing personal

health information), we have also included features that we hope will empower users by giving them the ability to access the information that they want to immediately see when using the patient portal, providing them with an ability to promote communications and discussion of their health with providers and family members, and supplying an additional personalized and persuasive way of enhancing the user experience while using the patient portal (Figure 1).

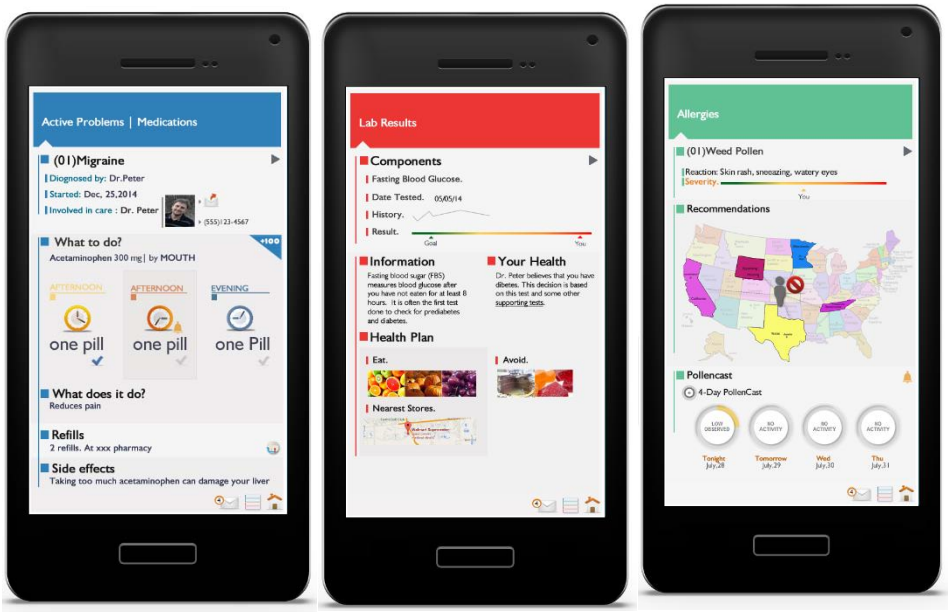


Figure 1. Some examples of UHealth's potential features empowering users and promoting healthier living.

The first image highlights a patient's ailment and the corresponding medication, pill dosage, and information for managing this ailment. In the second image, helpful hints and suggestions are given to the patient to help them manage a potentially healthier lifestyle. The third image provides users with their current location and awareness of potential health issues that may occur in that area due to their sensitivity to pollen, in this case [a].

Strengths and Weaknesses of Our Solution

Strengths

Our solution has several strengths over current features available in current patient portals that may increase the perceived benefits and actual use of patient portals. We are promoting greater occurrences for patients, their families, and providers in communicating and documenting health issues with one another in an asynchronous and synchronous way. This enhances the quality of the patient care experience for all involved in the patient care process.

UHealth also provides users with a more customized experience, not only by their own additions of health information, but also by giving them suggestions, motivations, and notifications on health choices, and use of a geographic information system (GIS) to help guide potential changes in health management. The suggestions and notifications would be generated from verified health-related resources and the motivations (*e.g.*, points earned for medication adherence towards a prize of the user's choosing) all help to promote trust in the system as it may better fit the user's needs and perception of the app's benefits.

In addition, UHealth is device agnostic, meaning it can be accessed in mobile format (*e.g.*, mobile phone or tablet) or over the Internet on a desktop computer. Being device agnostic helps the user to have the option to have their health information with them at all times (*i.e.*, mobile phone) or at their convenience on a larger screen (*i.e.*, desktop computer). Having the option for the extent of the app's convenience to the user can enable users to get a reminder for a medication refill and the closest location for the pharmacy based on the user's current location.

[a] More in depth explanations of UHealth's features can be found in our accompanied interactive PDF.

Weaknesses

A potential weakness of our solution is the amount of data exchanged through the app that might make users wary of how secure their information is within the app. Our attempts to mitigate these concerns include providing the users an option in the "Settings" tab to selectively restrict when or where their location status can be pulled by UHealth and by having all information sent and stored using a 256-bit AES encryption and Secure Socket Layer (SSL) transmission. Our solution may also not provide as much of a benefit to those that are less inclined to use technology or do not have an easy access to it. In addition, we must verify that we have correctly met all of the users' needs

Implementation and Dissemination Plan

Patient portals can be very beneficial to patients, patient's families, and providers, therefore, it is important to provide all users easier accessibility and knowledge of the product. Implementation and dissemination of our UHealth app has the following goals: 1) to promote downloads of UHealth and 2) to facilitate use of UHealth. In order to accomplish these tasks, suggestive marketing strategies will be used to increase UHealth's attractiveness to patients.

UHealth Application Downloads and Initial Use

The first step is to get UHealth into the hands of patients. When a provider sends email reminders for a checkup or appointment, a QR code or the UHealth logo can be included in the reminder that will direct the patient to go to the Google Play Store or Apple AppStore to download the patient portal app onto their smart phone. This information will also help to guide patients to UHealth's corresponding online app. In addition, providers can also include the QR code or UHealth logo on their business cards to be distributed to current patients and their families or potential patients. In addition, fliers with the main features offered by the UHealth app could be available in the provider's office for the patient's perusal during their appointment. These business cards can be attached to forms that the patients may need to fill out to update their personal information on file at the office.

Another possibility to promote downloads and initiate use is to incentivize downloads of UHealth. For instance, if a patient downloads the UHealth, they can be entered into a raffle to win a $50 gift card at their local grocery store or a free gym membership for a month. These incentives can further promote use of the features found in the UHealth patient portal such as the suggestions for healthier food substitutions and local grocery stores that may carry these items. With a gift card to that store, patients may find it easier to locate and purchase these foods and maintain healthier eating habits.

Verifying UHealth Use and Obtaining Patient Feedback

The UHealth app is only effective in helping the patient if he or she is using it and can tell where improvements are being made. Every two months, patients will receive an email or smart phone notification encouraging them to keep using UHealth. These alerts are meant to be non-invasive, but are also utilized to help increase adherence to using the patient portal on a regular basis. These notifications will also provide a healthy suggestion corresponding to a relevant health concern of the patient's. Acknowledging the notifications will provide the users with another chance to enter a raffle to win a gift card to a restaurant providing healthy food options.

Additionally, when patients return to their providers for another appointment, the providers or staff could ask them if they have downloaded UHealth. If the patient has not downloaded the app, they will be given the business card with the QR code and instructions on how to download the app. If the patient has downloaded UHealth, the patient will be given a comment card with the patient's contact information on one side and a space for comments on the other side. While the patient is updating his or her information, they can give their feedback on the comment card and put it in a drawing for another health-related prize. By obtaining patient feedback, we are both driving the dissemination and potential use of the app and motivating the iterative design process of UHealth to better answer patient's needs in their patient portals.

Moreover, patients could be asked to bring in the app, if they have already downloaded UHealth on their smart phone or as a print out if they prefer another mode of technology, in order to help assist the patient and provider communication during the visit. This can also potentially promote patient portal adherence.

Evaluation Plan

In order to gauge the effectiveness of UHealth, much analysis is needed. First, we must determine how many patients consistently use the app. By using the number of responses from the feedback cards or downloads of the app from Google Play Store or Apple AppStore, we could verify how many people downloaded and continued to use the app.

Secondly, to gauge how this app directly impacts the patients, before speaking to each patient, the provider could see if the patient has downloaded UHealth on the patient's EHR patient file. At the next visit, the doctor could also review the patient's current health compared with the prior visit and determine how many of the patients that used UHealth were healthier over time compared to the patients that did not use the app.

We will also ask all of UHealth's users to provide feedback on the app and fill out a System Usability Scale (SUS) survey in order to determine the subjective impression of satisfaction on the app[9]. The SUS is a ten question established survey traditionally used to provides a quick assessment of the app or device's usability[9]. Coupled with the feedback from all of the users, we can determine how well UHealth meets users' needs and how usable it is. Any omissions or extra features that are not required by the users' needs or tasks will be modified and pushed through to the next iteration of the app.

Additionally, we could implement an evaluation plan based on the Scenario Walkthrough and Inspection Method (SWIM) which is a method to characterize the technical, psychological, organizational, social, and economic outcomes of a collaborative system[10]. This method requires four parts. First, a requirements analysis must be conducted on all of the users and tasks that they might have to accomplish when using a patient portal[10]. This can be accomplished through ethnographic observation of the users. Second, scenarios that users currently must undertake in order to complete functions in the patient portal will be acquired as well as scenarios users wish to have will be collected through semi-structured interviews and coding of the scenarios and needs or claims by the users[10]. Third, these scenarios will be validated with both individual users and groups of users through focus groups[10]. Lastly, the results of the scenarios in relation to the claims stated by the users will be analyzed while keeping in mind the user's role, the task goal, system's features, and scenarios[10]. All four steps would be completed during the pre-UHealth implementation and during the post-UHealth implementation, six months later. A comparison of the patient care factors and user's responses will be analyzed.

Conclusion

Patient portals have the potential to help patients, their families, and providers impart more effective care and better management of health. Here, we have suggested a potential solution that is accessible across multiple platforms incorporating elements of CSCW, Persuasive Technology, and UCD as well as a potential implementation, dissemination, and evaluation plan to verify its viability. UHealth is a potential solution that may positively influence and increase use of patient portals.

References

1. Ammensworth E, Schnell-Inderst P, Hoerbst A. The impact of electronic patient portals on patient care: A systematic review of controlled trials. J Med Internet Res. 2012;14(6):e162.
2. Terry K. Patient portals: Beyond meaningful use. Physicians Practice Your Practice Your Way. Physicians Practice online; 2011 Jun 27 [cited 2014 Jul 15]. Available from: http://www.physicianspractice.com/technology/patient-portals-beyond-meaningful-use.
3. Wilkins S. If you build a patient portal, why won't they come? KevinMD. KevinMD.com online; 2013 Apr 13 [cited 2014 Jul 15]. Available on http://www.kevinmd.com/blog/2013/04/build-patient-portal.html.
4. Fitzpatrick G, Ellingsen G. A review of 25 years of CSCW research in healthcare: Contributions, challenges and future agendas. CSCW. 2013;22(4-6):609-665.
5. Oinas-Kukkonen H, Harjumaa M. Persuasive systems design: Key issues, process model, and system features. Commun Assoc Inf Syst. 2009;24(1):28.
6. Johnson CM, Johnson TR, Zhang J. A user-centered framework for redesigning health care interfaces. J Biomed Inform. 2005;38:75-87.
7. Lenhart A, Purcell K, Smith A, Zickuhr K. Social media & mobile Internet use among teens and young adults. Pew Internet & American Life Project. Pew Research Center online; 2010 Feb 3 [cited 2014 Jul 20]. Available on http://pewinterent.org/Reports/2010/Social-Media-and-Young-Adults.aspx.
8. Duggan M, Smith A. Cell Internet use. Pew Research Internet Project. Pew Research Center online; 2013 Sept 16 [cited 2014 Jul 20]. Available on http://pewinternet.org/Reports/2013/Cell-Internet.aspx.
9. Stanton NA, Salmon PM, Walker GH, Baber C, Jenkins DP. Human factors methods: A practical guide for engineering and design. Hampshire (ENG): Ashgate Publishing; 2005. p. 365-429.
10. Haynes SR, Purao S, Skattebo AL. Scenario-based methods for evaluating collaborative systems. CSCW. 2009;18:331-356.

A Patient Portal for Clinical Trials: Towards Increasing Patient Enrollment

Maurine Tong[1] and Mary McNamara[1]
[1]Department of Bioengineering, University of California, Los Angeles

Abstract

Clinical trials face delays due to the inability to enroll the necessary number of patients. A survey performed on cancer clinical trials demonstrated that American adults are willing to participate in clinical trials, however there is often an inability to recruit the required number of participants to a clinical trial study. To address this deficit, we propose an application that mines patient records, to extract diagnoses via ICD 9 codes obtained by natural language processing (NLP), and links patients with clinical trials relevant to their diagnoses. The patient portal searches relevant clinical trials tailored to a user's personal demographics, allows users to save trials from previous searches, provides a medium to initiate contact with a recruiter prior to enrollment with a click of a button, and presents additional information targeted towards patient needs such as options for continued care, cost of participation and required appointments and procedures. We theorize that this application will improve patients understanding of clinical trial information and potentially increase enrollment rates.

Introduction

Failure to recruit the required number of participants to a clinical trial is a significant hurdle impeding clinical trial research[1]. Achieving the required numbers of patients is necessary to ensure statistical power and that results will be applicable to populations outside the clinical trial participants. The failure is not due to a participant's unwillingness to enroll. 32% of American adults surveyed indicated they would be very willing to participate in a cancer clinical trial if asked to do so and an additional 38% of adults are inclined to participate in a clinical trial if asked, but hold some questions or reservations about participating[2]. The primary problem with accrual of patients is due to the lack of awareness of an appropriate clinical trial[3]. Specifically within the domain on cancer, less than 1 percent of cancer patients enroll in clinical trials. In addition, while over a third of the U.S. population identifies as a minority, less than 1% of those enrolled in cancer clinical trials are minorities. In comparison to other types of cancer, lung cancer patients are less likely to enroll in clinical trials than other cancer patients[3]. Although the average age of a person diagnosed with lung cancer is 71[4], patients over the age of 64 are underrepresented in clinical trials. A lack of participants affects research quality. A diverse group of participants is needed to design effective therapies[3,4].

The emergence of the internet has brought new opportunities for patients to be more active in their care. A particular group of patients, cancer patients, have been shown to utilize medical information online. A survey of a group of patients' self-reported Internet use before and after the diagnosis of cancer found that use of the Internet either directly (i.e., patient themselves doing the research) or indirectly (i.e., someone else searching for the patient) increased[5]. Similarly, another survey of cancer survivors, cancer patients, and patients with no history of cancer found that survivors and cancer patients tended to have a more positive view of electronic health records (EHRs) and the potential for access to its contents, as well as Health Information Technology (HIT) in general[6].

Across patient types, patient portals are being implemented[7-10], and are likely to become commonplace. Patient portals allow patients to electronically access health information managed by a healthcare institution. This is in part driven by the paradigm shift from patients as recipients of care to active members of their healthcare team. Portals are documented as having the potential to help patients make care decision[11]. They also contain the potential for patients to make health decisions that benefit others in addition to themselves, such as staying abreast of public health campaigns or enrolling in clinical trials.

Background

Other attempts to increase clinical trial enrollment use more traditional forms of patient outreach, including: using physicians as the contact point and the distribution of public services messages (e.g., T.V. and radio commercials) [12]. Previously, an EHR, the clinical counter-part of a patient portal, had been enhanced to alert physicians when they were treating a patient that fit clinical trial recruitment criteria[1]. The enhanced EHR resulted in a significant increase in the number of clinical trial referrals a physician wrote, and the number of patients who enrolled in a trial.

Similarly, in an intervention where patients were exposed to educational content about clinical trials[12], patients had a more positive attitude regarding enrolling in clinical trials after they watched an educational video.

While other efforts have helped to increase the number of patients enrolled in a trial, they do not provide a platform for disseminating clinical trial information directly to patients across trial types and patients. Using doctors to recruit patients does not enable patients to search for trials on their own. The website ClinicalTrials.gov does allow for patients to search for trials, but does not provide a means to contact the recruiter via the site, or make inclusion criteria explicit. Nor does ClinicialTrials.gov let patients save information about trials, enabling them to review the information later, and compare trials to one another. Using other media to educate patients about clinical trials as seen in Nelson 2013[12] can only provide either general information on the subject of clinical trials, or information about one specific trial. Our application provides specific information about trials, including inclusion criteria, and allows for the user to communicate with recruiters. Our application enables patients to refer back to clinical trial information at a later time and to compare trials. Providing this level of detail and the ability to save and compare trials empowers patients, encouraging patient involvement based on the information they have received.

Proposed Implementation

I. System Architecture

Our proposed system provides a link between the patient's clinical data and relevant trials from ClinicalTrials.gov. The portal is intended to summarize patient information already disclosed to the patient, highlight aspects of this information, and rate its relevancy to current clinical trials. The system architecture is shown in Figure 1. The portal fetches the required patient information. Personal medical record information is automatically searched, and ICD-9 diagnoses codes are annotated from patient reports. The extracted features (age, sex, diagnoses) are stored in a database on the server. These features are submitted to ClinicalTrials.gov and the search results are returned to the portal database. A user can access the patient portal page with the appropriate authentication. Upon logging in, a dynamically generated HTML page will display relevant clinical trials based on the search criteria of extracted features.

II. Components of Patient Portal

The main components of our patient portal are as follows (Figure 2): (1) a panel summarizing diagnoses extracted by the NLP system; (2) a panel displaying a ranked list of relevant clinical trials tailored to the patient based on age, sex, and diagnoses (these filters can also be removed); and (3) a panel summarizing the clinical trials that a patient has flagged as "of interest," each with a requirement checklist provided by the recruiter.

(1) Extracted Diagnosis

NLP modules are used to identify extracted diagnoses. A candidate list of health concerns is automatically extracted from the summary or conclusion section of radiology, pathology, or physician reports using the Mayo Clinical Text Analysis and Knowledge Extraction System (cTAKES) natural language processing (NLP) software in combination with Systemized Nomenclature of Medical Clinical Terms (SnoMED-CT) terminology[13]. Negated health concerns are deleted from the final set using cTAKES integrated negation detector. Examples of extracted diagnoses include tumor, chronic obstructive pulmonary disease (COPD), and smoking status in 5-predetermined categories. The final list of extracted diagnoses is displayed in the My Diagnoses section, and each diagnosis can be clicked on to see its definition. These diagnoses are also used as search terms to retrieve relevant clinical trial results.

(2) Relevant Clinical Trials

The query is performed using the search engine from ClinicalTrials.gov. There are preset filters on the left-hand-side of the webpage to identify clinical trials based on age, location, and conditions. Initial search results will also utilize the diagnoses extracted from the patient's medical record that are listed in the My Diagnoses section. The user can click on trials within the Search Menu's results (Figure 2, Module 2A) to see a trial's details displayed in the Clinical Trial Information Panel (Figure 2, Module 2B). At the top, Clinical Trial Information Panel summarizes general information including the condition, intervention, and primary outcome. In the middle of the panel, a table displays the inclusion and exclusion criteria organized by category and compares each criterion to the user's

characteristics. User characteristics are extracted from defined fields within the EHR. Other characteristics can be manually entered by the physician or the user. At the bottom, the panel summarizes to the user whether the user is eligible for the study based on initial search results.

Within the Clinical Trial Information Panel there is a Contact the Recruiter interactive button (Figure 2). The Contact Recruiter button lets the user send an email to the primary investigator to request more information about the trial. This request message also contains the patient's original search criteria (age, sex, diagnoses), to help the recruiter to determine how eligible the patient is for the trial in question. The recruiter requests more information regarding eligibility from the patient using the Determine Eligibility form (Figure 4) and can then reply with the additional Clinical Trial Requirements form (duration of study, cost of study, compensation, number of visits, etc.) (Figure 5). The Determine Eligibility and Clinical Trial Requirements forms are dynamic, based on the trial criteria. When the patient returns the Determine Eligibility form via the portal, the recruiter has extra information, from which to assess the patient's eligibility. This user and recruiter communication workflow is seen in Figure 3. The Add to My Clinical Trials button takes the current selected trial and adds it to the patient's My List of Clinical Trials (Figure 2 Module 3B). Although the Clinical Trial Requirements form has not been standardized, past research suggests patients are interested in learning about the cost of trials, additional tests, and follow-up care for trials[2]. Our evaluation will assess whether these information interests hold true for patients. The Clinical Trial Information Panel also contains a Request a Brochure interactive button. This button sends a request for a general overview of the study, similar to the hardcopy brochures available on clinical trials currently. Using the Request a Brochure button takes the current selected trial and adds it to the patient's My List of Clinical Trials, which is updated with the brochure content provided by the recruiter.

(3) My List of Clinical Trials and Menu Bar

This panel displays all trials selected by the user. After a trial is added using the panel above, the trial is indexed in the Clinical Trial Menu Bar (Figure 2, Module 3A). The Clinical Trial Viewer (Figure 2, Module 3B) is populated with additional information provided by the recruiter in the Clinical Trial Requirements form, such as cost of participation and any compensation; and requirements of the trial, including intervention, follow-up requirements, and options for continued care. The purpose of this panel is to track the details of each clinical trial study prior to the initial meeting with the recruiter.

Within the Clinical Trial Menu Bar, a user can check the status of each trial and perform actions if necessary. Each tile contains a customized name given by the user, and the title of the trial. The status is displayed in green (i.e., Complete Form), demonstrating that action is necessary; or red (i.e., Awaiting Recruiter Response), demonstrating that no action is necessary. A user can scroll through the tiles and click on one to populate the Clinical Trial Information Panel (Figure 2, Module 2B) and the Clinical Trial Viewer. Before enrollment, the Clinical Trial Viewer displays the checklist of steps required. After enrollment, a distinction is made between steps completed, such as initial visit completed, and steps remaining, such as follow-up appointment (not shown). The user can also export the list, print it out, and show it to their physician for more input.

Proposed Evaluation

We designed a patient portal to improve understanding, perform queries more easily, and facilitate communication. To determine if users report improved understanding, we will conduct usability testing of the portal. During this usability test, patients will compare the search results of the portal to the status quo, ClinicalTrials.gov. This usability testing will be a two arm cross over test. Patients will be randomly divided into two groups: Group A and Group B. Group A will first use our designed portal and then the website ClinicalTrials.gov. Group B will use the website ClinicalTrials.gov and then the portal. Both groups will complete a questionnaire (Table 1) after each step, based on the survey by Wu 2013[14], to determine which format they found better improved their understanding of what clinical trials were applicable to them, and which format they found easier to use.

In addition to a usability study, patients will also be followed for one year via email to determine rates of participants' enrollment in clinical trials. To do this, patients will be sent an email twice, once six months after completing the usability study, and the second time, twelve months after, to ask if they have enrolled or attempted to enroll in a clinical trial. They will be asked if they have enrolled or attempted to enroll, as some patients may try to enroll but find themselves disqualified.

Perceived Usefulness	
1	Using this portal/website can make me accomplish clinical trial information tasks quickly.
2	This portal/website is useful to manage my clinical trial information.
3	Using this portal/website can increase my productivity in managing my clinical trial information.
4	Using this portal/website can enhance my effectiveness in clinical trial information management.
Perceived Ease of Use	
5	This portal/website is easy to learn how to use.
6	This portal/website is easy to operate.
7	It should be easy to become skillful at using this portal/website.
8	This portal/website is not difficult to use.
Patient-Researcher Communication	
12	Using this portal/website can assist my communication with clinical trial researchers.
13	Using a portal/website can assist the communication between patients and clinical trial researchers in general.
Health Information Understandability	
14	Using this portal/website can improve my understanding of what clinical trials are relevant to me.
15	Using this portal/website can improve my understanding of a trial's inclusion and exclusion criteria.

Table 1. Usability Test Survey

Discussion

This portal allows for patients to search for clinical trials that match their demographics (age and sex) as well as diagnoses found in their medical records. While this portal can be used to directly provide patients with clinical trial information, future work can include integration with a tool like the enhanced EHR in Embi et al. 2005[1]. The enhanced EHR interface could add the function of letting a physician send a message to a patient regarding a trial that matches their criteria. Patients could then access these messages via this portal, with each message linked to the trial information provided in the Relevant Clinical Trials list.

Limitations of this work include the reliance on NLP to extract all diagnoses from the patient's medical records, the assumption that patients want these types of information on clinical trials, that recruiters will spend time filling out additional forms to provide patients with more information, that there is the potential to increase primary care physicians workflow should patients need help filling out the Eligibility Criteria form, and the difficulty in mapping ICD9 codes to eligibility criteria. While cTAKES has been used to extract findings from patient records[9], it may not prove as accurate in this domain. In the event that NLP does not capture all of a patient's diagnoses, manual entry and review by a physician will be necessary. This process can be streamlined, with the physician reviewing the annotations made by the system, to determine their correctness. The portal provides additional information to the patient to assist with making a better informed decision. However, a usability study is needed to better assess the types of information patients want. Results from the study will be used to refine the information visualized in the portal, to better reflect patient information needs and preferences. The portal clearly matches eligibility criteria with a patient's characteristics. However, there is no definitive mapping between ICD9 codes and eligibility criteria. A translation is needed between EHR vocabularies and their corresponding eligibility criteria.

A considerable contribution of this portal is it provides additional screening to find more patients with a higher chance of eligibility. However, additional tasks added to the recruiter workflow are a considerable effort to make. While there is a startup cost, these additional tasks can save time in the end. Creating these forms one time can help to save time later in the recruitment process, allowing for some of the patient-trial matching to occur asynchronously. Similarly, while some patients may need the help of their physician to fill out the Determine Eligibility form, this request could also be accomplished online, with an additional feature added to the portal to forward the form to the patient's physician.

Future directions for this work include: the usability study mentioned above, integrating this portal with an EHR system, and providing a dictionary within the portal to define difficult terms for patients. Integrating this portal with an EHR would have the benefits previously mentioned of allowing practitioners to flag relevant trials for patients,

and allowing patients to request help filling out the Additional Criteria form. A dictionary of clinical trial terminology with patient friendly definitions could be used to provide scroll over definitions within the portal.

Conclusion

The presented patient portal provides a solution to assist patients with finding an appropriate clinical trial. The portal automatically populates a search function with a patient's criteria and formats the results in tabular format alongside the patient's characteristics. Compared with the visualization for ClinicalTrials.gov, the presented patient portal's visualization filters information to present only what is relevant to user's diagnoses. It also provides information on care, financial, and scheduling, to assist patients deciding what trials fit their needs. In addition, a user can initiate and track communication with numerous recruiters to find the best clinical trial for them. Provision of this information can increase awareness and understanding of the clinical trial enrollment process, increase communication between recruiters and patients, and potentially increase rates of enrollment.

References

1. Embi P, Jain A, Clark J, Bizjack S, Hornung R, Harris M. Effect if clinical trial alert system on physicians. JAMA. 2005;165(19).
2. Comis RL, Miller JD, Aldige CR, Krebs L, Stoval E. Public attitudes toward participation in cancer clinical trials. Journal of Clinical Oncology. 2003;21(5):830-5.
3. Al-Rafie W, Vickers S, Zhong W, Parsons H, Rothenberger D, Habermann E. Cancer trials versus the real world in the United States. Annals of Surgery. 2011;254 (3):438-43.
4. Murthy V, Krumholz H, Gross C. Participation in cancer clinical trials: race-, sex-, age-based disparities. JAMA. 2004;291(22):2720-6.
5. Bass S, Ruzek S, Gordon T, Fleisher L, McKeown N, Moore D. The relationship of internet health information use with patient behavior and self efficacy: experiences of newly diagnosed cancer patients who contact the National Cancer Institute's Cancer Information Service. Journal of Health Communications. 2006;11:219-36.
6. Beckjord E, Rechis R, Nutt S, Shulman L, Hesse B. What do people affected by cancer think about electronic health information exchange? Results from the 2010 LIVESTRONG Electronic Health Information Exchange Survey and the 2008 Health Information National Trends Survey. Journal of Oncology Practice. 2011;7(4):237-41.
7. Silvestre A, Sue V, Allen J. If you build it, will they come? The Kaiser Permanente model of online health care. Health Affairs. 2009:334-44.
8. Winkleman WJ, Leonard KJ. Overcoming structural constraints to patient utilization of electronic medical records: a critical review and proposal for an evaluation framework. JAMIA. 2004;11:151-61.
9. Arnold C, McNamara M, El-Saden S, Chen S, Taira R, Bui A. Imaging informatics for consumer health: towards a radiology patient portal. JAMIA. 2013;20.
10. Greenhalgh T, Hinder S, Stramer K, Brantan T, Russell J. Adoption, Non-Adoption, and Abandonment of a Personal Electronic Health Record: Case Study of HealthSpace. British Medical Journal. 2010;341.
11. Krist A, Woolf S. A vision for patient-centered health information systems. JAMA. 2011;305(3):300-1.
12. Nelson R. Novel strategies may help boost clinical trial enrollment: Medscape; 2013 [cited 2014]. Available from: http://www.medscape.com/viewarticle/805618.
13. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo Clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association. 2010;17(5):507-13.
14. Wu H. Exploring Healthcare Consumer Acceptance of Personal Health Information Technology -- Personal Health Record Systems. Baltimore: University of Maryland; 2013.

Drawn Together: Enhancing Patient Engagement and Improving Diagnostic Tools through Electronic Draw-and-Tell Conversation

Deborah Woodcock MBA, Steven Williamson, Dana Womack MS,
Kimberley Anne Gray, Kate Fultz Hollis MS, Michelle Hribar PhD
Department of Medical Informatics & Clinical Epidemiology
Oregon Health & Science University, Portland, OR

Summary

Pediatric health professionals are able to assess children quickly and accurately when the children are engaged and empowered during an encounter. Research has shown that it is beneficial to have children draw their health-related symptoms and/or experiences before they speak about how they're feeling. However, capturing these children's "draw-and-tell" conversations in the electronic health record (EHR) has been challenging. We propose the creation of a mobile application ("app") named DrawnTogether™ that achieves this objective. An interactive prototype of DrawnTogether can be accessed at http://drawntogether.cooperativework.com/demo/ (please use patient Jane Smith).

The Challenge

Delivery of patient-centered care depends on the ability of providers to understand and respond to each individual's unique preferences, needs, and values[1]; eliciting and acting on this information requires providers to engage patients directly during an encounter's three primary activities: relationship-building, information exchange, and decision making.[2] These activities occur with all patients – including children.

While adults can speak directly with their healthcare providers, young children (ages 5-10 years) cannot. This does not mean that they are incapable of expressing themselves. Instead, they need a more child-centric approach to improve patient-provider communication. There are few of these approaches, so many professionals continue to defer to the child's parents, asking them to speak for their child.[3,4,5] Yet, no matter how well parents believe they know their child, they cannot fully represent the child's perspective and often relay a different story than the one told by their child. (Driessnack M 2014, personal conversation, May 15) Further, many of the current communication aids for children were originally developed for adults and later adapted for use with children by adding pictures or simpler language. For example, when children are asked to rate their pain, they are provided with a pain intensity scale of 1-10, accompanied by pictures and simple language. This is inappropriate for children as they do not yet have the cognitive skills to understand or use a scaled tool.[6]

The power of illustration. Children between 5 and 10 years of age communicate more accurately and completely when they are able to create drawings first, before they are asked to express their thoughts, feelings, and physical sensations. This child-centric approach capitalizes on children's (rather than adults') cognitive strengths.[7]

The Draw-and-Tell Conversation (DTC) approach has been shown to be an effective technique for capturing children's descriptions of their health experiences, offering insights not otherwise gained from existing tools. During the DTC, the healthcare professional invites the child to first draw how s/he is feeling, and then describe the picture. The drawing can be created on paper or electronically (e.g., using a tablet); regardless of the method used, the drawing's purpose is to facilitate the child's narrative. [6,7] However, storage and retrieval of children's drawings and accompanying narratives has proved challenging. (Driessnack M 2014, personal conversation, July 11)

Design objective. We chose to focus on the challenge of improving child-provider communication in patient-centered care. Our efforts center on the creation of a mechanism that allows for the active engagement of children at the point of care, facilitates relationship building, improves information exchange between the child and their providers, and makes it possible to incorporate child-generated data into decision making. Specifically, we endeavor to assure that children's drawings and narratives have a place both at the point of care and in the EHR.

Design Process

To understand the problem and deliver a solution, we decided to build a prototype on existing DTC research that served as an initial needs assessment, and to utilize rapid application development (RAD) software development concepts such as prototyping and frequent user involvement throughout the design and construction process. (RAD is a design philosophy that continuously incorporates user feedback to arrive at a solution more quickly than traditional "waterfall" methods.[8])

Problem definition and validation. In June 2014, Dr. Martha Driessnack, a researcher and developer of the Draw-and-Tell Conversation technique, presented an overview of her research as well as the clinical informatics challenges she continues to encounter. Through subsequent discussions with Dr. Driessnack and a review of her published research, the team identified electronic support for DTCs as the focus for the design challenge. The team ascertained the key stakeholders are children, parents, pediatric care providers and staff, and clinical informaticians.

Requirements planning. Once the problem had been defined and stakeholders identified, we further distilled design principles and workflows (Supplementary Materials Exhibits 1 and 2) and conducted a secondary qualitative appraisal of stakeholders' needs by observing Dr. Driessnack conducting DTCs in person and on video. To more deeply understand the clinical problem, team members held informal DTCs with children in their families, and conducted an extensive review of pediatrics and informatics literature (Supplementary Materials Bibliography).

User design and evaluation of application. Finally, we synthesized the design principles, observations, and literature reviews, brainstormed solution alternatives, and determined which option to develop. Representing the solution using interactive web-based prototypes (Supplemental Materials Exhibit 3), the team conducted 13 evaluative interviews to discover needs and solicit ideas from a wide variety of stakeholders (Supplemental Materials Exhibit 4), and iterated the design daily during development.

Evaluation of Possible Solutions

The team considered five different approaches for supporting DTC: 1) drawing on paper and storing in a paper chart; 2) drawing on paper and scanning into the EHR; 3) drawing using an existing mobile app; 4) using the EHR's image and annotation functionality for drawing and storing; and 5) creating a new custom mobile app. Table 1 highlights key points of comparison among these alternatives considered by the team, based on the design principles and clinical requirements we identified.

Paper creation and storage has the benefit of being low cost and low tech, but presents storage and retrieval complexity. Scanning drawings could also work, but requires more workflow steps than an app. We considered use of off-the-shelf drawing tools, but the lack of support for annotation and metadata limits their utility. EHR functionality could be extended, but requires prioritization by a vendor. Custom App Creation & Digital storage of DTC was selected because it meets all key requirements.

Proposed Solution

The proposed solution, DrawnTogether, is a mobile application that supports the capture, creation, and curation of children's drawings and, more importantly, their accompanying narratives. The name DrawnTogether reflects the linking of drawing and narrative, as well as the linking of child to provider. It is expected that children, who are now exposed to apps and devices at young ages, will find using DrawnTogether engaging, intuitive, easy to use, and fun. Based on DTC research, [6,7] we anticipate that providers using the app will be able to gather more complete and accurate information from children in the same amount of time that it would take to record a history. However, as Dr. Driessnack has shown, use of a DTC not only provides more information, but also provides information that is specific and insightful, thus having the potential to shorten time to diagnosis. Further, because visual and verbal information are brought together in DrawnTogether and accessible in the EHR, we also believe that the app will reduce the time providers spend locating and associating DTC information from multiple encounters.

DrawnTogether can also be used by parents, who can ask their child to make a drawing at home (pre-visit), while they are in the provider's waiting room, or during the office visit before the provider arrives (Supplementary Materials Exhibit 2). For maximum flexibility and widest distribution, the app is envisioned as separate from the EHR and available for multiple mobile platforms.

Table 1. Methods considered for creating and capturing drawing and narrative.

	Paper Creation & Storage	Paper Creation; Scanned Storage	App Creation; Digital Storage	EHR-Based Creation & Storage	Custom App Creation; Digital Storage *(selected approach)*
Supports paper drawing creation	Y	Y	N	Y	Y
Supports digital creation of drawings	N	N	Y	Y	Y
Digital image + text capture, storage, and retrieval	N	N	N	Y	Y
Child-friendly (intuitive, easy to use)	Y	Y	Y	N	Y
Touch-based "pinch and zoom" capability	N	N	Y	N	Y
Accessible by all users in home, ambulatory, acute settings	Y	Y	Y	N	Y
Ability to view image and text concurrently in EHR	N	N	N	Y	Y
Ability to view images and text sequentially over time	N	N	N	N	Y

The app will include the following features:

- **Zoomable touch screen for creating original images**. The drawing interface provides touch screen controls consistent with other drawing apps available for children.
- **Image capture of paper drawings.** For situations where touch screen drawing isn't desirable or available, children can draw using paper and art supplies and then take a picture of the drawing with the app. The image is loaded into the drawing screen, so additional markup can be done when the child discusses the drawing with their provider or parents.
- **Text entry for the drawing narrative**. The drawing interface provides an area for entering notes about the child's narrative. This part of the DTC is essential; the drawing cannot be saved or shared without an entered narrative.
- **Audio capture of child's narrative**. The app provides an audio recording feature that may be used when a child is describing his or her drawing. This is helpful for capturing the child's own words and speech characteristics. Because the text narrative is still important for sharing and retrieval within the EHR, the audio capture is not intended to replace text, but rather augment it.
- **Labeling with keywords and diagnoses**. Since the drawings and narratives will be referenced over time, the ability to label them with metadata is important. Doing so enables searching and easy identification of drawings in the future.
- **Display of all drawings done by a patient**. The app has a gallery feature that displays thumbnails of all drawings for easy retrieval of images. The gallery is also a way for children to view past artwork.
- **Upload the drawing and narrative or narrative only to the EHR**. The app sends the drawing and narrative to the EHR through a feature customized for each EHR: these might be as an image attached to a SOAP note; as a separate screen for DrawnTogether files (similar to the Media tab in some EHRs); or as a cloud-based web service. The app will also be built using the SMART-on-FHIR platform in anticipation of the FHIR standards for plug and play EHR applications in the future.[8]
- **Provide security and privacy controls for the app**. For the provider version of the app, authentication is required for accessing the app and for searching and accessing patients to ensure that patients drawing with the app won't be able to search or load another patient's drawings. The data from the app will be fully encrypted

using Advanced Encryption Standard (AES) and the app will be compliant with institutional protocols for secure tablets and mobile devices.
- **Provide online resources and support for the app.** The app will have a website (currently http://drawntogether.cooperativework.com) that provides tutorials on how to use the app, information about DTCs, and suggestions for incorporating the app into clinical practice.

The team believes DrawnTogether will provide a robust electronic method for initiating, annotating, retrieving, and sharing DTCs. In the initial 13 interviews, our stakeholders found the functionality of the app to be sound, but also identified challenges for its adoption, including:

- **The required tablet or smart device may not be available in clinics.** There is a workaround for this: the child can still use paper to draw and discuss their pictures while the provider can use a smartphone to capture the image and enter the narrative text.
- **Providers may not view the process of drawing as a worthwhile use of increasingly limited time available for encounters.** This can be addressed by allowing children to start their drawings before the provider is in the room, so they are ready to discuss them once the provider is present. The mobile app will allow flexibility: the drawing need not be done at the same time it is discussed. In addition, children and parents can create drawings and narratives at home that can be shared later with the provider. It is the team's hope that distributing the app to parents as well as providers will encourage the use of DrawnTogether in a clinic setting.
- **Providers may not feel confident about their ability to use DTC in their clinic practice.** A few of our stakeholders expressed concern about their ability to use the technique; specifically, they did not feel qualified to interpret the child's drawings. Since DTC does not require the provider to interpret the drawings, but rather involves listening to the child's interpretation and incorporating it into their assessment, we felt the best way to address this concern is to provide educational support during the implementation phase and through the website materials.

Implementation, Evaluation and Dissemination Plan

It is expected that many pediatric care clinics will benefit from use of the DrawnTogether app. Widespread implementation and clinical adoption requires a multi-phase implementation and dissemination plan encompassing pilot implementation with evaluation and publication of findings, development of an implementation toolkit, and educational campaigns. Once clinical effectiveness of the technique has been demonstrated, collaboration with the American Academy of Pediatrics to include electronically supported DTC in guidelines for pediatric medical homes is a suggested strategy to encourage widespread adoption.

Pilot implementation. Initial implementation of DrawnTogether will occur in a single pediatric clinic focused on management of children with chronic headaches. Stakeholders (children, parents, pediatric care providers and staff, clinical informaticians) will be engaged in all aspects of the pilot implementation. Change management will be supported by stakeholder education regarding the value and utility of DTC, and a communication plan will ensure that all stakeholders are engaged throughout the implementation. We have developed initial workflow recommendations (Supplementary Materials Exhibit 2), but local workflow redesign will be accomplished through observation of the local "as-is" visit workflow and subsequent redesign for the "to-be" workflow at each site, with stakeholder-led review and refinement. A technical readiness assessment will include software testing of the DrawnTogether app, integration testing to ensure that drawings and annotations accurately flow to the EHR, and security testing to ensure patient privacy and confidentiality. Software and workflow training will be provided to all end users, and plans for go-live will be communicated to clinic staff and patients/parents. On the day of go-live, extra staff will be on hand to provide user support and remediation training.

Pilot evaluation. Evaluation of the DrawnTogether pilot implementation can occur as follows: At the end of each DrawnTogether-supported visit during the pilot, patients/parents will be will be asked to answer a short set of questions focused on their perceptions of the ease of use of DrawnTogether, its value in facilitating child-provider communication, and its effect on child empowerment. Provider and clinical staff evaluation will occur via a post-pilot debrief/interview session to elicit experiences, perceptions, and suggestions regarding the use of DrawnTogether. Providers and clinic staff will receive a quantitative survey to elicit their perceptions of ease of use, value in facilitating child-provider communication, and value of incorporating child-contributed data in the EHR. Evaluation findings will be published in the professional literature to promote broader awareness of the innovation.

Application and workflow refinement: Consistent with a continuing RAD process, user feedback and findings from the pilot evaluation will be used to further refine the DrawnTogether app, associated workflows, end user training, and implementation processes.

Implementation toolkit. The DrawnTogether implementation toolkit will include the best practices, learnings, and artifacts from the pilot implementation and subsequent application and workflow refinement. The toolkit will also include educational materials for providers, clinic staff, and patients/parents, as well as sample documents that can be used as templates for stakeholder identification, communication planning, workflow redesign, end user training, go-live planning, and post-implementation evaluation. Additional checklists for cultural, training, and technical readiness assessments will be available in the toolkit.

Outcome evaluation and dissemination. Evaluation will occur at multiple points along the path from pilot implementation of DrawnTogether to widespread adoption of digitally supported DTCs in clinical practice. In addition to evaluation of the application, we anticipate the need for broader evaluation of the clinical effectiveness of DTC and its impact on pediatric care outcomes. Outcome evaluation planning will commence after completion of the pilot, but may include development of valid and reliable survey instruments to collect perceptions of utility, value, and impact of DTC on child-provider communication by patients, parents, and providers. Valid and reliable instrument(s) for assessing the clinical impact of DrawnTogether (e.g., long-term pain management and quality of life of pediatric patients with chronic migraines) may also need to be developed. Findings from all studies will be disseminated via peer-reviewed literature and professional conference presentations. Future directions for research include using DrawnTogether for pediatric conditions other than pain management, and for additional populations such as adults with limited verbal vocabularies.

Additional educational campaigns. We recommend a broad education campaign targeting health professionals, parents, and children in support of a wider effort to encourage providers to adopt DTC for pediatric pain management as well as other conditions. The focus of the campaign should be on engaging children in their health care and highlighting the value of using drawn images and narratives to meet this need. The campaign would invite parents to download DrawnTogether on their personal devices for use at home, and/or to create images at home or in the waiting room to support provider conversations in the clinic.

Summary

DrawnTogether is a novel solution designed to improve child-to-provider communication, increase child engagement, and allow for inclusion of child-generated information in the EHR. The project team has enjoyed this student design challenge, and we look forward to moving this project past the design stage and into clinical practice.

References

1. Institute of Medicine. Crossing the quality chasm: a new health system for the 21st century. Washington, DC: National Academy Press; 2001. p.6.
2. Cox ED, Nackers KA, Young HN, Moreno MA, Levy JF, Mangione-Smith RM. Influence of race and socioeconomic status on engagement in pediatric primary care. Patient Educ Couns 2012;87.
3. van Dulmen AM. Children's contributions to pediatric outpatient encounters. Pediatrics 1998;102.
4. Cox ED, Smith MA, Brown RL, Fitzpatrick MA. Effect of gender and visit length on participation in pediatric visits. Patient Educ Couns 2007;65.
5. Tates K, Meeuwesen L. 'Let mum have her say': turntaking in doctor-parent-child communication. Patient Educ Couns 2000;40.
6. Hourcade JP, Driessnack M, Huebner KE. Supporting face-to-face communication between clinicians and children with chronic headaches through a zoomable multi-touch app. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; 2012 May 5-12; Austin, TX. New York: ACM; 2012.
7. Driessnack M. Draw-and-tell conversations with children about fear. Qual Health Res 2006;16.
8. McConnell S. Rapid development: taming wild software schedules. Redmond, WA: Microsoft Press;1996. p.2.
9. SMART. Smart, FHIR and a plan for achieving healthcare IT interoperability [Online]. [2014?] [cited 2014 Jul 21]. Available from: URL: http://smartplatforms.org/2013/11/smart-fhir-and-a-plan-for-achieving-healthcare-it-interoperability/

The Use Of A Gamified Platform To Empower And Increase Patient Engagement In Diabetes Mellitus Adolescents

Guido Giunti, MD[1], Agustín Ciancaglini, MD PED[1], Carlos Otero, MD[1], Analía Baum, MD[1], Fernán Gonzalez Bernaldo de Quiros, MD[1]

[1]*Hospital Italiano de Buenos Aires, Ciudad Autónoma de Buenos Aires, Argentina*

Introduction

Effective chronic disease management can result in improved health outcomes and increased quality of life[1]. Such is the case that the World Health Organization has an indicator known as DALYs[2] (Disability Adjusted Life Years) to quantify the Burden of Disease from mortality and morbidity.

The International Diabetes Federation estimates that in the North American and Caribbean region there are close to 109 thousand children under the age of 15 with Type 1 Diabetes. In the United States it is estimated that the economic impact of this disease is such that a person with diabetes spends 2.3 times more than the average person's in medical expenditures[3]. In the South and Central America Region an estimated 45,6 thousand children under the age of 15 live in the region with Type 1 Diabetes[4].

Patients with type 1 diabetes are often diagnosed and start treatment at an early age. The incidence of type 2 diabetes among youth is increasing[5–8]. This is of particular concern because diabetes is a chronic disease that requires a high degree of adherence to the medical treatment plan through self-management[9–11], which may be difficult for youth of this age[12–15].So far, despite advances in technology that ease insulin delivery with pens or pumps, adherence to diabetes regimens is often problematic for patients of all ages, but most difficult for adolescents[16]. There are studies that show a lack of correlation between insulin regimen and glycemic control[17]. It is likely that self-care behaviors and educational models have substantial impact on outcomes and that increased attention to these factors may lead to improved blood glucose control. Finding ways to help adolescents engage in self-management is a critical issue[9–11].

Patient empowerment is the enhanced ability of patients to actively understand and influence their health status[18] and would be achieved through proper 'Patient Education'[19]. However, providing information alone, without the necessary incentives, does not insure a positive effect[20,21] as many young diabetic patients are well informed about their situation but don't act accordingly[22]. A systematic review of ecological interventions for health behavior suggests that the use of mobile technology-based ecological momentary interventions can be effectively implemented for a variety of health behaviors and psychological and physical symptoms[23]. Also, there is research that supports the integration of social elements to incentivize behaviors[24].

A challenging scenario

Younger generations are designated by some as ''digital natives''[25,26], for them, technologies such as personal computers, video games and mobile devices have always existed and are used as something that was always part of their lives[27]. Recently self-tracking has become rather popular[28]. Such self-monitoring and self-sensing can combine wearable sensors and wearable computing. There are already more than 1.08 billion smartphones of a total of 5 billion mobile phones around the world, with 80% of the population having a mobile phone[29]. There is great opportunity for mHealth in using these mobile devices and, in fact, a significant number of mHealth applications have been already developed.

Self-determination theory (SDT) is a macro-theory of human motivation that has been applied to identifying which factors sustain individuals' motivation within video games[30]. SDT postulates that the more often basic psychological needs for autonomy, competence, and relatedness are satisfied within a game context, making

both the experience more enjoyable and the motivation more sustainable[31]. There is also evidence that a meaningful and engaging narrative seems to help sustain engagement[32].

The use of game elements in non-game environments to enhance user experience and increase engagement is called Gamification[33]. This practice has been incorporated with commercial success into several platforms like Linkedin, Badgeville, and Facebook and this has led researchers to theorize that it could also be used as a tool to increase student engagement and to drive desirable learning behaviors on them[34]. Game elements provide engagement consistent with various other theories of motivation as well[35], such as positive psychology (e.g., flow)[36,37], and also provide instant feedback on actions. Feedback is more effective when it provides sufficient and specific information for goal achievement and is presented relatively close in time to the event being evaluated[38,39]. Feedback systems can be used to reference individual progress, make social comparisons, or they can refer to task criteria[40].

Our proposed solution

Taking the above into consideration we have proposed the development and use of a platform to promote self-management by enabling patients to monitor their condition and track their treatment adherence while creating positive feedback loops through game elements that reward desirable behaviors and provide health information.

Monitoring Aspects

Home telemonitoring (HT) represents a promising approach for enabling patients with chronic conditions to be followed up by clinicians more frequently, over longer periods of time, away from hospital settings[41–44]. Within the healthcare sector, apps are supporting the management of illnesses, thereby promoting health awareness and well-being[29,45,46]. Specifically, a multitude of apps have been developed to assist patients in the management of diabetes mellitus type 1 or 2[29,47].

The documentation function of the platform includes recording and monitoring of individual eating habits (eg, the bread unit intake); log the frequency of the user's physical activity and the individual medical therapy (type and frequency). Recording blood glucose values is another feature included although during the initial stages blood glucose values would only be inputted manually, a feature to transfer data wirelessly and automatically from Bluetooth measuring devices is scheduled for implementation. As El-Gayar et al point out[48], this will probably be an important driver for the perceived ease of use.

Integrating our platform with our Hospital's existing Personal Health Record (PHR) will allow the platform to extract previously stored data.

A detailed list of monitoring features can be found in *Appendix A. Table* 1. Features that require and benefit from PHR integration are described in *Appendix A. Table* 2.

Gamified Aspects

In order to incentivize adherence, a point system will be implemented based on operant conditioning and schedules of reinforcement[49–51]. A combination of different schedules of reinforcement will be employed. Continuous reinforcement takes place when the desired behavior is rewarded or "reinforced" every single time it occurs, this is used in order to create a strong association between behavior and response. In partial reinforcement, the response is reinforced only part of the time which causes the subject to continue trying. Learned behaviors are acquired more slowly with partial reinforcement, but the response is more resistant to extinction. There are different schedules of partial reinforcement based on the number of responses that have occurred (ratio) or the length of time since the last reinforcer was available (interval). These in turn can be either fixed or variable[49–51]

This point system provides feedback to patients in the form of different "scores", breaking the long arc of managing a chronic lifelong condition into smaller, more manageable units. Feedback loops are essential parts of all games, and they are seen most frequently in the interplay between scores and levels[52,53]. Levels and other progress mechanics such as "achievements" further add to this positive feedback loop. An assortment of visual elements will represent the different increments to the individual's point score.

Points System

Performing actions that are needed for managing their conditions will be considered a "required goal" and the patient will receive "points". "Points" will be rewarded based on Diabetes Mellitus' clinical and treatment objectives[54] with the possibility of customizing a patient's reward to specific activities to allow fine-tuning. During the patient's initial use points will follow a continuous reinforcement schedule (see *Appendix A. Table* 3).

An accumulation of a sufficient number of points increases a *patient*'s "Level", acting in this manner as a fixed ratio partial reinforcement. A "level" is a number that represents the patient's overall performance and treatment adherence. By gaining levels, patients will gain status and "achievements" within the gamified platform. The level progression is geometrical, to give the patient a better sense of progress and growth during the game. See *Appendix A. Table* 4.

Achievements System

"Achievements" are rewards to "optional goals" in the form of arbitrary challenges. These achievements may coincide with the inherent goals the treatment itself, such as having perfect adherence to the treatment for a specific period of time, or may also be independent of the clinical and treatment objectives, such as helping a fellow patient stick to the treatment. Achievements act as fixed-ratio or fixed-interval partial reinforcements (see *Appendix A. Table* 5).

Social Aspects

Social support is an important part of chronic disease management which is why we decided to include social elements to provide encouragement and positive feedback between users.

User Profile

Patients will have a User Profile which will aggregate information from the Monitoring Aspects and the Gamified Aspects. This is meant so that the patient can display his or hers Points, Achievements and Levels giving the user a "bragging rights" about how well they are taking care of themselves.

Whenever the user gets Achievements or Levels, that information will be highlighted on their User Profile. The user will have full control as to which information, whether it's personal, clinical or treatment related, becomes public through adjusting the user account settings. Information can be shared with anyone on the platform (public), only Friends or no one (private).

Friending and Groups

Friending is the act of adding someone to a list of "friends" on a social networking service. The notion does not necessarily involve the concept of friendship[55]. The act of "friending" someone grants that person special privileges on the platform with respect to oneself. Users will be able to cheer and like their friend's updates to provide encouragement.

Friends can invite other users to be part of a Group. Groups will have access to special Achievements and Points bonifications when all members meet their weekly goals.

Messaging and Commenting

Commenting and instant messaging are features common in Social Networks but it was decided against implementing them into the platform because of the difficulty of moderating polite behavior.

Patient Education

Diabetes education has largely been accepted in diabetes care. Our proposed platform provides the patient with health and wellness tips. These tips, also include interesting facts about the condition to increase the patient's awareness of the condition.

The content of the different Tips of the Day is developed by an interdisciplinary team of Pediatricians, Diabetologists, Child Psychiatrists and Patient Education experts to optimize communication. These Tips are less than 260 characters in length to keep the patient from feeling burdened with information.

The time of the day at which the Tips are displayed can be set by the patient and it's a feature that can be turned off. The patient can also explore the list of previously displayed Tips for later reference. A sample Tips of the Day can be found in *Appendix A. Table* 6.

Alternative solutions considered

There's been several tries to bridge the gap generated between prescription and patient compliance. A common theme among the different therapeutic approaches is the creation of a dynamic and interdisciplinary support framework for the patient.

Telemonitoring

Gómez el al.[56] describe the DIABTel telemedicine system to complement the daily care and intensive management of diabetic patients. There was a trend towards HbA1c improvement during DIABTel use with no incidence in the number of hypoglycaemias.

Bujnowska et al.[57] published a study where a telemedicine support system for diabetes management was compared with standard monitoring. There was no significant difference in haemoglobin A(1c) between telemonitoring and the traditional group of diabetic patients during the survey.

Family Environment

Wysocki et al.[58]'s work with Behavioral Family Systems Therapy for Diabetes (BFST-D) aims to assist parents and adolescents as they work on communication skills, problem solving, and minimizing family conflict in relation to diabetes. This paper showed significant improvement in the quality of family interactions, family communication, and problem solving with BFST-D.

Social Support

Grey et al.[59] studied a form of cognitive behavioral therapy that they called coping skills training (CST) in 12–20-year-olds with T1D who were beginning intensive insulin therapy. The authors reported better glycemic control and quality of life with the addition of CST to intensive therapy.

Peters et al.[60] shows that the adolescents with diabetes identified various supportive behaviors of friends, particularly concerning emotional support: treating them normally, showing interest, having fun, providing a distraction, and taking their diabetes into account. Fear of stigmatization and sense of autonomy withheld some adolescents with diabetes from soliciting more support.

Strengths and weaknesses analysis

The use of game elements as a way of motivating patients is an interesting approach as it provides patients with instant feedback and positive reinforcements. This, we believe, will allow patients to conceptualize their condition in more manageable units and will seem like a less tiresome task. Balancing the point system is a matter that requires cautions otherwise patients might place too much emphasis in obtaining points and feel inclined to exploit the system.

Integrating a home telemonitoring system with our Hospital's PHR adds value as the monitoring tool will not be a standalone device but part of a larger system that will significantly improve the quality of care offered to patients.

While reviewing the literature, however, we realized that at this time our platform lacks a way to allow family participation. This is the subject of an ongoing discussion within our team as to what would be the proper role and the right approach.

Although the alternative solutions are not without merit our platform attempts a more integrative approach to the patient's condition. An illustrative comparison between our platform's strengths and weaknesses against the alternative solutions explored can be found in *Appendix A. Table* 7.

Implementation and dissemination plan

Our Hospital's Chronic Disease Program (CDP), Department of Pediatrics and Endocrinology Service will work with the Healthcare IT department during the planning, structuring, software development and the process of implementation and validation of a pilot. Given possible legal and ethical implications both the Hospital's Legal Department and the Bioethics Committee will stay informed of the project's coming and goings.

Implementation of the platform will follow a three stage process. In Stage 1 we will perform a functionality test with 20 healthy adults in order to validate the process and operation of the platform. Stage 2 will consist of a pilot test with 20 teenagers assessing flow and ease of use, once this pilot test is completed we will move to Stage 3. Stage 3 involves the massive dissemination and large scale implementation of the platform to susceptible and willing patients.

Diffusion and dissemination of the platform will be carried out through the Hospital's usual channels of communication such as newsletters mailing, Hospital magazines, website, internal TV, etc. The use of other social media is being considered to capture the interest of adolescents participating in the study.

Evaluation plan

As this is a patient-oriented platform, their input and feedback is crucial. Once our solution is implemented the idea is to advance on several assessment lines. We will assess user satisfaction through a SUS questionnaire, and conduct qualitative studies using semi structured interviews. To optimize the processes of patient interaction with the tool and generate an intuitive and easy to use application, we will perform regular assessments of interface usability.

Finally, a randomized controlled trial will be conducted to evaluate clinical outcomes, based on clinical parameters involved in the protocol, in order to assess the platform's impact.

References

1. Lorig KR, Sobel DS, Stewart AL, et al. Evidence suggesting that a chronic disease self-management program can improve health status while reducing hospitalization: a randomized trial. *Med Care*. 1999;37:5-14. Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=10413387.

2. Murray CJ. Quantifying the burden of disease: the technical basis for disability-adjusted life years. *Bull World Health Organ*. 1994;72:429-445.

3. American Diabetes A. Economic costs of diabetes in the U.S. in 2012. *Diabetes Care*. 2013;36(4):1033-1046. doi:10.2337/dc12-2625.

4. International Diabetes Federation. *IDF Diabetes Atlas, 6th Edn*. Brussels, Belgium: International Diabetes Federation, 2013 Available at: http://www.idf.org/diabetesatlas.

5. Onkamo P, Väänänen S, Karvonen M, Tuomilehto J. Worldwide increase in incidence of Type I diabetes--the analysis of the data on published incidence trends. *Diabetologia*. 1999;42:1395-1403. doi:10.1007/s001250051309.

6. Karvonen M, Viik-Kajander M, Moltchanova E, Libman I, LaPorte R, Tuomilehto J. Incidence of childhood type 1 diabetes worldwide. Diabetes Mondiale (DiaMond) Project Group. *Diabetes Care*. 2000;23:1516-1526. doi:10.2337/diacare.23.10.1516.

7. Fagot-Campagna A, Pettitt DJ, Engelgau MM, et al. Type 2 diabetes among North American children and adolescents: an epidemiologic review and a public health perspective. *J Pediatr*. 2000;136:664-672.

8. Fagot-Campagna A, Saaddine JB, Flegal KM, Beckles GL. Diabetes, impaired fasting glucose, and elevated HbA1c in U.S. adolescents: the Third National Health and Nutrition Examination Survey. *Diabetes Care*. 2001;24:834-837.

9. Ahmann AJ. Guidelines and performance measures for diabetes. *Am J Manag Care*. 2007;13 Suppl 2:S41-S46. doi:17417932.

10. Albano MG, Crozet C, D'Ivernois JF. Analysis of the 2004-2007 literature on therapeutic patient education in diabetes: Results and trends. *Acta Diabetol*. 2008;45:211-219. doi:10.1007/s00592-008-0044-9.

11. Silverstein J, Klingensmith G, Copeland K, et al. Care of Children and Adolescents With Type 1 Diabetes: A statement of the American Diabetes Association . *Diabetes Care* . 2005;28 (1):186-212. doi:10.2337/diacare.28.1.186.

12. Holl RW, Swift PGF, Mortensen HB, et al. Insulin injection regimens and metabolic control in an international survey of adolescents with type 1 diabetes over 3 years: Results from the Hvidore study group. *Eur J Pediatr*. 2003;162:22-29. doi:10.1007/s00431-002-1037-2.

13. Silverstein J, Klingensmith G, Copeland K, et al. Care of Children and Adolescents With Type 1 Diabetes: A statement of the American Diabetes Association . *Diabetes Care* . 2005;28 (1):186-212. doi:10.2337/diacare.28.1.186.

14. Helgeson VS, Siminerio L, Escobar O, Becker D. Predictors of metabolic control among adolescents with diabetes: A 4-year longitudinal study. *J Pediatr Psychol*. 2009;34:254-270. doi:10.1093/jpepsy/jsn079.

15. Holmes CS, Chen R, Streisand R, et al. Predictors of youth diabetes care behaviors and metabolic control: A structural equation modeling approach. *J Pediatr Psychol*. 2006;31:770-784. doi:10.1093/jpepsy/jsj083.

16. Morris AD, Boyle DI, McMahon AD, Greene S a, MacDonald TM, Newton RW. Adherence to insulin treatment, glycaemic control, and ketoacidosis in insulin-dependent diabetes mellitus. *Lancet*. 1997;350(9090):1505-1510. doi:10.1016/S0140-6736(97)06234-X.

17. De Beaufort CE, Swift PGF, Skinner CT, et al. *Continuing Stability of Center Differences in Pediatric Diabetes Care: Do Advances in Diabetes Treatment Improve Outcome? The Hvidoere Study Group on Childhood Diabetes.*; 2007:2245-2250. doi:10.2337/dc07-2275.

18. D'Alessandro DM, Dosa NP. Empowering children and families with information technology. *Arch Pediatr Adolesc Med*. 2001;155:1131-1136. doi:10.1001/archpedi.155.10.1131.

19. Johansson K, Leino-Kilpi H, Salantera S, et al. Need for change in patient education: A finnish survey from the patient's perspective. *Patient Educ Couns*. 2003;51:239-245. doi:http://dx.doi.org/10.1016/S0738-3991%2802%2900223-9.

20. Fawcett SB, White GW, Balcazar FE, et al. A contextual-behavioral model of empowerment: Case studies involving people with physical disabilities. *Am J Community Psychol*. 1994;22:471-496. doi:10.1007/BF02506890.

21. Challenging the balance of power: patient empowerment. *Nurs Stand*. 2004;18(22):33-37. doi:10.7748/ns2004.02.18.22.33.c3546.

22. Keers JC, Blaauwwiekel EE, Hania M, et al. Diabetes rehabilitation: Development and first results of a Multidisciplinary Intensive Education Program for patients with prolonged self-management difficulties. *Patient Educ Couns*. 2004;52:151-157. doi:10.1016/S0738-3991(03)00019-3.

23. Heron KE, Smyth JM. Ecological momentary interventions: incorporating mobile technology into psychosocial and health behaviour treatments. *Br J Health Psychol*. 2010;15:1-39. doi:10.1348/135910709X466063.Ecological.

24. Anderson-Hanley C, Arciero P, Snyder. Social facilitation in virtual reality-enhanced exercise: competitiveness moderates exercise effort of older adults. *Clin Interv Aging*. 2011:275. doi:10.2147/CIA.S25337.

25. Prensky M. Digital Natives, Digital Immigrants Part 1. *Horiz*. 2001;9(5):1-6. doi:10.1108/10748120110424816.

26. Prensky M. Digital Natives, Digital Immigrants Part 2: Do They Really Think Differently? *Horiz*. 2001;9:1-6. doi:10.1108/10748120110424843.

27. Johnson L, Smith R, Willis H, Levine A, Haywood K. *The 2011 Horizon Report*.; 2011:36. doi:10.1002/chem.201001078.

28. Swan M. The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery. *Big Data*. 2013;1:85-99. doi:10.1089/big.2012.0002.

29. Martínez-Pérez B, de la Torre-Díez I, López-Coronado M. Mobile health applications for the most prevalent conditions by the World Health Organization: review and analysis. *J Med Internet Res*. 2013;15(6):e120. doi:10.2196/jmir.2600.

30. Rigby S, Richard R. Immersion and Presence. In: *Glued to Games: How Video Games Draw Us In and Hold Us Spellbound*.; 2011:81-96.

31. Sylvester BD, Standage M, Dowd a J, Martin LJ, Sweet SN, Beauchamp MR. Perceived variety, psychological needs satisfaction and exercise-related well-being. *Psychol Health*. 2014;29(9):1044-61. doi:10.1080/08870446.2014.907900.

32. Lu AS, Baranowski T, Thompson D, Buday R. Story Immersion of Video Games for Youth Health Promotion : A Review of Literature. *Heal San Fr*. 2012;1. doi:10.1089/g4h.2011.0012.

33. Kapp KM. *The Gamification of Learning and Instruction: Game- Based Methods and Strategies for Training and Education*. San Francisco, CA: Pfeiffer; 2012.

34. Lee JJ, College T, Ph D, Hammer E, Interdisciplinary M. Gamification in Education : What , How , Why Bother ? What : Definitions and Uses. 2011;15:1-5.

35. Ryan R, Deci E. Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. *Contemp Educ Psychol*. 2000;25:54-67. doi:10.1006/ceps.1999.1020.

36. Csikszentmihalyi M. The domain of creativity. In: *Theories of Creativity*.; 1990:190-212.

37. Davis MS, Csikszentmihalyi M. Beyond Boredom and Anxiety: The Experience of Play in Work and Games. *Contemp Sociol*. 1977;6:197. doi:10.2307/2065805.

38. Graesser AC, Person NK. Question Asking During Tutoring. *Am Educ Res J*. 1994;31:104-137. doi:10.3102/00028312031001104.

39. Prensky M. Digital game-based learning. *Comput Entertain*. 2003;1:21. doi:10.1145/950566.950596.

40. Wilbert J. Effects of Evaluative Feedback on Rate of Learning and Task Motivation: An Analogue Experiment. 2010;8(2):43-52.

41. Meystre S. The current state of telemonitoring: a comment on the literature. *Telemed J E Health*. 2005;11:63-69. doi:10.1089/tmj.2005.11.63.

42. Paré G, Poba-Nzaou P, Sicotte C. Home telemonitoring for chronic disease management: an economic assessment. *Int J Technol Assess Health Care*. 2013;29:155-61. doi:10.1017/S0266462313000111.

43. Anker SD, Koehler F, Abraham WT. Telemedicine and remote management of patients with heart failure. *Lancet*. 2011;378:731-739. doi:10.1016/S0140-6736(11)61229-4.

44. Roine R, Ohinmaa A, Hailey D. Assessing telemedicine: a systematic review of the literature. *CMAJ*. 2001;165:765-771.

45. Cafazzo J a, Casselman M, Hamming N, Katzman DK, Palmert MR. Design of an mHealth app for the self-management of adolescent type 1 diabetes: a pilot study. *J Med Internet Res*. 2012;14(3):e70. doi:10.2196/jmir.2058.

46. Eng DS, Lee JM. The promise and peril of mobile health applications for diabetes and endocrinology. *Pediatr Diabetes*. 2013;14:231-238. doi:10.1111/pedi.12034.

47. Chomutare T, Fernandez-Luque L, Arsand E, Hartvigsen G. Features of mobile diabetes applications: review of the literature and analysis of current applications compared against evidence-based guidelines. *J Med Internet Res*. 2011;13:e65. doi:10.2196/jmir.1874.

48. El-Gayar O, Timsina P, Nawar N, Eid W. Mobile applications for diabetes self-management: status and potential. *J Diabetes Sci Technol*. 2013;7:247-62. doi:10.1089/dia.2014.1507.

49. Ferster, CB, & Skinner B. *Schedules of Reinforcement*. New York, New York, USA: Appleton-Century-Crofts; 1957.

50. Skinner BF. *Science and Human Behavior*.; 1953:458.

51. Skinner B. *Contingencies of Reinforcement: A Theoretical Analysis*. New York, New York, USA: Appleton-Century-Crofts; 1969.

52. Zichermann G, Cunningham C. *Gamification By Design*.; 2008:208. Available at: http://medcontent.metapress.com/index/A65RM03P4874243N.pdf.

53. Schell J. *The Art of Game Design, a Book of Lenses*.; 2008:489. doi:9780123694966.

54. American Diabetes A. Standards of medical care in diabetes--2014. *Diabetes Care*. 2014;37 Suppl 1:S14-80. doi:10.2337/dc14-S014.

55. Drucker SJ, Gumpert G, Cohen HM. *Regulating Convergence. Communication Law.* Peter Lang International Academic Publishers; 2010.

56. G??mez EJ, Hernando ME, Garc??a A, et al. Telemedicine as a tool for intensive management of diabetes: The DIABTel experience. In: *Computer Methods and Programs in Biomedicine*.Vol 69.; 2002:163-177. doi:10.1016/S0169-2607(02)00039-1.

57. Bujnowska-Fedak MM, Puchała E, Steciwko A. Telemedicine for diabetes support in family doctors' practices: a pilot project. *J Telemed Telecare*. 2006;12 Suppl 1:8-10. doi:10.1258/135763306777978551.

58. Wysocki T, Harris MA, Buckloh LM, et al. Randomized, controlled trial of Behavioral Family Systems Therapy for Diabetes: maintenance and generalization of effects on parent-adolescent communication. *Behav Ther*. 2008;39(1):33-46. doi:10.1016/j.beth.2007.04.001.

59. Grey M, Boland EA, Davidson M, Li J, Tamborlane W V. Coping skills training for youth with diabetes mellitus has long-lasting effects on metabolic control and quality of life. *J Pediatr*. 2000;137(1):107-113. doi:10.1067/mpd.2000.106568.

60. Peters LW, Nawijn L, van Kesteren NM. How adolescents with diabetes experience social support from friends: two qualitative studies. *Sci*. 2014;2014:415849. doi:10.1155/2014/415849.

Sharing My Health Data: A Survey of Data Sharing Preferences of Healthy Individuals

Elizabeth A. Bell, MPH[1,2], Lucila Ohno-Machado, MD, PhD[1], M. Adela Grando PhD[3]
[1]University of California San Diego Division of Biomedical Informatics, [2]University of Minnesota School of Public Health, [3]Arizona State University Department of Biomedical Informatics

Abstract

We interviewed 70 healthy volunteers to understand their choices about how the information in their health record should be shared for research. Twenty-eight survey questions captured individual preferences of healthy volunteers. The results showed that respondents felt comfortable participating in research if they were given choices about which portions of their medical data would be shared, and with whom those data would be shared. Respondents indicated a strong preference towards controlling access to specific data (83%), and a large proportion (68%) indicated concern about the possibility of their data being used by for-profit entities. The results suggest that transparency in the process of sharing is an important factor in the decision to share clinical data for research.

Introduction

In parallel with the rapid accumulation of electronic health-related data, ethical considerations and public concern related to clinical data sharing for biomedical and behavioral research have been raised[1]. The debate on how ethical it is to use data and biospecimens—collected primarily for clinical care—for research purposes spans scientific, legal/ethical/regulatory, and patient circles. Many patients may not know how their data are being used for research. Some may receive a re-consent form, but this requirement can be waived by IRBs (institutional review boards) in certain circumstances. Consent is typically broad and binary (i.e., patients either consent or not), and tiered approaches to informed consent are seldom utilized. Innovative systems that inform patients about the current use of their data and allow them to select tiered opt-out options could offer an alternative to current practices. However, institutions may be hesitant to adopt such systems, since this might decrease participation in data sharing for research and potentially bias research results. The financial and political costs of implementing such systems, as well as their efficacy in terms of patient and provider satisfaction, are currently unknown. A recent survey indicates that many patients believe the use of electronic health records will improve care[2]. Technical obstacles also exist, as ways of ensuring compliance to patients' choices may be difficult to implement and upkeep. The ethical gains in implementing choices for patients, however, may justify the development towards systems of granular control should continue[3].

Input from patients may potentially drive future policies for data sharing. In Great Britain, an open consultation was launched in October 2013 to give citizens an opportunity to share their views about the collection, use, and analysis of their personal medical data. The goal of this initiative was to determine peoples' expectations towards privacy and anonymity, as well as to address the ethical concerns of the use of private information. Results are expected mid- 2014[5]. In addition, a leaflet campaign through the NHS system distributed information to all residents about sharing data to improve quality and care for everyone, and how they could opt out of such a system[6]. Reactions have been mixed: 4 out of 5 surveyed patients in favor of sharing their data yet there is suspicion and distrust for the project, and indications that the methods by which patients are invited to control their own medical records may impact their feelings about data sharing[7,8].

Some literature suggests that patients may want to have control over which institutions have access to their data for research[3-4]. There may be multiple factors influencing subjects' attitudes towards sharing their medical data for research[9]:

1. Type of information: subjects are less willing to share information that is highly personal, such as sensitive information about drug abuse, sexual-related diseases, or mental health disorders[9].
2. The type of recipient: subjects' willingness to share decreases when the recipient of the information is a commercial or for-profit entity[10,11]. In a study by Willison et al., participants who had specific targeted health conditions or were generally healthy individuals thought that health information should not be used for marketing purposes, and that re-consent should be needed for use in the case of research by for-profit organizations[10]. Focus groups and interviews about different scenarios found that most participants were concerned about for-profit uses of their information[12].
3. Level of anonymity: subjects are concerned about privacy and are more willing to share information that is de-identified[13].
4. Health condition of the subject: if the subject has a progressive or chronic illness, the individual tends to be more willing to share[10,14].

5. Perceived value: recent emergence of many active Internet communities for patients experiencing similar conditions shows that subjects may value peer-based sharing models and the possibility of hastening treatment and cure of their medical conditions[15].

These studies indicate that using a broad consent for personal data may not be what people desire. If a tiered consent model is used, it can provide subjects with more options and opportunity for involvement in the information sharing process. The U.S. Office of the National Coordinator supports granular control models for Health Information Technology, and suggests that patients should have a "greater degree of choice to determine, at a granular level, which personal health information should be shared with whom, and for what purpose"[16].

Technological difficulties make it challenging to track and ensure compliance with patients' choices even if they are given options for sharing medical data. Currently, most health providers do not have a system for patients to see who is accessing their data for research purposes, and also do not allow patients to select certain categories of information for sharing. The literature suggests that patients may want to have options for keeping their information private[4,12]. A study by Caine et al., investigated tiered sharing for patients through a system of cards and questionnaires[4], while Meslin et al. gave a series of guidelines to consider when creating electronic health records[18], and these studies formed the basis of our investigation into different categories and sharing options for patients. However, we are aware that theoretical surveys can be limiting[4] as patients may respond differently when it is not their actual medical data at stake.

We surveyed 70 healthy volunteers to establish a baseline on patient preferences who responded to an advertisement posted at several different locations on the UCSD campus for a period of 4 months. The surveys and interviews were conducted between 7/15/2013 and 10/30/2013. The survey was designed at as preliminary study of a larger project that will implement tiered patient preferences for use with a clinical data warehouse for research (CDWR). iCONCUR (informed CONsent for Clinical record and sample Use in Research) is a project of iDASH (integrating Data for Analysis, anonymization and SHaring), an NIH-funded National Center for Biomedical Computing (NCBC)[21]. iCONCUR will record the patient's choices for sharing medical information and this information will be transmitted to the CDWR where permissions will disallow sharing of corresponding data about subjects who register their preferences in iCONCUR. iCONCUR will begin in late summer 2014, and will enroll 400 patients from 2 sites – a general Internal Medicine Clinic and a specialty clinic that exclusively treats HIV positive patients. By recruiting a diversified group of patients, we hope to discover trends among patient sharing choices that could be generalized to a wider population.

Materials and Methods

Participant choices were implemented in a graphical user interface (GUI). From the GUI, participants proceed through three taxonomies where choices about sharing data can be made: (1) *What am I sharing? (*2) *Who am I sharing it with?* and (3) *Which type of funding do I allow?* Participants had both the GUI and the survey open in web browser windows so they could refer to both as needed. The study took place at UCSD, and participants met with the researcher (EAB) in person for 45-60 minutes to discuss their choices. During the process of using the GUI and filling out the survey, participants were encouraged to ask questions or give verbal feedback in addition to the feedback that was collected in the survey. No connection of these choices was done with the individual's CDWR data, as the intent was only to test the appropriateness of available choices.

Educational Materials

An introduction section provided an overview of research, legal requirements for disclosure of medical information for research purposes, and information about how data sharing could contribute to research. All the material was written at an 8th grade reading level. As an example of how medical records could lead to important research discoveries, a link was provided to a recent article that explored data mining of electronic health records[17]. The idea to include this example came from preliminary interviews. Additional links led to websites about data anonymization, IRB policies, and NIH pages detailing HIPAA.

What am I sharing?

The first taxonomy of choices addressed the content of the person's health records. As depicted in Figure 1, options included "*demographics*", "*test and lab results*", and "*diagnostic information*". The diagnostic information was classified as "*sensitive information*" or "*non-sensitive information*". The sensitive information was classified in four categories: "*mental health*", "*sexual and reproductive health*", "*alcohol and substance abuse*", and "*genetic information*".

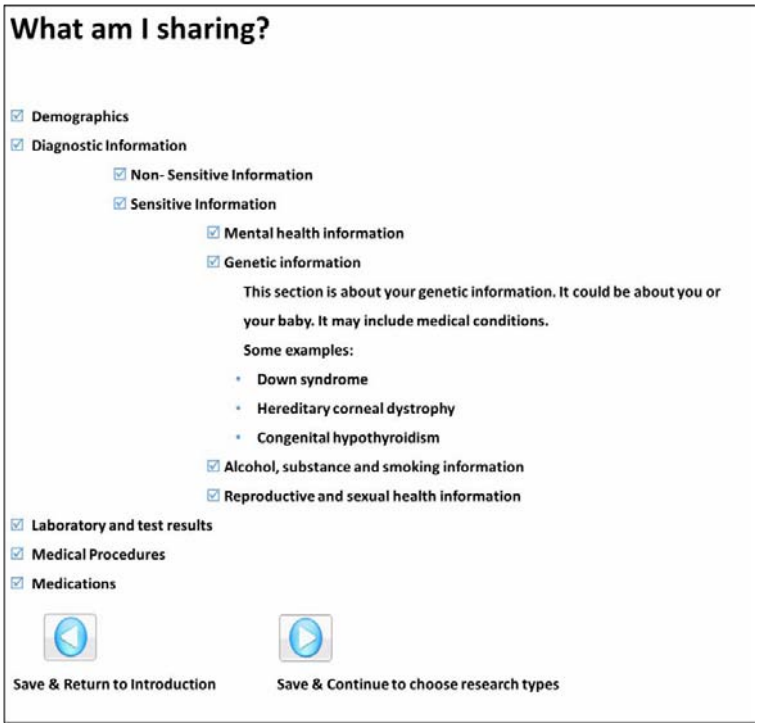


Figure 1. Screenshot of GUI with taxonomy of data sharing choices

Links were provided to the diagnoses for each category of sensitive information, which were selected by two clinicians. Each category expanded to give further descriptions, as depicted in Figure 1 with genetic information. A taxonomy of sensitive categories of information was chosen because previous investigations had shown that sharing sensitive medical data could be a concern to patients[4]. In previously published work when patients were given the option of choosing how to share their data, many chose to only share less sensitive data[4]. The definition of sensitive data is described by the National Committee on Vital and Health Statistics and includes five categories: domestic violence, genetic information, mental health information, reproductive and sexual health, and substance abuse[19]. For the development of the taxonomy, a group of UCSD clinicians used the definition of sensitive information from *Points to consider in ethically constructing patient-controlled electronic health records*[18] to categorize diagnosis codes from CDWR. This way, when a patient decides to share or not share a category of sensitive information, we can precisely determine the set of diagnosis codes available in the CDWR that the patient was selecting and comply with his/her choices. These categorizations were arbitrary, and we acknowledge that different clinicians could have selected the diagnosis codes differently. For the classification of sensitive information we decided to focus only on diagnostic codes. There may be sensitive information contained in other parts of a patient's medical record, but we have not addressed it in the current survey.

Who am I sharing it with?

In the second taxonomy, subjects had the opportunity to determine whether they wanted to share data with a research team that is: (1) entirely formed by UCSD or San Diego Veteran's Administration (VA) hospital researchers, or (2) led by UCSD or San Diego VA hospital researchers but involving members outside UCSD and San Diego VA hospital. These options were included because UCSD IRB protocols require that all the research plans that request access to clinical data from the CDWR should be led by UCSD or San Diego VA, and to indicate whether there are research members outside these institutions.

Which type of funding do I allow?

In the last taxonomy subjects had the option of deciding what types of institutions should have access to their medical data. The options included *no restriction on type of institution*. If this option was not chosen subjects could select some or all of (1) *commercial*, (2) *mixed*, (3) *non-commercial*, and (4) *unfunded research*.

Evaluation Study

Recruitment was conducted using flyers on the UCSD campus, the UCSD medical center, and the San Diego VA hospital. The inclusion criteria used were: (1) English speaker, and (2) Age 18 or older. Participants began by using the GUI to make their own personal choices about sharing preferences. All choices were saved for the evaluation. After participants completed their choices, a link led to a 28-question survey.

Eight questions collected demographic information, in order to detect possible trends in sharing choices based on gender, income, and educational level. Three questions evaluated the study to check reading comprehension, twelve questions captured the participant's opinions towards the sharing options and their motivations for sharing or not sharing medical data, and five questions related to the online GUI itself. In total, participants spent 45-60 minutes using the tool and taking the survey.

Participants choices for not sharing	
Sensitive Information	**N (%)**
Genetic information	9 (12.9)
Sexual and reproductive health information	12 (17.1)
Mental health information	10 (14.3)
Alcohol and drug use information	6 (8.6)
Types of Sponsorship	
For-profit	16 (22.9)
Mixed (for-profit and non-profit)	7 (10)
Non-profit	1 (1.4)
Unfunded	5 (7.1)
Researcher Types	
Teams composed of UCSD and VA researchers	0
Teams that included external researchers	8 (11.4)
Other Categories of Medical Information	
Demographics	3 (4.3)
Non-sensitive diagnostics	1 (1.4)
Laboratory and test results	0
Medical procedures	3 (4.3)
Medications	3 (4.3)

Table 1. Number and percentage of participants who selected 'do not share' for each data category.

Results

Choices about sharing were recorded in a database and are shown in Table 1. The numbers indicate the number and percentage of participants who would choose *not* to share data in each category. Not sharing with researchers from for-profit institutions was the most common choice. For all of the survey questions of interest, a chi-square test was applied to test for association of attributes and the outcome of interest. Representative comments for six of the survey questions that captured the participant's opinions towards sharing options and motivations for sharing are in shown in Table 2. Results that showed statistical significance at a level of of $p < 0.05$ are indicated.

The results show that participants were significantly more willing to have their health data shared for research if they were given choices about which aspects of their data they wished to share. Participants were also interested in knowing more about the researchers and indicated a significant desire to know who was accessing their data. Ninety-four percent of participants indicated that they wanted to be able to know what kind of organization the researcher belonged to, 89% wanted to know the aim of the research study, 84% were interested in being informed of the outcomes of the research, and 70% would like access to publications that resulted from using their data. Preferences for being notified of how and when their data are being used were quite varied. Forty-four percent of participants want to be informed each time a new researcher uses their data, 20% prefer once a month (even if there were no changes), 13% once a year (even if there were no changes), 17% never if they could go online to the site and find it themselves, and 6% had other suggestions. Participants correctly answered the three questions designed to check reading comprehension 91% of the time.

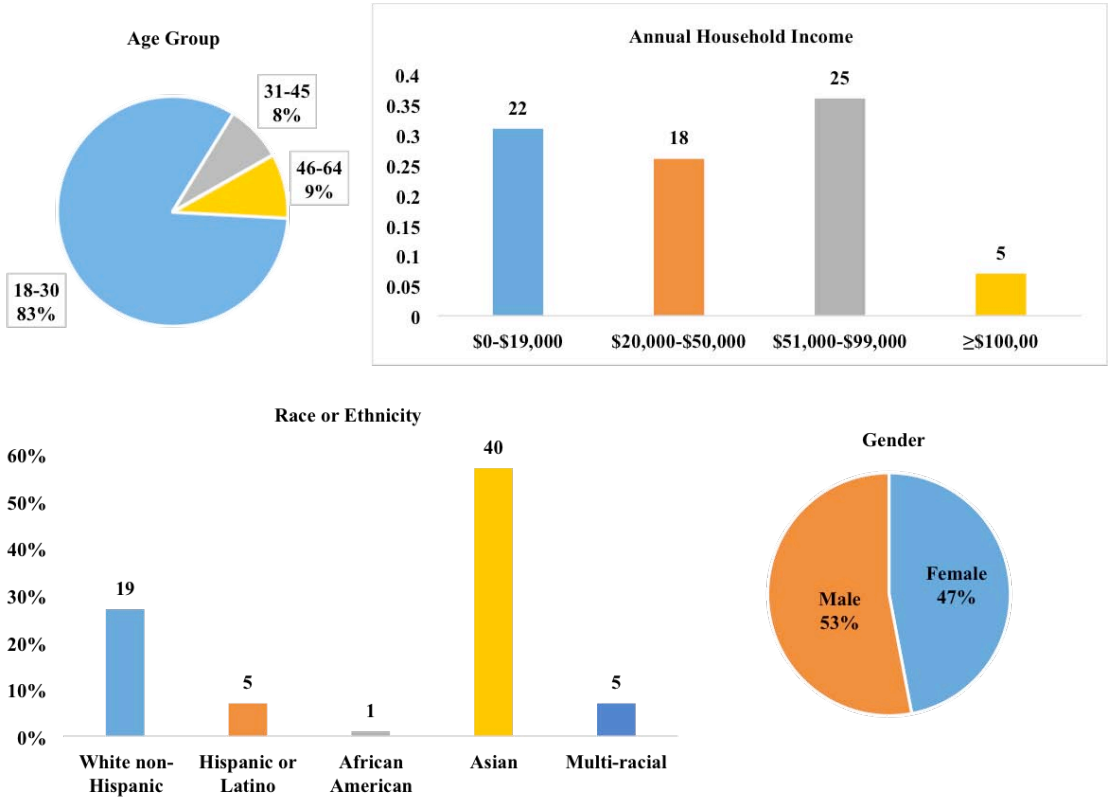


Figure 2. Demographics of participants

In addition to the demographics shown in Figure 2, information about education level, personal health status, weekly Internet use and health insurance status were collected. Participants were more highly educated than would be expected from a random sample of the US population[20]. Seventeen percent had a graduate or professional degree. When asked to describe health status, 14% said *fair*, 35% said *good*, 31% said *very good* and 20% said *excellent.* Four participants did not have health insurance. Regarding questions about the GUI and variety of choices presented, 94% of participants felt that the number of categories in *What am I sharing?* was adequate. However, 83% would like more granular control over sharing options, such as sharing mental health information with non-profit sponsored researchers but not for- profit sponsored researchers. Eighty-seven percent of participants were motivated to share in order to help other patients and contribute to science and research, and 44% would do it because they trusted UCSD and believed in the importance of the research being conducted.

Questions	Number of Participants	Some comments from participants
Are you more or less willing to share your data now that you had these choices?		
More	54[a]	"No change. I'm for research and would like to help either way. But, being explicit and providing this added information, makes me feel more at ease doing so."
Less	9	
Other	7	
Is it important to you to know whether your data are being shared with for-profit or non-profit institutions?		
Yes	48[b]*	"Slightly important because I would only want to share this information if I knew researchers were going to use it for well meaning purposes."
No	7	
I am indifferent	15	
If it were possible for you to know who is accessing your data, would you like to know this?		
Yes	62[a]*	"Yes, but more because of curiosity on what is going on with the research community, rather than actually feeling my privacy is being "invaded""
No	1	
I am indifferent	7	
Would you feel more comfortable sharing your information if you know who is accessing it?		
Yes	62[a]*	No comments
No	3	
I am indifferent	5	
If there were an option for you to control the sharing of biosamples, such as tissue, blood and urine, would you want to control this?		
Yes	34	No comments
No	10	
I am indifferent	26	
Is there anything else you would like to keep private in your medical record? (multiple choice)		
No	42	"DNA info" "potentially criminally-related illnesses or drug use" "STDs"
Chronic disease	5	
Acute disease	6	

Table 2. Selected survey question and responses. [a] $p < .001$ [b] $p < .05$.
* Indicates that this is a secondary hypothesis that was subjected to post-hoc analysis without correction for Type I Errors.

Discussion

Patients may not be aware of how their personal health information is being used for research. By presenting information and options, we sought to discover preferences for sharing health information. We hypothesized that participants would be more likely to share their health information if they were presented with choices, specifically for controlling categories of sensitive information. Having choices available did make participants more willing to share their data, and they expressed interest in keeping specific categories of information private. These results likely do not represent the opinions of all citizens, as recruitment was limited in scope. Since recruitment occurred on the UCSD campus, where most people are 18-30 years old and of Asian ancestry, the results may not generalize to the general population. Other important limitations include the fact that most of the participants had at least some college education and mostly were in *good* or better health status.

Furthermore, since this study was hypothetical, participants may choose differently than if they were actually making choices for their medical record. The results of the projected iCONCUR study on actual patients and with their preferences being implemented in data sharing for research should provide insights on those differences. During this study, the news regarding the NSA (National Security Agency)[22] and global security surveillance scandal was made public. Verbal comments from several participants of the study indicated that their decisions to share or not share their data were influenced by this event. Two distinct schools of thought were noted – some participants indicated that they were more willing to share because "the government has all our info anyways" and there was no longer a perceived benefit to keeping information private. Others stated they were less likely to share, as the government was potentially infringing upon their personal privacy and they would like control whenever possible.

Twenty-two percent of participants did not want to share health data with for-profit sponsored researchers. This finding was fairly consistent with the literature[10-12]. Although there were limitations in this study due to the small sample size, the sponsors of the research is important factor to participants. Concerns about sharing sensitive

categories of information were also consistent with the literature[4,9,18]. Some important new findings of this study are that participants appear to be more willing to share when given granular choices over what categories of information to share, as well as when they are given information about who is accessing their information. Many participants are interested in contributing to research but would like feedback about how their data are being used, and would like to be informed of the results from the research.

Acknowledgements

This study was supported by iDASH: Integrating Data for Analysis, anonymization, and SHaring (1U54HL108460). We would also like to acknowledge Mindy Ross, MD, MBA, and Robert El-Kareh, MD, MS, MPH, for classifying CDWR data into categories of sensitive information, Jihoon Kim, MS, for the statistical analysis of the study results, Claudiu Farcas, PhD, Michele Day, PhD and Hyeoneui Kim, PhD, MPH, RN, for feedback and Mona Wong for building the GUI used in this study.

References

1. Spriggs M, Arnold M, Pearce C, Fry C. Ethical questions must be considered for electronic health records. J Med Ethics. 38:535-9.
2. Ancker JS, Silver M, Miller MC, Kaushal R. Consumer experience and attitudes toward health information technology: a nationwide survey. J Am Med Inform Assoc. 2013;20:152-156.
3. Meslin EM, Alpert SA, Carroll AE, Odell JD, Tierney WM, Schwartz PH. Giving patients granular control of personal health information: using an ethics 'Points to Consider' to inform informatics system designers. Int J Med Inform. 2013;82(12):1136-1143.
4. Caine K, Hanania R. Patients want granular privacy control over health information in electronic medical records. J Am Med Inform Assoc. 2013 Jan 1;20(1):7-15.
5. Limb M. Nuffield council opens consultation on use of personal biological and health data. BMJ 2013;347:f6315.
6. Dixon WG, Spencer K, Williams H, Sanders C, Lund D, Whitley EA, et al. A dynamic model of patient consent to sharing of medical record data. BMJ 2014;348:g1294.
7. Sheather J, Brannan S. Patient confidentiality in a time of care.data. BMJ 2013;347:f7042.
8. Ipsos Mori/Association of Medical Research Charities. Public support for research in the NHS. 2001. www.ipsos-mori.com/Assets/Docs/Polls/amrc-public-support-for-research-in-the-NHS-ipsos-mori-topline.pdf. Accessed February 10 2014
9. Whiddett R, Hunter I, Engelbrecht J, Handy J. Patients attitudes towards sharing their health information. Int J Med Inform. 2006;75(7):530-541.
10. Willison DJ, Steeves V, Charles C, Schwartz L, Ranford J, Agarwal G, et al. Consent for use of personal information for health research: Do people with potentially stigmatizing health conditions and the general public differ in their opinions? BMC Med Ethics. 2009;10(1):10
11.Willison DJ, Swinton M, Scwatrz L, Abelson C, Charles C, Northrup D, et al. Alternatives to project-specific consent for access to personal information for health research: insights from a public dialogue. BMC Med Ethics. 2008;9(1):18.
12. Mamo LA, Browe DK, Logan HC, Kim KK. Patient informed governance of distributed research networks: results and discussion from six patient focus groups. AMIA Annu Symp Proc. 2013:920-929
13. Weitzman ER, Kaci L, Mandl KD. Sharing medical data for health research: the early personal health record experience. J Med Internet Res. 2010;12(2):e14.
14. Barrett G, Cassell JA, Peacock JL, Coleman MP. National surveys of British public's views on use of identifiable medical data by the National Cancer Registry. BMJ. 2006;332(7549):1068-1072.
15. Kaye J, Curren L, Anderson N, Edwards K, Fullerton SM, Kanellopoulou N, et al. From patients to partners: participant-centric initiatives in biomedical research. Nat Rev Genet. 2012;13(5):371-376.
16. Health IT policy committee, privacy and security tiger team, letter to David Blumenthal, Chairman of the Office of National Coordinator for Health IT. 2010 Aug.
http://www.healthit.gov/sites/default/files/hitpc_transmittal_p_s_tt_9_1_10_0.pdf Accessed Mar 5 2014.
17. Jaret P. Mining electronic records for revealing health data. New York Times. 2013 Jan
14. http://www.nytimes.com/2013/01/15/health/mining-electronic-records-for-revealing-health- data.html?pagewanted=all&_r=0. Accessed Oct 20 2013
18. Meslin EM, Alpert S, Carroll AE, Odell JD, Scwatrz PH. Points to consider in ethically constructing patient- controlled electronic health records. 2012. http://hdl.handle.net/1805/2936. Accessed August 20 2013.
19. Carr J. Recommendations regarding sensitive health information.
2010. http://www.ncvhs.hhs.gov/101110lt.pdf.Accessed October 25 2013
20. US Census Bureau. Educational attainment in the United States 2012.
http://www.census.gov/hhes/socdemo/education/data/cps/2012/tables.html. Accessed March 11 2014.
21. Ohno-Machado L, Banfa V, Boxwala AA, Chapman BE, Chapman WW, Chaudhuri K, et al. iDASH: integrating data for analysis, anonymization and sharing. J Am Med Inform Assoc. 2012; 19:196-201

22. The Guardian. Edward Snowden and the NSA files- timeline. 2013 July 5. http://www.theguardian.com/world/2013/jun/23/edward-snowden-nsa-files-timeline. Accessed March 5 2014.

Appendix 1

Survey questions

Learning about you

1. Which age group do you belong to?
    - 18-30
    - 31-45
    - 46-64
    - ≥65
2. What is your gender?
    - Male
    - Female
    - Unknown
3. What is your highest education level?
    - High school or less
    - Beyond high school or < 4 years of college
    - 4 year college graduate
    - Graduate or professional school
4. What race or ethnicity do you identify as?
    - White non-Hispanic
    - Hispanic or Latino
    - African American
    - Asian
    - Multi-racial
5. What is your annual household income?
    - <$5000
    - $5000-$14,999
    - $15,000-$19,999
    - $20,000-$49,999
    - $50,000-$59,999
    - $60,000-$99,999
    - ≥$100,000
6. How would you describe your personal health status?
    - Poor
    - Fair
    - Good
    - Very good
    - Excellent
7. How many hours per week do you use the Internet?
    - 1-10 hours/week
    - 11-15 hours/week
    - >15 hours/week
8. Do you currently have health insurance?
    - Yes
    - No

What is this study about?

9. How long will this study last?
    - One month
    - 6 months
    - 1 year
    - 2 years
10. Can you change your mind about sharing your information or participating in this study?
    - Yes
    - No
11. When your data is shared, will you be notified?

- Yes
- No

Your Suggestions for Improvement

12. Was the number of categories of information in "What am I sharing?" adequate?
    - No, it needed more categories [please explain in comment box]
    - Yes
    - No, there were too many categories [please explain in comment box]
13. Is there anything else that you would like to be able to keep private in your medical record? [multiple choice]
    - No
    - Chronic disease information
    - Acute disease information
    - Sensitive non-diagnostic information
14. If there were an option for you to control the sharing of biosamples, such as tissue, blood, and urine, would you want to be able to control this?
    - Yes, I would like to control the sharing of biosamples
    - No, I would not like to control the sharing of biosamples
    - I am indifferent
15. Do you feel more or less willing to share your medical information now that you had these choices?
    - More
    - Less
    - Other
16. What is your motivation for sharing your health information? [multiple choice]
    - Benefit future patients
    - Contribute to science and research
    - Trust in UCSD and a desire to contribute to the research they are doing
    - Establish a good relationship with UCSD
    - Other
17. Is it important to you to know whether your data is being shared with for-profit or non-profit institutions?
    - Yes
    - No
    - I am indifferent
18. The tool (GUI) does not allow different researchers to access different information from your medical record. For instance, the tool does not allow you to choose to share your mental health information with non-profit organizations but not to share it with for-profit organization. Would you like this option?
    - Yes, I would like to have the option to allow different researchers access to different information from my medical record.
    - No, I would not like to have the option to allow different researchers access to different information from my medical record.
19. Do you like to have control on what to share and with whom to share it?
    - Yes
    - No
    - Yes but not enough to go through the trouble of selecting options
    - Other
20. If it were possible to know who is accessing your data, would you like to know this?
    - Yes
    - No
    - I am indifferent
    - Other
21. Would you feel more comfortable sharing your information if you know who is accessing it?
    - Yes
    - No
    - I am indifferent
22. What would you like to know about the researchers who used your data? [multiple choice]
    - What kind of organization they belong to (e.g. a profit/non-profit organization, university or healthcare system)
    - What was the aim of their research?

- What papers were published using my data?
- What were the outcomes of their research?
- Other

23. How often would you prefer to be notified of updates about who is using your information?
    - Every time someone new uses it
    - Once a month even if there are no changes
    - Once a year even if there are no changes
    - Never, I can go to the site whenever I want to
    - Other
24. For each category of sensitive information (sexual and reproductive history, substance abuse, mental health, and genetic information), there was a paragraph explaining it and offering some examples. Do you feel like these paragraphs and examples were clear enough?
    - Yes, they were all clear enough
    - No, none were clear enough
    - No, some were unclear
25. Were the categories of sensitive information (sexual and reproductive history, substance abuse, mental health and genetic information) enough to cover your own preferences?
    - Yes, the categories were enough to cover my preferences
    - No, the categories were not enough to cover my preferences
    - Other
26. In this study, you are given the option of sharing or not sharing categories of information that are not considered sensitive. These include demographics, non-sensitive diagnosis information, laboratory and test results, medical procedures, and medications. Do you feel that these categories covered all types of information you want to control, or do you wish there were more categories for you to choose from?
    - Yes there were enough categories for me to control
    - No, I wanted more categories of information
    - Other
27. Did you have trouble understanding any of the information presented in the tool?
    - No, I was able to understand everything
    - Yes, I had trouble understanding parts of it
    - Yes, I couldn't understand it at all
    - Other
28. Did you have trouble understanding how to use the tool when selecting your choices on what to share and with whom?
    - No, I was able to understand how to make my choices
    - Yes, I had some trouble understanding how to make my choices
    - Yes, I couldn't make any choices because I didn't understand how to do it
    - Other

Automated Extraction of Family History Information from Clinical Notes

Robert Bill[1], Serguei Pakhomov, PhD[1, 2], Elizabeth S. Chen, PhD[4, 5], Tamara J. Winden, MBA[1,6], Elizabeth W. Carter, MS[4], Genevieve B. Melton, MD, MA[1, 3]
[1]Institute for Health Informatics, [2]College of Pharmacy, and [3]Department of Surgery, University of Minnesota, Minneapolis, MN; [4]Center for Clinical & Translational Science and [5]Department of Medicine; University of Vermont, Burlington, VT; [6]Division of Applied Research, Allina Health, Minneapolis, MN

Abstract

Despite increased functionality for obtaining family history in a structured format within electronic health record systems, clinical notes often still contain this information. We developed and evaluated an Unstructured Information Management Application (UIMA)-based natural language processing (NLP) module for automated extraction of family history information with functionality for identifying statements, observations (e.g., disease or procedure), relative or side of family with attributes (i.e., vital status, age of diagnosis, certainty, and negation), and predication ("indicator phrases"), the latter of which was used to establish relationships between observations and family member. The family history NLP system demonstrated F-scores of 66.9, 92.4, 82.9, 57.3, 97.7, and 61.9 for detection of family history statements, family member identification, observation identification, negation identification, vital status, and overall extraction of the predications between family members and observations, respectively. While the system performed well for detection of family history statements and predication constituents, further work is needed to improve extraction of certainty and temporal modifications.

Introduction

Family history information is essential for understanding disease susceptibility and is critical for individualized disease prevention, diagnosis, and treatment (1,2). With greater and more widespread adoption of electronic health record (EHR) systems and mandates with Meaningful Use Stage 2 to utilize structured family history functionality (3), there is an increasing opportunity to perform secondary analysis of family history information. Important applications include researching genetic influences between family history and disease, performing association mining of the EHR, and supporting public health research (4). For instance, family history can be used to calculate odds ratios and relative risks using genotypic and environmental interactions (5) to estimate the level of risk for certain diseases, as well as signal the need for further consultation (e.g., genetic counseling) or diagnostic workup (e.g., preventative health screening).

We have previously analyzed the representation of family history information in the EHR (6) and characterized family history free text comments in an EHR family history module (7). Most recently, we have expanded upon this work and have used multiple sources to develop a more comprehensive family history representation model (8). While two previous studies have described preliminary approaches for automated family history extraction from clinical texts (9,10), these approaches did not include functionality to classify family history from other non-family history texts, nor were high-level linguistic features like predication or linguistic chunking used. Currently, most established clinical natural language processing (NLP) systems (e.g., MedLEE(11–13) and cTAKES(14)) are primarily focused on extracting named entities such as diseases, medications, or procedures or contextual information (15). Family history information is frequently expressed via relations between named entities such as family members or diseases and also may contain contextual information such as certainty and negation, as well as information about vital status and age modifiers. We sought to incorporate family history functionality into an open-source NLP system, BioMedICUS (16) and to evaluate the performance of this module for family history extraction.

Methods

Figure 1 depicts a broad overview of the NLP pipeline and the process for its evaluation. The system was developed as part of our larger BioMedICUS NLP system that uses the Unstructured Information Management Architecture (UIMA) framework. UIMA, originally developed by IBM, is now an open source Apache document annotation framework that provides functionality to write sets of analysis engines (pipelines) that annotate and process unstructured text. Authoring UIMA-based analysis engines involves design and development of algorithms that read document text and add annotations to it in progressive stages through a pipeline where each analysis engine has access to annotations added to the common analysis structure (CAS) by upstream analysis engines.

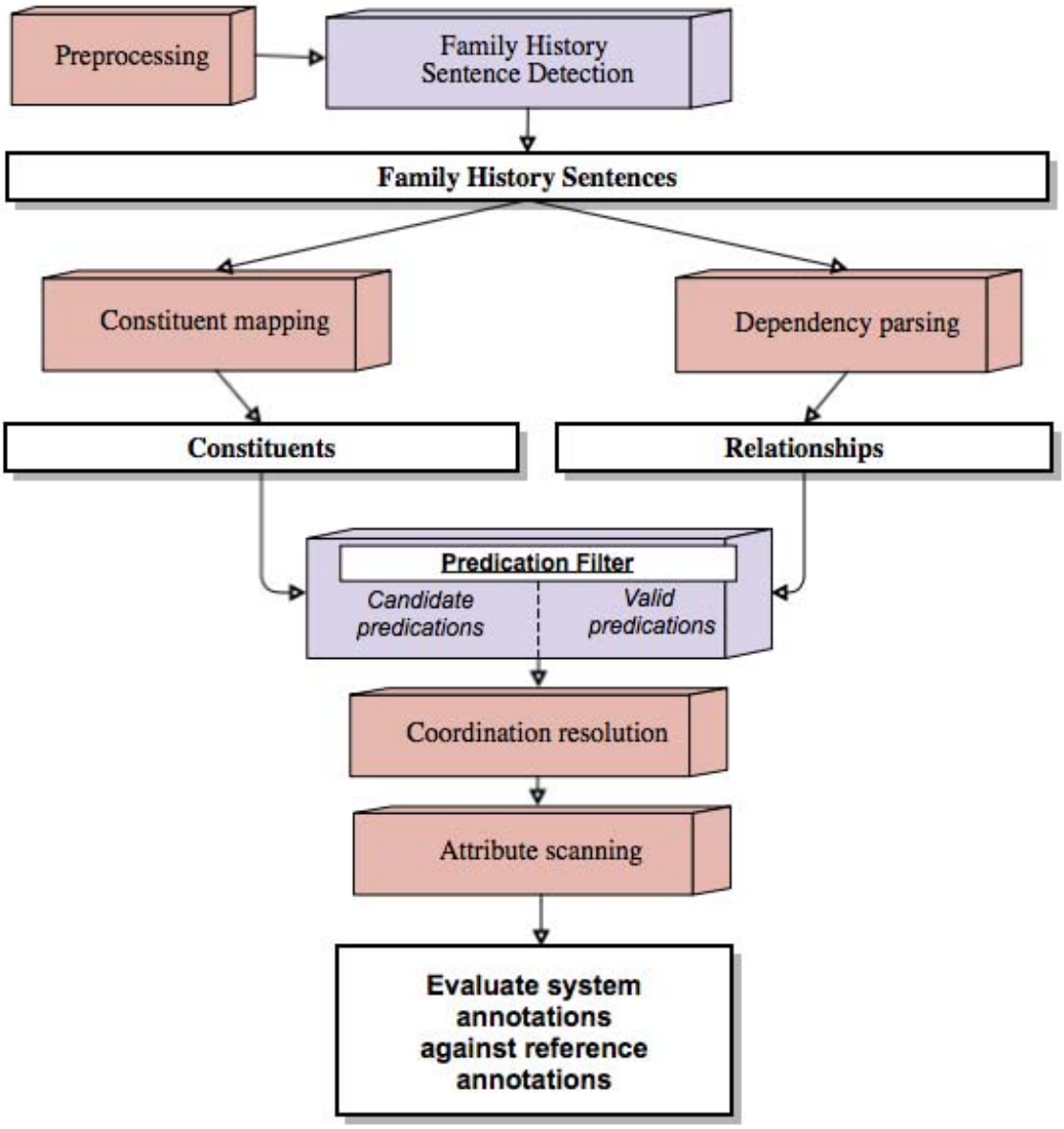


Figure 1. Overview of family history system and its evaluation

We used an approach to family history relationship extraction similar to the one developed by Rindflesch *et al.* for identification of semantic predications in biomedical literature (SemRep) (17). The NLP system shown in Figure 1 extracts family members and clinical observations as relationships where the relationship is usually identified by "indicator word(s)" (i.e., predication) appearing between the family member occurrence and clinical observation occurrence. After identifying the primary relationship, dependent phrases are then mined for attributes such as certainty, vital status and age. The pipeline was evaluated using an expert-based gold standard, as described below. Table 1 details each of the pipeline steps.

Corpora and Expert Annotation

The corpus used to develop and evaluate the family history NLP system was from a publicly available resource, MTSamples.com (18). For the MTSamples corpus, all History and Physical (H&P) notes, which were not behavioral health notes, were included (N=491). The family history pipeline was then developed using 329 MTSamples notes

(training set). As part of the evaluation, the model built for detecting family history statements was evaluated with 10 by 10 cross validation of the training set of notes. For this, the model was built with 90% of the notes and evaluated on the remaining 10%, with the remaining 10% being a different set of notes for each of the 10 iterations. Other portions of the pipeline using rules instead of models were evaluated on the entire training and test sets of notes. The system was then also evaluated on the remaining 162 MTSamples notes (evaluation set) as a more objective evaluation "hold out" set. The reference standard was developed by having annotators first identify family history statements in 20 notes to establish inter-annotator agreement prior to annotating the remaining set of MTSampels notes. After review of the general annotations completed in the first step, annotators added specific attributes. The specific attributes included family member, observation, side-of-family, vital status, certainty, and other temporal information. Table 2 shows an example set of annotations for the family history text for the reference standard.

Table 1: Family History Pipeline Components

Component	Description
Family History Sentence Detection	Determines if the current sentence contains a family history statement
Constituent mapping	Annotates the family history members from the HL7 clinical genomics family history model as family-member entities, and annotates all the disorders or procedures from SNOMED CT as observation entities
Relationship extraction	Identifies relationships between family member(s) and observation(s)
Coordination and list resolution	Transforms relationships that have lists of family members and/or observations into a list of relationships that specify a single family member and single observation. Also, adds attributes based on rules (e.g., adds "paternal-side" to relationships with "father" as the family entity)
Attribute scanning	Search patterns to identify the essential attributes of a relationship: vital status, certainty, age of diagnosis, side of family

Four informatics experts, including a physician, one PhD, one informatics graduate student, and one informatics researcher, provided manual annotations. Two of the experts also had previous experience in biomedical standard evaluation. Annotators arbitrated ambiguous sentences by discussing similar sentences to agree on presence of family history information. The resulting sets of annotations were provided in GATE (19) XML format and then converted to the UIMA XMI/CAS format for using in the UIMA pipeline developed for this study.

Table 2: Annotation Examples

Sentence 1: "He says his father might have died of heart disease."

Sentence 2: "Significant for epilepsy on the father's side of the family."

Sentence 3: "Mother with colon cancer but no other cancers."

	Family Member	**Observation**	**Vital status**	**Side of Family**	**Negation**	**Certainty**
Sentence 1	Father	heart disease	died			might
Sentence 2		epilepsy		paternal		
Sentence 3	Mother	colon cancer				
Sentence 3	Mother	other cancers			no	

Table 3 summarizes the annotations and the frequency of their occurrence in the corpus. If a family history statement included multiple clinical observations for a family member, then the UIMA annotators created all possible member-to-observation pairings. For example, the statement "*father died of stroke due to uncontrolled hypertension*" would result in the following pairs: *father-stroke* and *father-uncontrolled hypertension*.

The automated system was evaluated based on the accuracy with which it could extract these family member-observation pairings from text. Evaluation was challenged by a number of different complexities in the text, including the finding that many of the pairings were missing an explicit mention of a family member (e.g., "Significant for Alzheimer's.") The annotation for this statement does not include a family member; in this case, the automated system would then generate a pairing with a null value for the family member slot or position.

Table 3: Reference Corpus Annotations

Sentences	***Number***
Total Sentences	*23,155*
Family History Sentences	*284*
Predications	
Family member to observation predications	*364*
Attributes	
Side of Family	*15*
Family member	*417*
Observation	*745*
Vital status	*131*
Negation	*73*
Certainty	*55*

Family history sentence detection

Section and subsection headings contribute significant information when identifying statements of family history. For instance, short phrases such as "*Significant for asthma"* could be difficult to identify as family history statements when considering only the sentence. In these cases, the section heading, and sometimes the subsection heading provided the context necessary to improve accuracy in classifying family history statements. The current version of the BioMedICUS section detection module uses regular expressions to identify the section and subsection boundaries and headings. This is effective and efficient for corpora that have a limited set of consistent spacing and punctuation patterns that identify the sections and subsections such as the MTSamples corpus, although future versions could potentially benefit from augmented section heading tagging techniques. The heading label for sections and subsections are included with the sentence as predictors for family history sentence detection.

Using the section heading, subsection heading and sentence text as features together, sentences were classified as family history sentences. The system used a classification approach for family history statements detection using the SGD module in WEKA (20). WEKA is software for machine learning and predictive modeling, and the SGD module implements the stochastic gradient descent learning model for use in a support vector machine (SVM) (or

other linear models). N-gram-tokenized text is used as predictors to the SVM. This was compared to a simple lexical match that utilized all family words from the HL7 vocabulary RoleCode - PersonalRelationshipRoleType as part of the HL7 Clinical Genomics family history model (21). If a section or subsection heading contains the text "Family History," then all sentences within that section were then classified as family history statements. Only those sentences automatically determined to be family history statements were then evaluated further in the pipeline.

Constituent mapping and relationship mapping for core family history named entities and indicator words

The syntactic units that were candidates for the family history annotations were generally phrases found in either the HL7 Clinical Genomics family history model for family relation entities or in SNOMED CT as observation entities, which were extracted as described below. The family and observation entities were then linked together by an indicator word or phrase thus forming a predication relation. Indicator words are those words or phrases that signal or indicate a relationship between entities. They usually appear between the family entity and the observation entity; therefore, mapping indicator words was key to identifying family history statements. A lexicon of indicator words and phrases was constructed from the annotations provided in the corpus training set. Examples of these words are those that indicate possession or experience. The lexicon of possession words include *has*, *is*, *with*, and the tense variations. Experience indicators include *suffered a*, *died of*, *recovered from* and the tense variations of those words. Words and phrases in this indicator lexicon are also mapped during the mapping stage of the pipeline.

Together, the identification of the triples composed of the predication indicator, family member, and observation arguments, along with subsequent attribute extraction, formed the core functionality of the system, The SNOMED CT observation candidates were limited to only disorders and procedures using a disorder/procedure subset created by limiting the SNOMED CT concepts to those with semantic types defined by the UMLS Semantic Groups (22) as disorders and procedures. When searching for matches to family or observation in the corpus, only the longest matches (greedy matches) were kept. The mapping results for family member were also mapped to relative entities defined by HL7 family member codes.

Negation

For all sample statements that contained a negation term in the training corpus, negation occurred within a window to the left of an observation. Guided by this discovery, the negation detection implementation scans terms to the left of the constituent object in question. All negative results were then annotated. For example, phrases such as *no significant history* are annotated with the relationship between family and observation being noted as negated. Additionally, in the example, "*father, but not mother*", the negation component in the pipeline will add a negation attribute to only the *mother* entity of the phrase. The latter example also clarifies how scanning terms to the right of a constituent to determine their negation status can risk incorrect attachment of negations, such as the *father* constituent being identified as negated. Instead of attaching the negation to specific constituents of the sentence, the predication/relationship was annotated as having negation or no negation. Thus, the two sentences "*Neither parent has diabetes*" and "*Parents without diabetes*" would produce the same predications between parents and diabetes that have the same negation attribute.

Predication filter to convert family history constituents to valid relationships

As previously described, the family history annotation structure is composed of a relationship triple involving three entities: family member, observation and indicator word. The chunks on either side of the indicator word are assumed to be the concepts in the relationship. With initial exploration of the MTSamples training dataset, we observed using the family and observation syntactic units on either side of the indicator word or phrase was a reliable means of identifying entities defining a family history statement. Many of the family history statements in the training corpus include coordination in the family member and the observation. For example, the statement *Mother and father had diabetes and asthma* indicates four family history predications. The combinatorial considerations also include lists, such as *father, mother* and *uncle* or the disease list of *cancer, high blood pressure*, and *glaucoma*. These examples require coordination resolution to create all the necessary family history triples. The resolutions required in the previous examples are lexical. An additional issue that had to be addressed for family member chunks is that there were sometimes semantic lists created by statements such as *parents, three brothers*, and *all grandparents*. In order to manage these statements, a simple lexicon is used to lookup the appropriate combinations of family members for family member constituents that indicate multiple family members. The predication filter noted in Figure 1 removes all the candidate predications that are not valid because the dependency parse cannot confirm a relationship between the constituents and indicator.

At the end of this portion of the pipeline, a set of relationships, or triples, were generated and resolved -- each

having one family member, one observation, and one indicator per triple. The final modification to the relationship status is to evaluate the negation status of each entity. If any of the entities are negated, then the predication was negated. The training set did not contain instances of double negation; therefore, any negation in entities is sufficient to annotate the predication as negated. This defines only the relationship; therefore, subsequent steps in the pipeline extract attribute information from the indicator, and dependent phrases.

Attribute extraction and evaluation

Additional relationship attributes include vital status, certainty, age of diagnosis, and side of family. Some attributes are determined by the relationship itself, as in the example when the indicator chunk is an experiential frame such as *died of* (indicating the vital status attribute). Other attributes are found in dependent words or phrases. In order to accurately attach attribute words and phrases to the correct entity, the attribute extraction component used the Link Grammar dependency parser to identify dependent phrases. Each token in the dependent phrases is compared to a lexicon of terms that define attributes. Successful scans for certainty words, vital status words, and temporal references become attribute assignments. Figure 2 depicts an example sentence with the mapping from a sentence to chunks, followed by attribute assignment. Attribute results are reported as precision, recall, and F-score where the totals are cumulative across the corpus. The test corpus was evaluated as a whole as a hold out unseen set of notes.

Sentence: The patient states that both parents had heart disease and thinks that a maternal grandmother might have died of cancer.

Entities: familyEntity{both parents}
indicatorEntity{had},
observationEntity{heart disease}
familyEntity{maternal grandmother}
indicatorEntity{died of},
observationEntity{cancer}

Coordination: Rel1 = relationship{father, had, heart disease}
Rel2 = relationship{mother, had, heart disease}
Rel3 = relationship{grandmother, died of, cancer}

Attribution: sideOfFamily{Rel1, paternal}
sideOfFamily{Rel2, maternal}
vitalStatus{died of, Rel3}
certainty{might have, Rel3}

Figure 2. Relation and attribute identification

Results

Evaluation separated each of the stages of the pipeline. The first step of the pipeline, after general pre-processing such as tokenization and sentence segmentation, is the binary classification of sentences as a family history statement or not. This classification of family history sentences had good performance (Table 4), with the caveat that keyword detection was equally effective as the SVM approach, and that the performance depended upon the accuracy of sentence detection in general as well as accurate section heading detection. Sentences outside of the family history section were detected; however, section headings proved to be significant because statements within the family history section were frequently abbreviated with context derived only by the heading. For example, many family history statements included only the word "*noncontributory*," or observation lists such as "*diabetes, heart disease*." Section headings are the only means of classifying these abbreviated sentences, but family history statements in other sections do not appear in abbreviated forms. The results for accurately classifying sentences as family history statements include sentences within and without the family history section of the clinical note. A high number of false positive classifications of a sentence as a family history statement occur among sentences outside the family history section, particularly statements regarding family dynamics that include many family member terms.

Once sentences that contain family history statements are identified, the subsequent step is identifying the family history constituents, including the family member, observation and the indicator word used to identify predications. The different constituents of family history statements were evaluated using the set of sentences known to contain family history statements (Table 5). We found family member detection results of 90.8% precision and 94.0% recall (F-score 92.4%) for a sample of text containing family history statements and 177 family member names. Observation detection had 80.2% precision and 85.7% recall (F-score 82.9%) on the evaluation set containing 325 annotated clinical observations. Following the identification of constituents, the next step is extracting the relationship between constituents to form predications. Predication detection performance achieved an F-score of 65.1% with precision of 70.3% and recall of 60.6% in the training corpus, with lower results for the test corpus, as shown in Table 4. Negation performance showed 49.5% precision and 68.2% recall (F-score 57.3%) over a sample of family history statements that contained 37 negation annotations. The performance of vital status and age of death are also summarized in Table 5.

Table 4. Evaluation of family history NLP module

	Training Corpus Results (10 by 10 cross validation)			Evaluation Test Corpus Results		
	Precision	Recall	F-score	Precision	Recall	F-score
Sentence detection	84.5	60.0	70.2	83.2	55.8	66.9
Predication detection	70.3	60.6	65.1	66.3	58.1	61.9

Table 5. Evaluation of individual components on family history statements

	Precision	Recall	F-score
Family detection	90.8	94.0	92.4
Observation detection	81.1	85.7	83.3
Negation	49.5	68.2	57.3
Vital status	96.3	99.2	97.7
Age of death	94.0	87.5	90.6

Discussion

In this study, we constructed a family history UIMA-based NLP module, specifically an annotation engine that classifies sentences that contain family history statements and a pipeline to extract important information about family history from clinical texts. The novel portion of the study was the system's conversion of narrative text into a family history data type for each family history statement with the family-member plus observation in the clinical note as the core relationship and then leveraging the use of predication in our NLP module. We found that the initial classification of sentences depended significantly on section information and, as such, the success of classification was dependent on accurate section detection, sentence boundary detection, and constituent mapping (family members and clinical observations). Individual sentences about family history frequently were found to have insufficient information to classify them without considering the section headings. Statements such as "*noncontributory*", "*unknown*", and "*nonsignificant*" are poor predictors for family history classification by themselves. Sentence segmentation errors also presented challenges for accurate detection of family history statements because the erroneous segmentation can be key in separating family member and observation entities that otherwise would be part of the same family history relationship. Furthermore, incorrect clinical and family mappings contributed to errors in classifying family history sentences.

Overall, we found that the core functionality of extracting the triple of indicator word or phrase, family member entity, and observation entity had reasonable performance. We also identified mechanisms for future system improvement by differentiating patient history, social history and family history early in the pipeline. The family member and observation detection results lag the performance reported by Goryachev *et al.* (10) and Friedlin *et al.* (9) although these studies were more limited in their evaluation. For instance, the former achieved 93% sensitivity and 97% positive predictive value; however, comparability is limited because they used a broader classification of family member. The Friedlin study was a limited report (poster) describing a potential family history system with limited results. The work by Goryachev *et al.* is more directly comparable and achieved higher reported results with family member detection of 85.12% precision and 86.93% recall; observation detection of 96.30% precision and 92.86% recall; and correct family member assignment to diagnosis of 92.31% precision and 92.31% recall. The previous work limited 'diagnosis' to 8 UMLS semantic groups as opposed to the 19 UMLS semantic types encompassing the disorder and procedure semantic groups used in this study. The system also did not address other modifications such as vital status, age of death, and certainty.

The lower performance of family history sentence detection results with the test set compared to training set may indicate some over-fitting to the training data. An additional limitation of this study is that the effectiveness on different sets of notes is unknown, particularly if section heading detection is troublesome. The mapping process of observations was limited to SNOMED CT observations and procedures, and that appeared to be sufficient for the corpora used in the tests. The largest influence on classifying and mapping statements was the section detection. Section headings supply significant information in identifying family history statements, and an important step to our future work will be improving the section detection phase and concept mapping phases of this system. We also observed the following pattern of errors when classifying sentences: the longest and shortest sentences were misclassified most often. The reason for the misclassification of the longer sentences was due to the errors in sentence segmenting, and misclassification of the very short sentences was due most often due to exclusion of either the family member or the observation entity. We also observed a high rate of false negatives for predication detection attributable to errors in finding the arguments for the predication. Furthermore, in many cases the misclassification was due to presence of family related information in statements reporting social rather than family history (e.g., *Mother was a smoker*). Future work with social history extraction may be helpful in reducing these types of errors as well as extending family history functionality for extraction of additional attributes such as temporal elements to improve the extracted family history information at a more detailed level.

Conclusion

Detecting family history statements and mapping them to regularized annotation structures holds promise for information extraction from clinical notes for important downstream uses. The results of our pilot study show that extracting family history information from clinical notes is a complex and challenging task that lends itself to NLP approaches developed for relation extraction. Specific challenges consist of a number of hard NLP problems including resolution of coordination and co-reference that, in turn, rely on the accuracy of lower-level processes such as sentence and phrase segmentation and negation detection. In this pilot, we were able to achieve reasonable performance and to identify a number of areas for improvement. Our next steps will include further refinement of both lower-level and higher-level components.

Acknowledgements

The National Institutes of Health (1 R01 LM011364-01 NIH-NLM, 1 R01 GM102282-01A1 NIH-NIGMS, U54 RR026066-01A2 NIH-NCRR) and Clinical and Translational Science Award (8UL1TR000114-02) supported this work. The content is solely the responsibility of the authors and does not represent the official views of the National Institutes of Health.

References

1. Guttmacher AE, Collins FS, Carmona RH. The family history--more important than ever. N Engl J Med. 2004;351:2333–6.
2. Dick R, Detmer DE, Steen EB. The Computer-Based Patient Record: An Essential Technology for Health Care, Revised Edition [Internet]. NAP. 1997. Available from: http://www.nap.edu/catalog.php?record_id=5306
3. Stage 2 - Centers for Medicare & Medicaid Services [Internet]. [cited 2014 Feb 20]. Available from: http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/Stage_2.html

4. Kukafka R, Ancker JS, Chan C, Chelico J, Khan S, Mortoti S, et al. Redesigning electronic health record systems to support public health. J Biomed Inform. 2007;40:398–409.
5. Ottman R. Gene-environment interaction: definitions and study designs. Prev Med (Baltim) [Internet]. 2010;25(6):764–70. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2823480&tool=pmcentrez&rendertype=abstract
6. Melton, Genevieve B and Raman, Nandhini and Chen, Elizabeth S and Sarkar, Indra Neil and Pakhomov, Serguei and Madoff RD. Evaluation of family history information within clinical documents and adequacy of HL7 clinical statement and clinical genomics family history models for its representation: a case report. J Am Med Informatics Assoc. 2010;17(3):337–40.
7. Chen ES, Melton GB, Burdick TE, Rosenau PT, Sarkar IN. Characterizing the use and contents of free-text family history comments in the Electronic Health Record. AMIA Annu Symp Proc [Internet]. 2012;2012:85–92. Available from: http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=prem&NEWS=N&AN=23304276
8. Chen ES, Carter EW, Winden TJ, Sarkar IN MG. Development of a comprehensive family health history information model. AMIA Annu Symp. 2013;
9. Friedlin J, McDonald CJ. Using a natural language processing system to extract and code family history data from admission reports. AMIA Annu Symp Proc. 2006;925.
10. Goryachev S, Kim H, Zeng-Treitler Q. Identification and extraction of family history information from clinical reports. AMIA Annu Symp Proc. 2008;247–51.
11. Friedman C. A broad-coverage natural language processing system. Proc AMIA Symp. 2000;270–4.
12. Friedman C, Hripcsak G, Shagina L, Liu H. Representing Information in Patient Reports Using Natural Language Processing and the Extensible Markup Language. J Am Med Informatics Assoc. 1999;6:76–87.
13. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. J Am Med Informatics Assoc. 2004;11:392–402.
14. Savova, Guergana K and Masanz, James J and Ogren, Philip V and Zheng, Jiaping and Sohn, Sunghwan and Kipper-Schuler, Karin C and Chute CG. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Informatics Assoc. 2010;17(5):507–13.
15. Chapman WW, Chu D, Dowling JN. ConText: an algorithm for identifying contextual features from clinical text. Proc Work BioNLP 2007 Biol Transl Clin Lang Process [Internet]. 2007;81–8. Available from: http://dl.acm.org/citation.cfm?id=1572392.1572408\npapers2://publication/uuid/48523D28-904A-48D8-AF74-BE9CFD2904F4
16. BioMedICUS [Internet]. Available from: https://bitbucket.org/nlpie/biomedicus
17. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: Interpreting hypernymic propositions in biomedical text. J Biomed Inform. 2003;36:462–77.
18. Transcribed Medical Transcription Sample Reports and Examples - MTSamples [Internet]. [cited 2014 Mar 4]. Available from: http://www.mtsamples.com/
19. Cunningham H, Tablan V, Roberts A, Bontcheva K. Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. PLoS Comput Biol. 2013;9.
20. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA Data Mining Software: An Update. ACM SIGKDD Explor Newsl [Internet]. 2009;11:10–8. Available from: http://dl.acm.org/citation.cfm?id=1656278
21. Shabo, Dr. Amnon D, Hughes, Kevis S. D. HL7 Clinical Genomics Family History Model Abridged. 2007.
22. McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. Stud Health Technol Inform. 2001;84:216–20.

The EHR's roles in collaboration between providers: A qualitative study

Dian A. Chase, RN, FNP, MSN, MBA[1], Joan S. Ash, PhD, MBA, MLIS[1], Deborah J. Cohen, PhD[1], Jennifer Hall, MPH[1], Gary M. Olson, PhD[2], David A. Dorr, MD, MS[1]
[1]Oregon Health & Science University, Portland, OR; [2]University of California Irvine, Irvine, CA

Abstract

Objective: Examine how the Electronic Health Record (EHR) and its related systems support or inhibit provider collaboration.

Background: Health care systems in the US are simultaneously implementing EHRs and transitioning to more collaborative delivery systems; this study examines the interaction between these two changes.

Methods: This qualitative study of five US EHR implementations included 49 interviews and over 60 hours of provider observation. We examined the role of the EHR in building relationships, communicating, coordinating, and collaborative decision-making.

Results: The EHR plays four roles in collaboration: a repository, a messenger, an orchestrator, and a monitor. While EHR performance varied, common themes were decreased trust due to poor quality documentation, incomplete communication, potential for increased effectiveness through better coordination, and the emerging role of the EHR in identifying performance gaps.

Conclusion: Both organizational and technical innovations are needed if the EHR is to truly support collaborative behaviors.

Introduction

For over a decade, experts have agreed that more collaborative, team-based care will be required to meet the increasing burden of chronic disease.[1,2] Unlike the acute care issues that dominated medical practice in the twentieth century, treating chronic disease in the twenty-first century will require multiple visits to providers in different disciplines. Not only will increasingly specialized medical expertise be required, but chronic disease treatment also involves changing lifestyles and navigating a complex web of treatments. Different health care professionals with different skills in different locations will need to collaborate to provide a cohesive care team. This increased collaboration is likely to constitute a disruptive change in the delivery of healthcare services.

In its broadest sense, collaboration simply means "to work jointly with others or together especially in an intellectual endeavor." [3] This definition may be too broad for use in healthcare situations; we found twenty-seven different definitions for healthcare collaboration in the literature. These definitions vary from a willingness to work with another party in any manner [4], to teamwork (defined roles and a common goal) [5], to shared decision-making.[6] For this study we adopted a taxonomy of collaboration behaviors based on the works of William Clancey[7,8] and Eduardo Salas [9] who suggest that collaboration includes the behaviors of trust and respect, communications, coordination, and adaptive collaboration. In table 1, we list the four behaviors and identify how the behavior can support a more effective healthcare process.

Table 1: the collaborative behaviors and their benefits

Collaborative Behaviors	Benefits
Trust and respect: willingness to rely on work of others	Less need to repeat diagnostics and procedures, more willingness to hand off or delegate
Communication: information flow, contextual background, understanding	Increased awareness and understanding, less mistakes due to missed data or context
Coordination: managing the timing and order of activities	More effective processes and increased efficiency in workflows
Adaptive Collaboration: changing the actual work content, tailoring solutions	Increased understanding across disciplines, when needed provides ability to tailor plans to meet patient circumstances

Team-based models of care based on collaboration have been implemented with improved results in primary care [10, 11], intensive care units [6], and operating rooms.[12] In the primary care setting, one prominent move towards collaboration and team-based care is the implementation of the Patient Centered Medical Home (PCMH). The configuration of the PCMH varies from system to system [13, 14] , but it consistently includes increased teamwork and collaboration within the clinic. In many cases it also includes improved collaboration with providers outside the clinic with whom they have working relationships (within the "Medical Neighborhood"). This is a disruptive change with significant effects on how care will be provided in the United States.

Another disruptive but potentially positive change is the introduction of the electronic health record (EHR) and its internal and related functions – computerized provider order entry (CPOE), clinical decision support (CDS) and health information exchanges (HIE). The EHR was originally designed as an electronic implementation of the paper chart.[15] Supporting collaboration (other than improved legibility) was not one of the original functional goals of this tool. [16] As the EHR has been deployed, the implementation has resulted in many examples of outcomes that are different from those anticipated when the systems were first designed. These are referred to as "unintended consequences." [17] Unintended consequences are not necessarily negative. For example, the EHR is being used to support care coordination, a process not included in the original vision of a documentation process to support medical decision-making. [16] The use of the EHR to support collaboration is, in many ways, a collection of these unintended consequences.

The relationship between the EHR and collaboration behaviors is evolving. This has been understudied. This paper presents the first multi-system qualitative study to delve deeply into the nature of this relationship and the roles the EHR plays.

Methods

Purpose: This qualitative study examines how the EHR and related systems affect the collaboration behaviors among providers in five systems that are implementing Patient Centered Medical Home delivery models.

Methodology: We used a modified Rapid Assessment Process [18,19] for data collection, and a grounded theory approach[20] for analysis. At the end of each site visit we summarized our initial themes and reviewed them with the site sponsor to verify our interpretations. Each aspect of the method is described below.

Sample: We studied five leading-edge multi-site organizations with EHR installations in the United States that were in the process of implementing their PCMH. These organizations were participating in SAFER project (Safety Assurance Factors for EHR Resilience [21]), which focused on developing guidelines for the safe implementation and use of EHRs and related systems. Sites were purposively selected to provide a cross section of successful EHR implementations in medium to large healthcare delivery organizations. As such they enabled a broad view of the potential advantages and challenges for collaboration in state of the art systems using an EHR. The characteristics of the sites are summarized in table 2. The visits occurred between May and November of 2012.

Table 2: Sites visited

	Location (US)	Structure	Number of physicians/providers	EHR
Site 1	Southeast	Augmented family practice/for profit	50 to 100	Centricity
Site 2	Mid Atlantic	Integrated System/ not for profit	More than 1000	EPIC
Site 3	Midwest	Community Health Center	Less than 50	Centricity through service provider
Site 4	Midwest	Community Health Center	50 to 100	Centricity through service provider
Site 5	Northeast	Academic Integrated Health System	More than 1000	Proprietary/ "Best of Breed"

Within each site we identified key individuals who were involved in the implementation, use, and monitoring of the EHR. These individuals included providers, clinical leaders, and informaticists. We then worked with our sponsor

at the site to schedule interviews with these individuals and field observations of providers working with patients using the EHR and other technology tools. The length of our visits varied from three to five days. The deployment of a multi-disciplinary team (see below) and a rapid assessment process [18] enabled us to reach saturation within a limited time period.

The data used for this collaboration study consisted of 49 interviews and 60 hours of field observation. This was selected from the larger SAFER project by focusing on how providers collaborate and communication with each other; over 500 pages of transcripts and field notes were annotated. The primary author and another analyst independently selected the interviews and observations included in this study from the larger set of interviews and field notes (over 2000 pages) gathered for the SAFER project.

IRB approvals: The Institutional Review Boards at Oregon Health & Science University and each of the sites approved the study.

Team: A multi-disciplinary team with experience in interviewing and observation conducted each visit. At a minimum, each site visit included a combination of professional qualitative researchers, informaticists and clinicians; several visits also included an expert researcher in communications and a human factors expert. The team met daily to debrief and prepare for the next day's encounters.

Collection Methods: During each site visit, we conducted field observation and semi-structured interviews. The semi-structured interview guides for this segment of our study focused on how the interviewees (or those working for the interviewees) collaborated with other professionals both within and without the clinic, the role the EHR played in facilitating or inhibiting the transaction, and implementation effects. All interviews were tape recorded and subsequently transcribed with the consent of the interviewees.

For the field observation, teams of two to six trained observers went to each site. We used a template for field observations which covered broad categories of foci of interest; each observer was responsible for gathering information on topics outlined on the template. Observation periods were three to eight hours, during that time we typically shadowed a provider to see how she used (or did not use) the EHR. In addition to shadowing providers, we also observed the work flow and meetings within a unit, and in one case we followed a patient. We were particularly interested in both safety issues (for the SAFER study) and how and when they collaborated with other providers (this study). Each day we documented their observations and then discussed them with the other researchers at a daily debriefing.

At the conclusion of each site visit, we prepared a summary of findings for the entire SAFER visit and met with our site sponsors and other leaders from each organization to confirm the veracity of our data.

Data Analysis: For the data analysis, we familiarized ourselves with the data, generated initial codes, then searched for and consolidated themes based on the codes. We produced reports consisting of quotations from interviews and field notes; these were translated into results based on "sense-making" sessions with key team members.

Coding: Two of the authors independently coded data for each site visit using NVivo 10. (22)Data tagged included portions of interview transcripts, field observation notes, and written artifacts collected in the site visits. After each site visit was coded, the analysts met to review codebook categories (codes) and relevant data elements. Discrepancies were analyzed and reconciled, and a new version of the codebook was prepared. This codebook was then used by both analysts for the analysis of the next site visit's data. New codes were added and old data elements reclassified after the analysis of each visit.

Theme generation: Following the initial coding, two of the authors met several times to identify emerging preliminary findings and implications from the analysis. During this period of reading and rereading (Miller and Crabtree's immersion phase[23]) the primary author drafted "memos" that helped her bring together key ideas.

Sense-making: Other co-authors with subject matter expertise in informatics, communications, and clinical delivery systems reviewed the themes during "sense-making" sessions. In these sessions, the team reviewed and refined findings, working to organize the findings, and sought to make connections between what we were seeing in these data and the current literature. Miller and Crabtree[23] call these steps in the analysis process the organizing and connecting phases. During these sessions, it became apparent that the collaboration models of Clancey and Salas [8,9] would provide a useful practice lens for organizing the emerging findings.

The data analysis resulted in 49 initial codes related to collaboration, and these categories were subsequently combined into 17 themes. These themes, in turn, were related to four roles played by the EHR discussed below.

Results

The four roles: We found the EHR and related systems play four distinct roles in collaboration: Repository, Messenger, Orchestrator, and Monitor. These roles are summarized in Table 2 below. The repository role was the most established at all sites; in all of the systems there was also a robust role for the messenger role. The degree of sophistication in the orchestrator role varied significantly from system to system. All of the sites we visited were actively working to improve and implement the EHR's performance in the monitor role.

Table 2: Four identified collaboration roles for the EHR

Role	Purpose	Examples
Repository	Contain all of the quality, accessible data needed for use by healthcare providers	Encounter and phone notes, lab and imaging results, and information from providers outside the system. Input may be by providers, by scanned documents, and increasingly, by vendor specific EHR to EHR communications (e.g. CareEverywhere) and Heath Information Exchange (HIE) mediated communications.
Messenger	Enable information transfer and communication between providers also between providers and other members of the healthcare team	Transmitted copies of encounter notes, pinned notes and flags sent to other providers, secure email, pop-ups, broadcast messages, clinical paging.
Orchestrator	Ensure that the right person is doing the right thing at the right time for the patient	Input templates that drive workflows, the use of standardized order sets, bundles and smart ordersets that implement best practice algorithms. Tickler messages generated by the system.
Monitor	Identify care gaps for patients and populations; provide a benchmark for measuring the performance of providers and teams	Registries, data warehouses and analytic tools, reporting tools for use by the clinic teams, dashboards, incentive performance displays.

In our interviews and observations, we found that the EHR can have varied, sometimes conflicting effects on the four collaboration behaviors (trust and respect, communication, coordination, and adaptive collaboration).

The EHR and collaboration behaviors: Each of the four roles had a primary effect on a single collaboration behavior (figure 2, below). They often had broader effects. One aspect of the EHR can and did affect several different behaviors at once (e.g. a lack of trust due to poor data quality cascaded throughout all four behaviors), but we have tried to avoid repetition.

Repository ←→ Trust and Respect
Messenger←→Communication
Orchestrator←→Coordination
Monitor ←→Collaboration

Figure 2: Each roles of the EHR had a primary effect on one collaboration behavior

Repository

The EHR, in its repository role had mixed effects – primarily on trust and respect. As provider told us, the major advantage to the EHR was that "*it's all there*" Because it is all there, the EHR can provide proof that one discipline can trust another: "*And our physicians saw [the data from the EHR] and realized 'wow, that really works… I can reliably delegate to my nurses*." But other times poor data quality would reduce their trust in other providers: "*This note that's 15 pages and it looks like they spent two weeks with the patient when the reality is they spent five minutes … and everything in there is either a fabrication and or was correct two years ago but not today*." Providers also complained about lack of context –the patient's "story" could not be adequately told with a series of check boxes.

A lack of quality in the notes can result from many factors. Time pressures are often a contributing factor: "*there is a trade-off between patient safety… and efficiency.*" "*If you are a busy PCP (primary care provider) with …one*

minute to prepare… you don't do it as thoroughly." It is not clear, however, how much this decrease in perceived note quality is simply an artifact of increased note visibility: "*people who wrote good notes on paper would write good notes in the electronic world, and people who wrote lousy notes …would write lousy electronic [notes]*."

As a repository, however, the function is not just to store data, but to make it available, When a provider could not find a specialist's note*, she became frustrated and began to blame the other provider for poor communications. We later found the note, but finding the note involved paging through three screens of poorly indexed lists of chart notes, consultant's reports, and laboratory data* (from our field notes). The ability to find necessary information is a key element in data accessibility.

The repository role also has an effect on the other collaboration behaviors. Although it did not guarantee good communications, it provided a common dataset that could speed and simplify communications. This was true, however, only when both parties had access to the same EHR. Commonly, we found that a copy of a chart note would be sent within a referral or other inter-provider communication; this provided the basis for other communications, but this often this transfer was not enough by itself, more contextual data was needed. For example, we were told that the radiologists in one system told us that the primary reason for poor implementation was the poor quality of the contextual data provided. The common data set also facilitated collaboration in multidisciplinary team meetings: "*As she describes the case the psychiatrist is displaying parts of the patient's history and progress notes from the EHR on the screen."* [from our field notes]

The repository function can improve coordination through increasing awareness: "*you can see a list of items that you need to complete for the patient. The nurse can see it, the physician can see it*." This increased awareness is, however, a two edged sword – "*now we are unearthing scope of practice issues … that we didn't have to deal with before*." At one organization a pharmacist could no longer complete a medication order with default parameters if they couldn't reach a provider, at another site the transition to an EHR unearthed the existence of hitherto unknown "personal order sets." These variations from standards were much more visible with the EHR than with paper records.

Messenger

The EHR and related clinical systems have significantly expanded the number of communication channels available to providers. Now, in addition to in-person, analog written, and phone/voicemail communications, providers can transmit information and communicate with other providers using secure email, clinical messages using a paging system, messages within the EHR (which can be attached to patient records), and pop-ups and general "broadcast notices." This multiplicity of channels can result in the choice of channels that are easier to use but providers considered less effective. In our visits, we were consistently told that sending a copy of the encounter note was the most frequent means of communication. And it is easy to do. As the one provider saved his encounter note, a routing box popped up and he clicked the name of the primary care provider to automatically send a copy: "*I just write a brief note and click.*" This was not uniform, there were also variations in practice between providers within the same healthcare settings. Some were more inclined to pick up the phone; some used paging or email.

In its role of messenger the EHR and related systems affected all of the collaboration behaviors, but its primary effect was on communication. While it could speed information transfer, it could reduce communication by inhibiting feedback. Asynchronicity facilitates information transfer by allowing EHR-related communications to bridge time and space; to facilitate providers messaging others on different shifts or in different offices or clinics. It also can increase awareness: *we know he's checking the vitals, he'll come when he needs to*."

But there are downsides to the default asynchronicity of information transfer. The ability to bridge time also came with a draw back – the lack of simultaneity made it difficult to give feedback and clarify, which could take more, rather than less time. Across all of the sites we heard: "*we call it brain freeze where with the technology implementation,* ***they think that they can no longer talk to one another."*** We also heard about the "communication illusion", when providers thought they were communicating, but weren't. Often a provider would send another a copy of their note, and assume the second would continue the plan. But the second provider might not find the plan or might misunderstand it due to a lack of contextual data. Delivery failure – particularly for lab results *("I've seem systems where 30% of the lab results weren't delivered to the right person"*) continues to be a problem. This was a problem that all of the sites we visited were trying to correct: "*You need to have that person communicate with the primary doctor …. and then have [the primary] them assume that responsibility and acknowledge it.*"

Orchestrator

The use of templates, smart order sets, and bundles has the potential to increase trust and respect, facilitate information transfer, and significantly improve coordination. The primary effect of the orchestrator role was, understandably, on coordination behaviors. Leaders at all of the sites were trying to use the EHR to improve quality by encouraging or 'orchestrating' "*the right person doing the right thing at the right time.*" Often this mean having someone work at "*the top of their license*." The degree of sophistication shown varied from site to site; for some it was templates that encouraged a given workflow, for others it was sophisticated smart order sets or bundles that created prompts and initiated workflows based on patient circumstances.

These smart order sets/bundles existed at more advanced sites only for the most prevalent ten to twenty-five conditions. "*We have lots of initiatives … most of them boil down to sophisticated checklists*." At one site they used "*automation to manage the preventative care needs of 220,000 patients … delegation of tasks [to non-physicians] as well as protocol*." As one of our key informants there told us, this does not happen well right after EHR implementation: "*We figure it takes about three months for the dust to settle…and [then we say]: What can we improve about the process? About the EHR? …The most valuable thing that happened … was that throughout the organization there are people who, when we say process redesign, say 'Okay, when and where?'"*

With smart order sets and bundles, there can be automatic dissemination and notification of results and follow-up. "*We look at [who]… can perform the task. The alerts fires to them and only that person. … We make it actionable so … they can address the alert at that time. … The nurse can see it. The physician can see it. And then, further in the work flow …, if they did not address the alert, it will display again*." The ability to generate tickler (reminder) notes based on a pre-agreed plan of care is a powerful communication tool within the orchestrator role.

Unfortunately, even the best thought out and planned bundles won't work if the providers don't accept the process. "*If we ever got any of ours 50 percent [of the new processes] accepted, I don't know what we'd do. We would be so excited*." One of the challenges for implementing EHR mediated coordinated care is getting physicians to trust in others: "*our physicians have been trained that really the buck stops with you. … [to offset this] we pilot everything…and then we show them it works.* Time is also an issue. All of the providers we met with felt time and productivity pressures, making them resistant to additional tasks: "*One the statements that can lead to failure is 'let's make the doctor do it …. Regardless of topic, I'll give you about a 30% chance … [of getting it to work when assigned to the doctors]."*, and "*you can have a great tool for asthma, but if it takes ten minutes to do it's not gonna happen*."

For many of the providers delegation helped relieve the time pressure. Care plans have long been a means of facilitating collaboration and team work. A more subtle point is the orchestrator role requires consolidating fragmented care plans, about which one participant noted: "*A human being should only have one care plan. This is a legitimate IT role."*

The EHR can only do so much to support collaboration, however; at every site we were told that when the situation was complex, they needed to meet with or "*pick up the phone and call*" another provider. One clinical leader told us that, as far as collaboration was concerned, "*the computer is giant hinder[ance]*" When collaboration was important, so was the person to person contact – either a warm handoff, a face to face meeting, or a phone call.

Monitor:

These systems of care were all striving to improve quality. With an EHR in place, data can be collected and analyzed to reveal care gaps. Several of the systems we visited were using these data to evaluate the performance of teams and individual providers; these data were also used to determine incentives and guide compensation decisions. By providing dashboard performance summaries to teams, and individual results to providers, systems were able to give feedback that could, in turn, facilitate performance improvement. It was in this role that we found the widest variation in system capabilities. None of the sites were completely satisfied with their EHR and the associated data warehouses that they were building to facilitate analysis. Given that, however, the capabilities ranged from automating data collection for mandatory reporting (e.g. HEDIS and meaningful use data) to enabling teams to use custom data extracts as they tried to improve their performance on quality metrics.

As a monitor, the EHR makes it possible to do ongoing reviews of process and outcomes. Perhaps because this is the least developed role, we found the most variation in effects. Uniformly organizations found that setting appropriate goals was difficult: "*If the A1c target is 6.9 and we are at 7.0, I'm not sure that is a fail*." And getting

good data could be difficult "*Our chlamydia and gonorrhea screening rates were three percent, and we said 'no way'…. It was a problem with the coding*." But once the goals were accepted, they could be used as an impetus for change: "*We roll out the measures…. [showed] there is a disconnect there… and so that allow us to change the workflow*."

When the monitoring through EHR data showed improvement, this willingness to change was reinforced. The use of dashboards, graphical indicators of provider and team performance, provided a sense of progress. The EHR as monitor could also trigger action: "*every month we pull a report [of diabetics] that includes nine different measures…and we then use this report to trigger telephone outreach…, vaccinations …and screenings*." Sometimes, however, the goals between providers would differ, and this would lead to conflict – for example when the endocrinologists were treating to an LDL goal of 100, but the primary care providers were only targeting a decrease to 120 ng/dl.

Overall:

We found four roles for the EHR – repository, messenger, orchestrator, and monitor; and that these four supported or inhibited collaboration behaviors and processes differently. Although they conceptualized it more in terms of the specific systems ("data warehouse", "clinical messaging," etc.), the leadership groups at every clinic we visited were actively working to improve the performance of their electronic health record in each of its roles. Table 3 summarizes the principle effects discusses above. In each role, we found evidence that the EHR affects all of the collaboration behaviors – both supporting and inhibiting collaboration. The inhibiting actions, however, appear to be more known, while the supporting factors are potential changes.

Table 3: The key issues for each collaboration behavior by role

	Repository	Messenger	Orchestrator	Monitor
Trust and respect: Enhancing positive relationships between providers	Increased awareness, but cut and paste and other quality issues decrease trust.	Asynchrony helps, but lack of richness in channel can result in misinterpretation	Particularly strong in establishing clear expectations	Key appears to be common goals and measurement
Communication: Providing the information and mutual understanding needed to care for patients	Facilitates information transfer "It's all there" (potentially), but "it's hard to find"	But doesn't guarantee communication Multiple channels can speed message delivery, but issues with "closing the loop"	Some successes, but clinical information is often not accessed/ignored by provider	When implemented can communicate gaps where practice improvement needed
Coordination: Having the right person do the right thing at the right time	A record of what actions and plans were, but each document frozen in time.	Issues due to variations in communications practice between providers	Bundles and "smart" worksheets particularly effective, but not implemented for enough conditions	Can facilitate team-based actions; if one member slips, another can fill in for them.
Collaboration: Facilitating collaborative decisions	Lack of interaction, one document per provider	No real time discussions, everything is lagged	Creates new boundaries, but doesn't encourage adaptation	Using dashboard and incentives to focus on common goals promotes dialogue

Limitations

Like any qualitative study, our results may not be generalizable. These sites were purposively selected to provide a cross section of successful EHR implementations in the United States. Readers will need to make their own judgments about the transferability of the results.

Discussion

Recent Research: Our results suggest that the EHR has evolved from its original role as a paper chart replacement (a repository) to also serving as messenger, an orchestrator, and a monitor. Issues with the performance of the EHR threaten its ability to support collaboration. Of particular concern were data quality and accessibility problems, which threaten the foundational behaviors of trust and respect. We also are concerned about the communication illusion, and issues with the multiplicity of communication channels. Other researchers have found similar issues at other locations using different EHRs and different qualitative perspectives.

Other studies with similar findings: One of our principal findings was that issues with the quality and accessibility of data threaten the EHR's utility for collaborative use. Weir, Hammond, Embi et al[24] used a lens based on Clark's theory of communication, joint action and common ground to examine the effects of computerized documentation on coordination and collaboration. Data were collected from focus groups at four different VA sites. Like our study, they found that the EHR could create a shared awareness and common database from which to act. They also found that "cut and paste" and failure to close the communications loop could create unintended consequences. Their work also discusses the value of narrative (richer contextual data) in building and maintaining shared mental models. This data were expanded to five sites and re-analyzed by Embi, Weir, Ehthiminiadis et al[25] using a grounded theory approach. Our work provides further evidence for the emergent themes from this expanded analysis. These included the inadequacy of the EHR as a sole communication channel, difficulties in finding relevant information, a need for better support for coordinated care, and disruptions in both trust and workflow due to problems with the EHR.

Our concerns about the communication illusions created by the EHR were foreshadowed by Lanham, Leykum, and McDaniel.[26] They used a complex systems approach to examine the effects of communication patterns on practice relationships. These relationships included trust and respect as well the appropriate use of communication channels. In a sample of six family medicine and specialty practices within a single system using the same EHR, they found that increased heterogeneity in communication patterns within each practice appeared to be related to increased practice fragmentation. We discuss a similar finding – the multiplicity of channels can result in communication patterns that inhibit collaboration.

Further research directions: EHR-related barriers to collaboration could be classified as predictable unintended consequences. The EHR was originally intended to replace the paper chart as a repository of data that would support medical reasoning and communication. But as healthcare is changing, so are the demands placed on the EHR and its related systems. It is unreasonable to expect an EHR to meet undefined needs. To meet these new demands we need a clear vision of what is needed. Both technological and organizational changes are needed. Processes need to be redesigned as well as technologies changed. We also need to be able to measure the effects of the changes once they are implemented.

Improved technology can help. Our study sites were developing better data warehousing and improved data exchange which would improve monitoring. We saw technology that could identify when a process step was missed and notify the proper person to get the gap filled to improve coordination. There is much innovation – and it is not confined to the sites we visited. The authors' informal discussion at both the AMIA student design competition and with EHR developers have shown possibilities for improving the EHR's ability to support collaboration. Better interface and data input technologies can provide time for better communications[27]. Curation can be improved by using plagiarism tools to identify inappropriate cut and paste[28]. Bundles that allow for more effective pathways and delegation can also free up time for providers[29].

More interestingly, the paradigm can be revised to fit a coordinated and collaborative process, rather than an individual practitioner. Models from other fields offer some possibilities. Ratings ("Amazon") for quality and usefulness might increase the quality of notes. Better identification of team members and their capabilities ("Facebook") could increase visibility and trust. With the advent of care managers, several "add on systems" have been developed to create a common care plan. What is needed is a reconfiguration of the basic structure of the

EHR. The ideal EHR for chronic disease management would allow for the integration of multiple care plans from different providers. This integrated care plan could be implemented as a wiki with the primary care provider or designee as the curator, or it could incorporate "column" care planning. It would be one care plan for one person.

Implementation and training are also key to improving collaboration through technology. As we saw in our visits, it is important to train users so they remember to keep talking, and not rely exclusively on leaner electronic media. Designing implementation processes that build buy-in from providers significantly increased adherence to the new protocols. And tying incentives to the data in the EHR – especially the problem lists – increased the value of the EHR data. There are other possibilities. Interventions to be tested include evaluating providers on the clarity and comprehensiveness of their notes, rather than on the number of billing factors; training providers on when to use what communication channels, and creating a "curators" who annotate and index of the documents for future use, accessibility would improve. Process innovations to gain acceptance are also understudied.

Finally, we need a tool to measure collaboration. The extant tools (cite) focus on supporting teams – collocated groups with a common goal. But collaboration between providers crosses organizational, geographic, professional and time boundaries. A more generalized tool would help us measure and manage the effects of the interventions discussed above.

Conclusion

In this study, we have identified four roles the EHR can and should play in the collaborative process. The organizations we visited demonstrated the potential for the EHR to support increased collaboration. These same systems also demonstrated barriers to collaboration presented by the implementations of current systems. Only by explicitly considering the different collaboration roles played by the EHR can designers and implementers develop the informatics tools needed for the twenty-first century. With these tools, the EHR can fill its role in the collaborative process.

Acknowledgements: In addition to my co-authors, the primary author would like to thank the POET/SAFER team (especially Adam Wright PhD, Dean Sittig PhD, and Hardeep Singh MD); Paul Gorman MD, Rachel Dresbeck PhD, and the Department of Medical Informatics and Clinical Epidemiology at OHSU; and the National Library of Medicine (training grant 5T15-LM007088-22) for their contributions and support.

References

(1) Wagner EH. The role of patient care teams in chronic disease management. BMJ 2000 Feb 26;320(7234):569-572.

(2) Bodenheimer T, Wagner EH, Grumbach K. Improving primary care for patients with chronic disease. JAMA 2002;288(14):1775-1779.

(3) Merriam-Webster. Merriam-Webster Online Dictionary. Available at: http://www.merriam-webster.com/dictionary/collaborate. Accessed August 4, 2013.

(4) Zillich AJ, Doucette WR, Carter BL, Kreiter CD. Development and initial validation of an instrument to measure physician-pharmacist collaboration from the physician perspective. Value in Health 2005 Jan-Feb;8(1):59-66.

(5) Stock R. Developing Team Based Care. 2013; Available at: http://www.pcpci.org/resources/webinars/developing-team-based-care-in-patient-centered-primary-care-home. Accessed March 11, 2014.

(6) Baggs JG. Nurse-physician collaboration in intensive care units. Crit Care Med 2007 Feb;35(2):641-642.

(7) Sierhouse M, Clancey W. Modeling and simulating practices, a work method for work systems design. Intelligent Systems, IEEE 2002;17(5):31.

(8) Clancey WJ, Sierhuis M, Damer B, Brodsky B. Cognitive modeling of social behaviors. In: Sun R, editor. Cognition and multi-agent interaction New York, New York: Cambridge University Press; 2005. p. 151-185.

(9) Salas E, Wilson KA, Murphy CE, King H, Salisbury M. Communicating, coordinating, and cooperating when lives depend on it: tips for teamwork. Joint Commission Journal on Quality & Patient Safety 2008 Jun;34(6):333-341.

(10) Fields D, Leshen E, Patel K. Analysis & commentary. Driving quality gains and cost savings through adoption of medical homes. Health Aff 2010 May;29(5):819-826.

(11) Rosenthal TC. The medical home: growing evidence to support a new approach to primary care. Journal of the American Board of Family Medicine: JABFM 2008 Sep-Oct;21(5):427-440.

(12) Gittell JH, Fairfield KM, Bierbaum B, Head W, Jackson R, Kelly M, et al. Impact of relational coordination on quality of care, postoperative pain and functioning, and length of stay: a nine-hospital study of surgical patients. Med Care 2000 Aug;38(8):807-819.
(13) National Center for Quality Assurance. Patient-centered medical home recognition. Available at: http://www.ncqa.org/Programs/Recognition/PatientCenteredMedicalHomePCMH.aspx. Accessed March 11, 2014.
(14) Oregon Health Authority. Patient Centered Primary Care Home. 2013; Available at: http://www.oregon.gov/oha/pcpch/Pages/index.aspx. Accessed March 11, 2014.
(15) Shortliffe EH. The evolution of electronic medical records. Academic Medicine 1999 Apr;74(4):414-419.
(16) Tang PC, McDonald CJ. Computer-based patient-record systems. Medical Informatics: Springer; 2001. p. 327-358.
(17) Sittig DF, Ash J editors. Clinical information systems: Ovecoming adverse consequences. Sudbury, Mass.: Jones and Bartlett; 2011.
(18) Beebe J. Rapid Assessment Process, An Introduction. Walnut Creek, CA: Altamira Press; 2001.
(19) McMullen CK, Ash JS, Sittig DF, Bunce A, Guappone K, Dykstra R, et al. Rapid assessment of clinical information systems in the healthcare setting: an efficient method for time-pressed evaluation. Methods Inf Med 2011;50(4):299-307.
(20) Strauss A, Corbin J. Basics of Qualitative Research, Grounded Theory Procedures and Techniques. Thousand Oaks, CA: Sage Publications; 1990.
(21) Office of the National Coordinator for Health Information Technology. The SAFER Guides. Available at: http://www.healthit.gov/policy-researchers-implementers/safer. Accessed March 10, 2013.
(22) QSR International. NVivo. 2011;9.
(23) Crabtree BF, Miller WL. Doing Qualitative Research. 2nd ed. Thousand Oaks, CA: Sage Publishing; 1999.
(24) Weir CR, Hammond KW, Embi PJ, Efthimiadis EN, Thielke SM, Hedeen AN. An exploration of the impact of computerized patient documentation on clinical collaboration. Int J Med Inf 2011 Aug;80(8):e62-71.
(25) Embi PJ, Weir C, Efthimiadis EN, Thielke SM, Hedeen AN, Hammond KW. Computerized provider documentation: findings and implications of a multisite study of clinicians and administrators. Journal of the American Medical Informatics Association 2013 Jul-Aug;20(4):718-726.
(26) Lanham HJ, Leykum LK, McDaniel RR,Jr. Same organization, same electronic health records (EHRs) system, different use: exploring the linkage between practice member communication patterns and EHR use patterns in an ambulatory care setting. Journal of the American Medical Informatics Association 2012 May-Jun;19(3):382-391.
(27) Hoyt R, Yoshihashi A. Lessons learned from implementation of voice recognition software in the military electronic health record. Perspect Health Inf Manag 2010;7(1e):1.
(28) Wrenn JO, Stein DM, Bakken S, Stetson PD. Quantifying clinical narrative redundancy in an electronic health record. J Am Med Inform Assoc 2010 Jan-Feb;17(1):49-53.
(29) Bloom F, Graf T, Anderer T, Stewart W. Redesign of a Diabetes System of Care Using an All-or- None Diabetes Bundle to Build Teamwork and Improve Intermediate Outcomes . Diabetes Spectrum 2010;20(3):165.

MedMinify: An Advice-giving System for Simplifying the Schedules of Daily Home Medication Regimens Used to Treat Chronic Conditions

Allen J. Flynn, PharmD[1], Predrag Klasnja, PhD[1], Charles P. Friedman, PhD[1,2]
[1]School of Information; [2]School of Public Health, University of Michigan, Ann Arbor, MI

Abstract

For those with high blood pressure, diabetes, or high cholesterol, adherence to a home medication regimen is important for health. Reductions in the number of daily medication-taking events or daily pill burden improve adherence. A novel advice-giving computer application was developed using the SMART platform to generate advice on how to potentially simplify home medication regimens. MedMinify generated advice for 41.3% of 1,500 home medication regimens for adults age 60 years and older with chronic medical conditions. If the advice given by MedMinify were implemented, 320 regimen changes would have reduced daily medication-taking events while an additional 295 changes would have decreased the daily pill burden. The application identified four serious drug-drug interactions and so advised against taking two pairs of medications simultaneously. MedMinify can give advice to change home medication regimens that could result in simpler home medication-taking schedules.

Introduction

To minify is to reduce the amount of something[1]. We studied how to use information technology to help minify the number of times consumers have to take medications each day and the quantity of pills they have to take. It is important to simplify home medication regimens in these ways to facilitate medication adherence. While simpler home medication-taking schedules may diminish medication mishaps[2], this research is justified on the basis that better adherence resulting from simplified medication regimens improves the health and well being of consumers.

For individuals diagnosed with high blood pressure, high cholesterol, or diabetes, adherence to home medication regimens is known to improve health[3]. Conversely, lack of adherence to home medication regimens for chronic conditions causes considerable harm[4]. The overall cost of non-adherence to medication regimens has been estimated to be $100 billion per year in the United States[5]. Several categories of strategies to increase medication adherence have proven modestly effective[5]. These include strategies to improve communication of what to take and why, strategies to support behavior change such as self-monitoring of adherence, strategies to decrease out-of-pocket costs of medications, and strategies to cue consumers to take their medications such as reminder systems and calendars[5]. The most effective approaches combine strategies from multiple categories[5]. However, in terms of effect size, the evidence indicates that greater adherence improvements result from decreasing the complexity of home medication regimens than from combining educational, behavioral, and cueing interventions[5]. One strategy to decrease the complexity of medication regimens is to standardize the times of daily medication-taking events and limit their number to a maximum of four[6]. Another strategy is for health care providers to conduct medication regimen reviews intended to simplify medication regimens[7]. To facilitate these latter two strategies, we have developed an advice giving system to assist in the process of simplifying home medication regimens.

This paper describes the architecture and early stage evaluation of MedMinify, a new application that offers advice on how a home medication schedule could be made simpler by reducing the number of daily medication-taking events and the daily pill burden while accounting for drug-drug interactions. Because problems with medication adherence are widespread, MedMinify was developed to interoperate with a broad range of electronic health record systems. Our intent is to avoid the need to recreate the functionality of MedMinify in every different electronic health record (EHR) system. For the purpose of simplifying regimens, MedMinify advises users as to the availability of sustained-release or fixed-dose combination drug products and indicates whether certain drug interactions pertain. The literature supports the potential effectiveness of regimen simplification to improve medication adherence[7,8].

Background and Significance

In a systematic review of the association between the number of daily medication-taking events and adherence, Claxton, Cramer, and Pierce found mean adherence to a medication regimen declined as daily medication-taking events increased. Mean adherence for one, two, three, and four daily medication-taking events was 79% (SD 14), 69% (SD 15), 65% (SD 16), and 51% (SD 20) respectively[8]. The mean adherence differences between one daily medication-taking event and either three or four daily medication-taking events were statistically significant as was the mean difference between two daily medication-taking events and four daily medication-taking events[8]. In light

of more recent concurring evidence, these findings indicate that decreasing daily medication-taking events results in improved adherence[9,10].

Controlled studies of a standardized home medication schedule where patients take all of their medications either at breakfast, lunch, dinner or bedtime, have shown that in most cases consumers do not need to take medications for many common chronic conditions more than four times a day[11-13]. These studies suggest there is an opportunity to minify the number of daily medication-taking events[12].

Adherence to home medication regimens is also influenced by the number of pills taken per day or daily pill burden. To minify the number of pills taken daily can also be seen as a necessary adherence intervention[10,14]. In a recent meta-analysis, higher pill burden was associated with lower medication adherence to medication regimens that treat human immunodeficiency virus infection[10]. In another study agreement with the statement, "I am already taking too many medications", was a predictor of very low adherence amongst women being treated for osteoporosis[15].

One difficulty in minifying daily medication-taking events and daily pill burden to simplify home medication schedules is that consumers may actually increase the number of scheduled daily medication-taking events on their own for a variety of reasons[16]. A partial explanation for this behavior could be concern over drug-drug interactions[16]. Consumers may assume that taking oral medications at different times of day mitigates adverse events from drug interactions. Actually, relatively few drug interactions result from simultaneous oral ingestion[17]. More often drug interactions result from systemic effects at the sites in the body where drug metabolism, pharmacologic action, or drug excretion take place[17]. To counteract their largely false beliefs, what consumers may need is explicit information regarding which drugs they should and should not ingest simultaneously.

Previous research has examined the use of information technology to assist in reducing medication regimen complexity and polypharmacy[18]. Others have reported on the automated generation of pill cards and printed home medication schedules[13]. We have used information technology differently to give advice on how to minify daily medication-taking events and daily pill burden while ensuring that the advice given is in concordance with what is known about the relatively few serious drug-drug interactions that result from the simultaneous ingestion of drugs.

This research is significant for several reasons. While the majority of patients take multiple medications for multiple conditions, many interventional trials have focused on adherence to single medication therapies or to treatments for one condition. We intentionally sought to analyze prescriptions for all scheduled tablets and capsules consumers must take to adhere to their home medication regimens. In addition, whereas pharmacist-led medication regimen review helps to simplify medication regimens, a lack of pharmacist time to conduct home medication regimen reviews limits this intervention[7]. An information resource that gives advice on how to minify daily medication-taking events and daily pill burden may decrease the amount of time required to review home medication regimens for simplification opportunities. Such advice may also apply during medication reconciliation tasks. Finally, to facilitate future integration with EHR systems we set a goal to build MedMinify so that it would interoperate with a variety of EHR systems from the outset.

Motivated by these research findings and the clinical significance of the problem of medication adherence, this paper reports whether an advice giving system capable of offering potentially useful advice to simplify medication regimens could be built using the Substitutable Medical Apps, Reusable Technology (SMART) platform, how frequently potentially useful advice to assist in simplifying medication regimens could be generated, and in what ways and by how much the advice generated, if heeded, could change actual home medication regimens.

Research Questions

The following three research questions were investigated.

1. Is the Substitutable Medical Apps, Reusable Technology (SMART) platform robust and extensible enough to support the architectural and functional requirements of a system capable of giving advice on how to potentially minify daily medication-taking events and daily pill burden?

2. What types of advice and how much of each type can be generated by analyzing home medication regimens with an information resource capable of identifying candidate regimens for minification of daily medication-taking events and daily pill burden while accounting for drug-drug interactions from simultaneous ingestion?

3. To what extent would the requirements of home medication regimens change if advice given by a computer application on how to minify daily medication-taking events and daily pill burden were implemented?

Methods

SMART Application Development

MedMinify was developed to meet the following requirements. The application had to be able to access and analyze prescription data organized as medication regimens for individuals. To provide drug product advice, MedMinify had to be able to identify the active ingredients in all current prescriptions for an individual and check for other substitutable drug products containing the same active ingredients. To ascertain whether drug-drug interactions pertain, MedMinify required a drug-drug interaction knowledge table and the routines to check whether the active ingredients in the current prescriptions for an individual could be found in the drug-drug interaction knowledge table. Finally, to facilitate regimen review by the investigators, MedMinify had to render key facts about each individual's regimen within a browser window along with any advice that the application generated.

Software design and application development of MedMinify involved the SMART platform (smartplatforms.org), the Django Web framework (Version 1.5, djangoproject.com), and the RxNorm application programming interfaces (rxnav.nlm.nih.gov/RxNormAPIs.html#). The organization of the MedMinify application is depicted in Figure 1. The SMART Reference EMR was used for development, testing, and evaluation of MedMinify. SMART afforded data models for individuals and their corresponding prescription records with an application programming interface (API) for the SMART Reference EMR. The Django Web framework afforded a Python-based Web development environment integrated with SQLite database technology. The Django Web framework also included a host of ready to use tools that enabled the development, testing, and rapid iterative refinement of the application. The RxNorm APIs afforded automated lookup of drug products marketed in the United States and their ingredients. MedMinify used Javascript executed in a browser along with Hyper Text Markup Language and Cascading Style Sheets to call the SMART API, render and format application output, respectively (Figure 1).

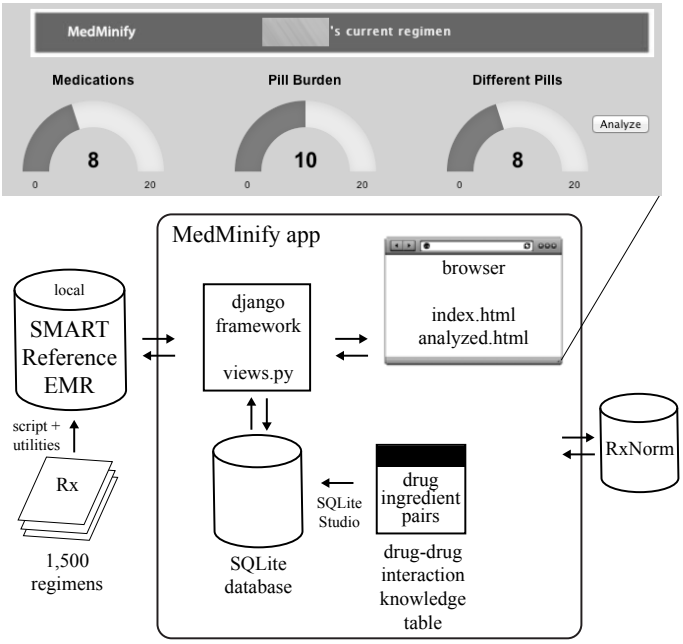


Figure 1. The MedMinify application architecture with a partial screen shot of its user interface. At the top of the image is a cutout of part of a user's view in a browser with counts of the Medications, Pill Burden, and Different Pills for one regimen. A local instance of the SMART Reference EMR platform was loaded with 1,500 sampled multi-drug enteral medication regimens (Rx). MedMinify used the Django Web framework and a SQLite database. Drug-drug interaction knowledge was loaded into the MedMinify SQLite database.

Several field value conversions were undertaken to convert data from the University of Michigan's EMR to the formats used by the SMART Reference EMR. These data conversions were made using rules in a spreadsheet. For example, the quantity of each dosage form to be administered was not a discrete field in the University of Michigan's EMR prescription data but the SMART EMR accepted a discrete value for the quantity of pills to be

administered. The quantity of pills to take at each medication-taking event was extracted from the instructions field in the original University EMR prescription data. A prescription instruction "Take 2 tablets daily" resulted in an extracted numeric quantity value of "2" for use by SMART. Similar data conversions were conducted for the frequency data from the University's EMR to transform the frequencies provided into the format used by SMART.

To support the drug interaction identification function of MedMinify, knowledge from drug interaction monographs was used (First DataBank, Inc. (FDB), San Francisco, CA). FDB provided the investigators with select fields from a complete set of their highest severity (FDB severity level 1) drug interaction monographs. The fields provided were the drug interaction monograph title, drug interaction mechanism, and clinical effects fields. Keyword analysis of the FDB knowledge source was conducted to select only those drug interactions related to simultaneous ingestion of medications. The 34 keywords included "gut", "intestine", "bioavailability", "absorption", "gastric", and "pH". A list of 59 unique, severity level 1 drug-drug interactions that are mechanistically associated with simultaneous ingestion was developed. After determining the drug pairs associated with each of the selected 59 drug-drug interactions and removing duplicates, a list of 559 interacting, active drug ingredient pairs for interactions due to simultaneous ingestion was created. For each active drug ingredient involved, the RxNorm ingredient (IN) code was identified. A free software application was used to load this drug interaction ingredient pair knowledge into the MedMinify application database (SQLiteStudio 2.1.5 by Pawel Salawa). An example row from this drug-drug interaction knowledge table includes the antacid omeprazole (RxNorm IN = 7646) and the antiviral drug atazanavir (RxNorm IN = 343047). It is known that omeprazole lowers the acidity of the gastric juices decreasing the solubility of atazanavir, disrupting atazanavir absorption, and thereby diminishing the potential efficacy of atazanavir.

After application development and testing using the publicly available SMART Reference EMR online sandbox, a local copy of the SMART Reference EMR was installed and made operable on an Apple Mac Mini computer running Mac OS version 10.8.5. A fictional name generator (http://homepage.net/name_generator/) was used to generate 1,500 fake names for loading the 1,500 randomly sampled regimens into the local SMART instance for use with MedMinify. After the 1,500 fictitious patients and corresponding regimens were loaded into SMART, data validation occurred by examining the SMART database. MedMinify application testing and use was conducted using an Apple MacBook Pro laptop running Mac OS version 10.8.5 connected via a local area network to the local SMART instance. The Safari browser was used (version 6.1.1).

Data source and sample

This study evaluated the function of a computer application we developed called MedMinify. To conduct the evaluation we used deidentified prescription data from the electronic medical record (EMR) system at the University of Michigan Hospitals and Health Systems (EpicCare Ambulatory Electronic Medical Record, Epic, Verona, WI). The Institutional Review Board of the University of Michigan reviewed and approved this study. Up to date medication regimens, provided as records of prescriptions for 41,903 unique individuals, were selected via query using the following criteria, a-c:

a. Each individual whose medication regimen was selected was alive and of a chronological age greater than or equal to 60 years on the date of the query, January 15, 2014.

b. Each individual whose medication regimen was selected had at least one medication list update documented in the University of Michigan's EMR during calendar year 2013 or the first two weeks of 2014 to ensure regimen currency.

c. Each individual's University of Michigan EMR profile had an International Classification of Diseases (ICD version 9) coded diagnosis for hypertension (401.x, 405.x, 416.x, 459.3x) and/or diabetes (249.x, 250.x, 253.5) and/or hypercholesterolemia (272.x).

These three criteria were specifically chosen to provide real medication regimens for adults using multiple medications to treat common diseases. This query of the University of Michigan's EMR resulted in data for 380,932 prescriptions. Besides patient identifiers such as age and gender, these data included prescription fields for ordering date, prescription start date, prescription end date, generic drug name, drug product strength, frequency, medication dosage form, route of administration, quantity to dispense, number of refills, medication therapeutic subclass, prescriber instructions, and RxNorm Semantic Clinical Drug code. Not all of these data were usable. A data exclusion diagram for the 380,932 prescriptions is given in Figure 2.

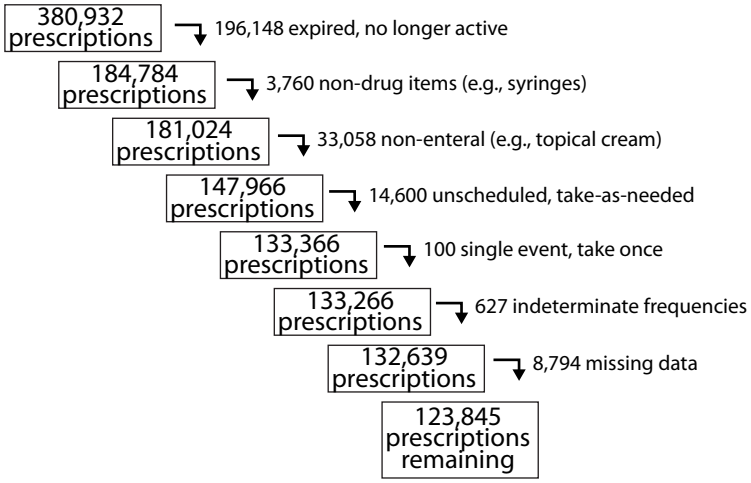


Figure 2. Prescription data was excluded for expired prescriptions, prescriptions for non-drug items, non-enteral prescriptions, take-as-needed prescriptions, single event prescriptions, and prescriptions with indeterminate frequencies. Prescriptions missing frequencies, product strengths, or RxNorm Semantic Drug codes were also removed.

After exclusions were applied and removal of prescriptions missing data, 123,845 active prescriptions remained for 35,042 unique individuals for scheduled and recurring home medications with known product strengths administered via an enteral route (e.g., orally or sublingually). Next, 8,105 complete medication regimens consisting of only one medication were also excluded leaving 115,740 prescriptions comprising the home medication regimens of 26,937 individuals with more than one scheduled and recurring medication of known product strength administered via an enteral route. These data included enteral medications prescribed for use on a weekly (e.g., "three times a week") or monthly (e.g., "once a month") schedule. For this study, these 26,937 medication regimens represent the population of qualified multi-drug enteral medication regimens used to treat chronic hypertension, hypercholesterolemia and diabetes. The 26,937 regimens include regimens for 14,403 females (53.5%) and 12,534 males (46.5%). The 26,937 individuals included had an average age of 71 years (SD 8.3, Range 60 to 103).

Because each regimen required approximately 10 seconds of time to analyze with MedMinify and 30 seconds more to assess if advice was given, instead of analyzing all 26,937 regimens a random sample of 1,500 regimens consisting of 6,241 prescriptions was taken from the population of 26,937 qualified regimens using SPSS (version 21). Random sampling was preferred over a selective sample of the most complex regimens because we believed random sampling to be the most conservative test of MedMinify's capabilities. The sample size of 1,500 regimens was chosen to give a desired margin of error of 0.025 for 95% confidence intervals for Chi-squared tests with one degree of freedom[19]. Sampling was done without replacement resulting in a hypergeometric distribution. However because the population of 26,937 is much larger than the sample size of 1,500 it was assumed that the binomial distribution provided a reasonably good approximation for calculating proportion confidence intervals. Therefore the Agresti-Coull adjusted Wald interval for large sample sizes was used to calculate the confidence intervals for sample proportions reported in this paper[20].

Generation of advice to minify daily medication-taking events or daily pill burden using a computer application

All 1,500 randomly sampled regimens were analyzed using MedMinify. For each regimen, MedMinify calculated the number of prescriptions in the regimen, the daily pill burden, the number of unique pills in the regimen, and the maximum number of daily medication-taking events required by the regimen. Next, for every immediate-release pill taken more than once a day, MedMinify checked RxNorm to see whether a sustained-release product was available as a potential substitute. If so, MedMinify recommended that the sustained-release product or products be considered as substitutes. Subsequently, for every pair of active ingredients prescribed as two single ingredient products, MedMinify checked RxNorm to see whether a fixed-dose combination drug product was available that combined two active ingredients in one pill. If so, MedMinify recommended that the fixed-dose combination product or products be considered as substitutes. Next, MedMinify screened every pair-wise combination of active drug ingredients in each regimen against the table of known FDB severity level 1 interacting drug-drug pairs where the drug-drug interaction is associated with simultaneous ingestion. When a potential drug-drug interaction was

identified, MedMinify gave advice not to take the two interacting drugs at the same time of day. Finally, MedMinify counted all instances when a single active ingredient in a regimen appeared in the drug-drug interaction knowledge table (i.e., one drug of an interacting pair) even where no drug interaction was found. Because this was a lab function study, there were no consumer end users and none of the advice generated was ever implemented.

Quantification of how much home medication regimen requirements would change if the computer-generated advice were implemented

All advisories given by MedMinify to change sampled medication regimens were documented and counted. Assuming MedMinify's advice was accepted and implemented, the potential change or changes to each regimen where advice was generated was assessed in two ways. First, if a sustained-release product or products was recommended, the pharmacist investigator (AJF) checked to see whether the maximum number of medication-taking events per day could be reduced, and, if so, by how many daily events. (The opportunity to minify the maximum number of medication-taking events per day depends on whether the recommendation to substitute a sustained-release product applies to the prescription with the most daily medication-taking events.) Next, if either a sustained-release or a fixed-dose combination product or both were recommended by MedMinify, the pharmacist investigator checked to see whether the daily pill burden could be reduced, and if so by how much. When overlaps implementing this advice arose, for example when the same active ingredient could be taken via a sustained-release drug product or as a component of a fixed-dose combination drug product, preference was given to minifying the number of daily medication-taking events over minifying the daily pill burden.

Results

1. Is the Substitutable Medical Apps, Reusable Technology (SMART) platform robust and extensible enough to support the architectural and functional requirements of a system capable of giving advice on how to potentially minify daily medication-taking events and daily pill burden?

The MedMinify application is an example of a SMART Connect application (Figure 1). SMART Connect provides the SMART JavaScript client for the purpose of communicating from a SMART container via the SMART API when connecting to the SMART Reference EMR with a browser. To meet MedMinify's requirement for access to prescription data organized as regimens, SMART permitted MedMinify access to prescription data stored in its database with one minor limitation. The SMART Reference EMR data model could not fully represent prescriptions with specific days and times in coded fields (e.g., "Take 1 tablet at 2:00 p.m. on Tuesdays and Thursdays."). A few prescriptions had to be represented in less detail with respect to dose timing upon conversion to SMART. To meet the requirement for RxNorm integration, procedural code written in a high-level programming language, Python, was used to manage calls and responses via the RxNorm prescribable APIs. The Django Web framework played a central role in the architecture of MedMinify. To meet the requirement for integration of drug-drug interaction knowledge, the Django Web framework provided an integrated SQLite database. Finally, to meet the requirement for rendering output to the screen within the SMART container, the Django Web framework provided a web server to serve various files including index.html and analyzed.html (Figure 1). The combination of SMART and the Django Web framework afforded a stable, extensible platform upon which MedMinify was successfully built.

2. What types of advice and how much of each type can be generated by analyzing home medication regimens with an information resource capable of identifying candidate regimens for minification of daily medication-taking events and daily pill burden while accounting for drug-drug interactions from simultaneous ingestion?

Three sets of results address the second research question, quantification of the overall amount of advice generated, categorization of the generated advice by type, and an analysis of MedMinify's drug-drug interaction checking function. Descriptive statistics and results quantifying the number of recommendations given as advice for minifying the 1,500 sampled regimens are given in Table 1 where the sampled regimens are categorized according to the maximum number of daily medication-taking events.

Table 1. Sampled qualified multi-drug enteral medication regimens (n = 1,500) data for individuals of age greater than or equal to 60 years with diagnoses for hypertension or diabetes or hypercholesterolemia (alone or in combination) according to the maximum number of daily medication-taking events per regimen.

Number of Daily Medication-Taking Events Category	Number of Sampled Regimens	Number of Sampled Regimens with Recommendations from MedMinify	Average Number of Prescriptions per Sampled Regimen	Average Daily Pill Burden per Sampled Regimen
Max. 8 events daily	1	0	7.0	26.0
Max. 6 events daily	3	2	5.3	15.6
Max. 5 events daily	4	2	6.3	15.8
Max. 4 events daily	44	22	5.4	11.3
Max. 3 events daily	178	126	5.7	10.4
Max. 2 events daily	643	388	4.6	6.7
Max. 1 event daily	626	79	3.2	3.4
Max. 0* events daily	1	0	2.0	0.0
TOTALS	1,500	619		

*One regimen included only prescriptions for medications to be taken weekly and no medications to be taken daily.

MedMinify identified 619 (41.3%) of the 1,500 randomly sampled regimens, 95% CI [38.7, 43.7] as candidates for minifying daily medication-taking events or daily pill burden.

In Table 2 below, the 619 regimens where at least one recommendation to change the regimen was given are further categorized by the type of recommendation. Population estimates for the proportion of regimens subject to each type of advice are given as 95% confidence intervals in Table 2.

Table 2. Counts and percentages of sampled regimens (n = 1,500) are provided by recommendation type with confidence intervals for population estimates of the percentages of regimens for which MedMinify would generate recommendations for the study population.

Recommendations Generated per Sampled Regimen	Number of Sampled Regimens	% of Sampled Regimens (95% CI)
Any Recommendation	619	41.3% (38.8, 43.8)
Sustained Release Product Recommendations Only	407	27.1% (24.9, 29.4)
Fixed-Dose Combination Product Recommendations Only	114	7.6% (6.3, 8.9)
Sustained Release and Fixed Dose Combination Products	94	6.3% (5.0, 7.5)
Sustained Release Product and Drug-Drug Interaction	3	0.2% (0.0, 0.4)
Drug-Drug Interaction Only	1	0.1% (0.0, 0.3)

For the 1,500 sampled regimens, MedMinify identified at least one drug on the FDB drug-drug interaction list in 1,209 (81%) of regimens, 95% CI [78.6, 82.6]. MedMinify recommended against taking two drugs at the same time in four cases due to drug-drug interactions related to simultaneous ingestion of drug products. Three of the four drug interactions involved the drugs cyclosporine and simvastatin. (There is a risk that competition for influx transporter proteins in the small intestine and liver could change the serum levels and efficacy of either drug[21]. It is not clear whether taking cyclosporine and simvastatin at different times of day mitigates the risk of harm from this interaction.) The fourth case of a drug-drug interaction detected by MedMinify involved the two drugs rosuvastatin and gemfibrozil. (There is a risk of muscle damage when using these drugs simultaneously. The exact mechanism of this drug interaction is unknown but there is a suggestion in the literature that simultaneous ingestion may play a role in the interaction[22]. It is not clear whether taking rosuvastatin and gemfibrozil at different times of day mitigates the risk of harm from this drug interaction.)

3. To what extent would the requirements of home medication regimens change if advice given by a computer application on how to minify daily medication-taking events and daily pill burden were implemented?

If MedMinify's advice were accepted and implemented for all of the regimens for which recommendations were generated, 320 out of the 1,500 regimens (21.3%, 95% CI [19.3, 23.4]) would have at least one fewer daily

medication-taking event. Three regimens would actually gain an additional daily medication-taking event to mitigate an identified drug-drug interaction by taking the interacting drugs at two different times of day. The daily pill burden for an additional 295 of the 1,500 regimens (19.7%, 95% CI [17.7, 21.8]) would be minified by an average of 1.4 pills per regimen. Table 3 includes more detail about these regimen changes.

Table 3. Changes for sampled qualified multi-drug enteral medication regimens (n = 1,500) according to the maximum number of daily medication-taking events per regimen if all of MedMinify's advice were implemented.

Number of Daily Medication-Taking Events Category	Number of Sampled Regimens	Number of Sampled Regimens with Fewer Daily Medication-Taking Events	Number of Sampled Regimens with Decreased Daily Pill Burden Only	New Average Daily Pill Burden per Sampled Regimen if Advice Implemented (Difference)
Max. 8 events daily	1	0	0	26.0 (0)
Max. 6 events daily	3	1	1	13.3 (-2.3)
Max. 5 events daily	4	1	1	14.8 (-1.0)
Max. 4 events daily	44	10	12	10.2 (-1.1)
Max. 3 events daily	178	80	45	9.3 (-1.1)
Max. 2 events daily	643	228	157	5.8 (-0.9)
Max. 1 event daily	626	N/A[†]	79	3.3 (-0.1)
Max. 0[*] events daily	1	N/A	N/A	0.0 (0)
TOTALS	1,500	320	295	

[*]One regimen included only prescriptions for medications to be taken weekly

[†]One is a minimum daily medication-taking event. MedMinify does not provide advice to discontinue medication therapies.

Discussion

Taking medications every day at home to treat chronic diseases is difficult for most people. Whether and how well one adheres to a home medication schedule is a result of a complex set of influences including economic factors, social support, individual dedication to the task, and beliefs about the utility of prescribed medication treatments[5,23]. For individuals with means who have supportive relationships and believe that taking medications is beneficial, adherence to complex medication regimens is still difficult[8]. More attention on simplifying regimens is needed to make every home medication regimen as easy to execute as possible[6].

Pharmacists have long been champions of simplifying medication regimens and reducing polypharmacy[24]. However, many consumers are unaware that their pharmacist can assist them to simplify their home medication regimens. Relatively few consumers are assured to have the simplest possible medication regimens. MedMinify could improve the efficiency of time-consuming home medication regimen reviews provided unsystematically to consumers today.

In this initial evaluation of the function of the MedMinify advice-giving system in the lab, advice to minify daily medication-taking event and daily pill burden was generated for 41.3%, 95% CI [38.7, 43.7] of actual home medication regimens for adults with one or more of three common chronic diseases. This result is bolstered by a previous analysis we did where we found that similar drug product substitution advice could be generated for 38.4% of 2,944 home medication regimens of retirees surveyed in 2007 and by results from Elliott that potential changes to reduce complexity were identified in 45.7% of reviewed medication regimens at hospital discharge[7,25].

Based on a careful analysis of the 1,500 sampled regimens we believe the 41.3% figure may overestimate the actual percentage of candidate regimens subject to applicable simplification advice. However, if for reasons of consumer preference, clinical indication, drug product cost, or perceived low marginal utility from changing prescriptions only 1 out of 5 home medication regimens for which advice was given by MedMinify would actually be simplified in practice, the fraction of regimens changed would be approximately 8% in this population. Routine screening with an information resource to try and minify daily medication-taking events or daily pill burden may be justified if 8 out of every 100 regimens could be made simpler. This is a particularly important finding when one considers that 15.3%, 95% CI [13.5, 17.2] of home medication regimens in this population require individuals to take medications more than two times a day (Table 1). We found that 151 of these 230 complex regimens (65.6%) were candidates for simplification, and that the greater fraction of the advice given to simplify these regimens would have resulted in diminishing the number of daily medication-taking events (Table 3). Because the regimens studied were real and current regimens for actual individuals, it is reasonable to generalize these findings to similar consumer populations.

Specific recommendations to substitute sustained-release products for immediate-release ones were the most common recommendations made by MedMinify (Table 2). An example would be to use the antihypertensive beta-blocker carvedilol as a sustained-release product once a day instead of twice a day as an immediate-release product. MedMinify generated fixed-dose combination recommendations for a total of 208 out of the 1,500 regimens (Table 2). An example of a fixed-dose combination recommendation would be to substitute a pill that contains both the antihypertensive drugs hydrochlorothiazide and lisinopril instead of taking these two drugs as separate pills.

Consumers may separate their medication-taking tasks throughout the day due to confusion over prescription instructions or concerns about drug-drug interactions[16]. It was expected that very few drug-drug interactions would be identified in real regimens like those studied because of prior drug interaction screening. In this study, while most regimens included one drug in the drug interaction knowledge table, rarely were two interacting drugs found in the same regimen. While commonly used medications such as the "cholesterol-lowering statins" are involved in severity level 1 drug-drug interactions, to cause an interaction these drugs had to have been paired with other drugs that are rarely used in this population. By checking for the rare drug-drug interactions that can be managed by taking medications at different times of day, MedMinify can be used to indicate how most medications prescribed to treat high blood pressure, diabetes, and high cholesterol can safely be taken together at the same time of day.

This study has several limitations. With respect to the methods used, the clinical appropriateness of the advice offered by MedMinify was not assessed. Also prescriptions for non-enteral drug products and take-as-needed prescriptions were excluded making the regimens studied less complex than they actually are. With respect to limits on MedMinify's functionality, the application cannot currently reconcile medication-taking events with individual preferences for meal times, work hours, or sleep schedules. MedMinify does not yet account for the differing cost of various medication products neither for prescription insurance coverage. MedMinify does not recognize the difference between temporary or trial prescriptions and other prescriptions. Also, by limiting the scope of drug-drug interaction assessment to only severity level 1 interactions as defined by FDB, we surely overlooked some less severe but important drug-drug interactions.

Future work will address some of the limitations. We also intend to study having MedMinify give other types of advice including advice about duplicative prescriptions, potentially inappropriate medications[26], and polypharmacy. We also look forward to further developing and evaluating a consumer-oriented user interface for MedMinify that will gather details about an individual's daily schedule and provide advice that is informed by it.

To achieve an ideal level of adherence to home medication regimens requires much more than decreasing the number of medication-taking events or the daily pill burden. However, adherence is improved when home medication-taking schedules become simpler. Minifying daily medication-taking events and daily pill burden is necessary, but not sufficient, to improve adherence to medication regimens used to treat chronic conditions.

Conclusion

A new advice-giving system developed for the SMART platform, and the advice it gave about drug products and drug-drug interactions, was studied. Advice to minify the number of medication-taking events per day and the number of pills taken daily was generated for 41.3% of home medication regimens for adults with chronic conditions. The percentage of regimens for which advice was given increased as the complexity of the home medication regimens increased. Even if only a fraction of the advice given resulted in regimen changes, a considerable number of home medication regimens would include fewer daily medication-taking events involving fewer pills. This study provides preliminary evidence that an information resource could be used routinely to help simplify home medication-taking schedules for adults with chronic medical conditions.

References

1. "minify, *v.*", Oxford English Dictionary, 3rd Edition, OED Online, The Oxford University Press; 2014.
2. Willson MN, Greer CL, Weeks DL. Medication regimen complexity and hospital readmission for an adverse drug event. Ann Pharmacother. 2014 Jan;48(1):26-32.
3. Deaton A. The great escape: health, wealth, and the origins of inequality. Princeton University Press; 2013.
4. Mennini FS, Marcellusi A, von der Schulenburg JM, Gray A, Levy P, Sciattella P, Soro M, Staffiero G, Zeidler J, Maggioni A, Schmieder RE. Cost of poor adherence to anti-hypertensive therapy in five European countries. Eur J Health Econ 2014;Jan 5 PMID: 24390212.
5. Bosworth HB, Granger BB, Mendys P, Brindis R, Burkholder R, Czajkowski SM, Daniel JG, Ekman I, Ho M, Johnson M, Kimmel SE, Liu LZ, Musaus J, Shrank WH, Whalley Buono E, Weiss K, Granger CB. Medication adherence: a call for action. Am Heart J. 2011. Sep;162(3):412-24.

6. Isham G, chair. Roundtable on health literacy, Institute of Medicine. Standardizing medication labels: confusing patients less, workshop summary. 2008.
7. Elliott RA, Reducing medication regimen complexity for older patients prior to discharge from hospital: feasibility and barriers, J Clin Pharm Ther. 2012. (37):637-642.
8. Claxton AJ, Cramer J, Pierce A. A Systematic Review of the Associations between dose regimens and medication compliance. Clinical Therapeutics. 2001. 23(8):1296-1310.
9. Cohen C, Elion RA, Frank I, Kloser P, Sherer R, Squires KE, Corklin S, Tebas P. Once-daily antiretroviral therapies for HIV infection: consensus of an advisory committee of the international association of physicians in AIDS care. JIAPAC. 2002 (1):141-145.
10. Nachega JB, Parienti JJ, Uthman OA, Gross R, Dowdy DW, Sax PE, Gallant JE, Mugavero MJ, Mills EJ, Giordano TP. Lower pill burden and once-daily antiretroviral treatment regimens for HIV infection: a meta-analysis of randomized controlled trials. Clin Infect Dis. 2014;Mar 5 PMID 24457345.
11. Kripalani S, Jacobson TA. Illustrated medication schedules improve medication adherence in at-risk patients with coronary heart disease [abstract]. J Gen Intern Med. 2010;25(S3):S301.
12. Gazmararian J, Jacobson KL, Pan Y, Schmotzer B, Kripalani S. Effect of a pharmacy-based health literacy intervention and patient characteristics on medication refill adherence in an urban health system. Ann Pharmacother. 2010; 44(1):80-87.
13. Schnipper JL, Roumie CL, Cawthon C, Businger A, Dalal AK, Mugalla I, Eden S, Jacobson TA, Rask KJ, Vaccarino V, Gandhi TK, Bates DW, Johnson DC, Labonville S, Gregory D, Kripalani S. PILL-CVD Study Group. Rationale and design of the pharmacist intervention for low literacy in cardiovascular disease (PILL-CVD) study. Circ Cardiovasc Qual Outcomes. 2010;3(2):212-219.
14. Stange D, Kriston L, von-Wolff A, Baehr M, Dartsch DC. Reducing cardiovascular medication complexity in a German university hospital: effects of a structured pharmaceutical management intervention on adherence. J Manag Care Pharm. 2013 Jun;19(5):396-407.
15. Solomon DH, Brookhart MA, Tsao P, Dundaresan D, Andrade SE, Mazor K, Yood R. Predictors of very low adherence with medications for osteoporosis. Osteoporos Int. 2011 (22):1737-1743.
16. Wolf MS, Curtis LM, Waite K, Bailey SC, Hedlund LA, Davis TC, Shrank WH, Parker RM, Wood AJ. Helping patients simplify and safely use complex prescription regimens. Arch Intern Med. 2011 Feb 28;171(4):300-5.
17. Hansten PD, Horn JR. The top 100 drug interactions. H&H Publishing, 2013.
18. Farrish S, Grando A. Ontological approach to reduce complexity in polypharmacy. AMIA Annu Symp Proc. 2013 Nov 16;2013:398-407.
19. Krejcie RV, Morgan DW. Determining sample size for research activities. Ed Psych Meas. 1970;30:607-10.
20. Brown LD, Cai TT, DasGupta A. Interval Estimation for a proportion. Statistical Science 2001;16:101-133.
21. Kalliokoski A, Niemi M. Impact of OATP transporters on pharmacokinetics. Br J Pharmacol. 2009;Oct;158(3):693-705.
22. Bergman E1, Matsson EM, Hedeland M, Bondesson U, Knutson L, Lennernäs H. Effect of a single gemfibrozil dose on the pharmacokinetics of rosuvastatin in bile and plasma in healthy volunteers. J Clin Pharmacol. 2010 Sep;50(9):1039-49.
23. Norell, SE, Improving medication compliance: a randomised clinical trial. Br Med J. 1979;2(6197):1031-3.
24. Lenaghan E1, Holland R, Brooks A. Home-based medication review in a high risk elderly population in primary care-the POLYMED randomised controlled trial. Age Ageing. 2007 May;36(3):292-7.
25. Flynn AJ, Klasnja P, Friedman CP. Taking it easy – a needs analysis for computer-generated advice to simplify home medication regimens, AMIA Annu Symp Proc. 2013 Nov 16.
26. Shade MY, Berger AM, Chaperon C. Potentially Inappropriate Medications in Community-Dwelling Older Adults. Res Gerontol Nurs. 2014 Feb 19:1-15.

Acknowledgments

This research would not have been possible without support from the following individuals and their organizations: Dr. Joan Kapusnik-Uner of First Databank, Inc. who provided the drug interaction knowledge, Lalitha Natarajan and James Law of the University of Michigan who queried the medication regimen data, Dr. John Kilbourne, Head, Medical Subject Headings, National Library of Medicine and the RxNorm team, and Josh Mandel, Dr. Kenneth Mandl, Dr. Pascal Pfiffner, and Nikolai Schwertner of the Substitutable Medical Apps, Reusable Technologies (SMART) platform project, a Strategic Health IT Advanced Research Program (SHARP) project funded by the Office of the National Coordinator for Health IT of the U.S. Department of Health and Human Services.

Could Patient Self-reported Health Data Complement EHR for Phenotyping?

Daniel Fort, MPH[1], Adam B. Wilcox, PhD[2], Chunhua Weng, PhD[1]

[1]Department of Biomedical Informatics, Columbia University, New York City, NY
[2]Intermountain Healthcare, Salt Lake City, UT

Abstract

Electronic health records (EHRs) have been used as a valuable data source for phenotyping. However, this method suffers from inherent data quality issues like data missingness. As patient self-reported health data are increasingly available, it is useful to know how the two data sources compare with each other for phenotyping. This study addresses this research question. We used self-reported diabetes status for 2,249 patients treated at Columbia University Medical Center and the well-known eMERGE EHR phenotyping algorithm for Type 2 diabetes mellitus (DM2) to conduct the experiment. The eMERGE algorithm achieved high specificity (.97) but low sensitivity (.32) among this patient cohort. About 87% of the patients with self-reported diabetes had at least one ICD-9 code, one medication, or one lab result supporting a DM2 diagnosis, implying the remaining 13% may have missing or incorrect self-reports. We discuss the tradeoffs in both data sources and in combining them for phenotyping.

Introduction

The vast amounts of clinical data made available by pervasive electronic health records (EHRs) presents a great opportunity for reusing these data to improve the efficiency and lower the costs of clinical and translational research[1]. One popular use case is to identify patients for care management or research, prospectively, or as part of retrospective cohort for study. In this context, cohort identification using EHR data is known as EHR phenotyping.

The Electronic Medical Records and Genomics (eMERGE) consortium is a current multi-site research network sponsored by the National Institutes of Health of the United State. This network develops precise and portable phenotyping algorithms using heterogeneous EHR data[2]. To improve algorithm portability across different EHR systems, the design and evaluation of EHR phenotyping algorithms have relied on collaboration across institutions. For example, the eMERGE Type 2 Diabetes Mellitus (DM2) Case and Control algorithms were developed collaboratively by five institutions, resulting in the identification of over three thousand cases and controls to support a genome-wide association study (GWAS) on diabetes patients[3, 4]. The algorithm uses commonly captured EHR data elements for diagnosis, medications, and lab values to identify Type 2 diabetics. The emphasis on portability imposes a tradeoff due to the inherent data quality issues of those commonly captured EHR data elements. For example, ICD-9 billing codes are a coarse representation for nuanced narrative notes, medication orders do not necessarily reflect medication adherence, and as reported by Wei et al., EHR data fragmentation could negatively impact clinical phenotyping[5]. Moreover, while EHR data like lab values may be objectively correct, they may not actually reflect patient awareness of their own health status.

The eMERGE DM2 algorithm was originally validated using chart review. The expense of chart review typically limits sample size and only 50-100 each for cases and controls were reviewed in this example[3, 4]. Moreover, the chart review process does not sample from patients excluded from the case and control groups, meaning that a true sensitivity for identification of diabetes cases may not be established. Finally, chart review is still internal validation, implying the reference standard is still limited to information captured within the EHRs of related institutions[5]. Richesson et al. compared the identified individuals from different diabetes phenotyping algorithms[6]. While different algorithms might be created for different purposes, for example maximizing sensitivity for a registry versus specificity for a genetic study, the results do suggest that any given algorithm may fail to identify all diabetics in a database.

With the increasing emphasis on patient and community engagement for clinical research, self-reported diseases status has risen as an alternative data source for clinical phenotyping. These data are usually collected directly from patients, as opposed to EHR data that reflect the encounters of a patient with a single institution. Prior studies

checked the self-reported diabetes status against EHR data and achieved sensitivities around 0.75, and specificities around 0.9 [7-10].

While pieces of patient self-reported data have informed specific elements of clinical data used for phenotyping, such as self-reported smoking rate[11] and date of diagnosis[12], little is known about how self-reported disease status data might be useful for clinical phenotyping. Both EHR data and patient self-reported health data have advantages and disadvantages for patient identification. We faced an unusual opportunity to address this research question.

The Washington Heights/Inwood Informatics Infrastructure for Community-Centered Comparative Effectiveness Research (WICER) Project has been conducting community-based research and collecting patient self-reported health information[13]. A subset of surveyed individuals have clinical information stored at the Columbia University Medical Center, allowing direct comparison of diabetes status derived from clinical data to the self-reported diabetes status. Therefore, in this study we will validate the eMERGE DM2 Case algorithm using patient-reported diabetes status. This study is part of a larger research effort to use research data to verify clinical data accuracy.

Table 1: eMERGE DM2 algorithm criteria and definitions

Criterion	Definition	Query Terms
DM1 Diagnosis	Patient has ICD-9 codes indicating Diabetes Type I.	250.x1, 250.x3
DM2 Diagnosis	Patient has ICD-9 codes indicating Diabetes Type II.	250.x0, 250.x2 excl 250.10, 250.12
Control Diagnosis	Patient has ICD-9 codes indicating diabetes, conditions which may lead to diabetes, or family history of diabetes	250.xx, 790.21, 790.22, 790.2, 790.29, 648.8x, 648.0x, 791.5, 277.7, V18.0, V77.1
DM1 Medications	Patient has medication history for drugs treating Diabetes Type I.	insulin pramlintide
DM2 Medications	Patient has medication history for drugs treating Diabetes Type II.	acetoexamide tolazamide chlorpropamide glipizide glyburide glimepiride repaglinide nateglinide metformin rosiglitazone pioglitazone troglitazone acarbose miglitol sitagliptin exenatide
Control Medications	Patient has medication history for drugs treating diabetes.	Combination of DM1 and DM2 Medications
DM Lab	Patient has recorded lab value for HbA1c > 6.5, Fasting Glucose >= 126, Random Glucose > 200	HbA1c, Fasting Glucose, Random Glucose

Methods

1. Data Collected by WICER

Through cluster and snowball sampling methodologies, the WICER Community Survey collected data from residents in Washington Heights, an area of Northern Manhattan in New York City with a population of approximately 300,000 people. Surveys were administered to individuals over the age of 18 who spoke either English or Spanish. Survey data was collected and processed from March 2012 through September 2013. A total of 5,269 individuals took the WICER Community Survey in either the household or clinic setting.

The survey collected information about social determinants of health and health seeking behaviors as well as established some baseline health information. Survey participants were explicitly asked whether they had been told they had diabetes, high blood sugar, or sugar in the urine when not pregnant. The answer to this question was extracted as the self-reported diabetes status.

2. Data Collected by the Columbia University Clinical Data Warehouse

The Columbia University Medical Center's Clinical Data Warehouse (CDW) integrates patient information collected

from assorted EHR systems for about 4.5 million patients for more than 20 years. Commonly available structured EHR data elements include visits, medications, diagnostic codes, lab values, and clinical notes.

3. The eMERGE DM2 Case and Control Algorithms

As stated above, the eMERGE DM2 Case algorithm consists of three sets of criteria: diagnosis, medications, and lab values[4]. Diagnosis and medication criteria have components which indicate Diabetes Mellitus Type I (DM1) or Type II. Only patients with DM1 ICD-9 codes were completely excluded from the Case algorithm. For the purpose of this study, any patient reporting positive diabetes status who also had DM1 ICD-9 codes had their status reset to negative. DM1 medications only denote insulin dependence, which may also be found in DM2, and so additional logical criteria are required.

In contrast, the criteria for the eMERGE DM2 Control algorithm are very similar to the case algorithm, albeit inverted. Controls must have at least two visits recorded, a normal glucose measurement, and no evidence of either diabetes or conditions which might lead to diabetes. The other differences are that no effort is made to distinguish between the types of diabetes (i.e., I or II), and the range of ICD-9 codes for the diagnostic criteria is expanded to include observations that co-occur with Type 2 diabetes. Criteria and their definitions are presented in Table 1.

4. Cohort Identification

Patient data were extracted for every patient in the CDW for 2009-13. We chose this time window to replicate the time scale used by Richesson, et al. and to accommodate the fact that the medication data in our data warehouse are not complete prior to 2009. A subset of CDW patients who also have a WICER-recorded diabetes status was identified for validation of the eMERGE DM2 Case algorithm. The remainder of the CDW population was used to investigate potential differences between the self-reported population and the general data population.

5. Data Element Extraction for Each Cohort

Table 2: Patient level data variables and definitions

Variable	Definition
Sex	Sex of the patient.
Age	Age in years on 1/1/2014.
Visits	Number of visits between 2009 and 2013.
Span	Length of time in days between first and last recorded visit.
DM1 Diagnosis	Number of ICD-9 codes meeting the Diabetes Type I diagnostic criteria.
DM2 Diagnosis	Number of ICD-9 codes meeting the Diabetes Type II diagnostic criteria.
Control Diagnosis	Number of ICD-9 codes meeting the Control algorithm diagnostic and family history exclusion criteria.
DM1 Medication	Earliest prescription date for medication meeting the Diabetes Type I medication criteria.
DM2 Medication	Earliest prescription date for medication meeting the Diabetes Type II medication criteria.
Control Medication	Number of medication orders meeting the control algorithm exclusion criteria.
Glucose Tests	Number of glucose test values recorded for the patient.
Abnormal Labs	Number of lab results high enough to indicate diabetes.
Diagnosis Criteria	1 if the patient meets the diagnostic criteria for Diabetes Type II, 0 otherwise.
Medication Criteria	1 if the patient meets the medication criteria for Diabetes Type II, 0 otherwise.
Lab Value Criteria	1 if the patient meets the labs criteria for Diabetes Type II, 0 otherwise.
Case	1 if the patient is identified by the eMERGE Case algorithm, 0 otherwise.
Control	1 if the patient is identified by the eMERGE Control algorithm, 0 otherwise.
Survey Diabetes	1 for a positive patient-reported diabetes status, 0 otherwise. Exists only in Matched Data

Table 2 presents the variables and definitions required for cohort identification and comparison using the eMERGE Case and Control algorithms. For each patient in a dataset, the data elements in Table 2 were either extracted or calculated. The self-reported diabetes status for each individual was extracted from their survey response and included in the patient level data. For the purpose of this study, any patient reporting positive diabetes status who also had DM1 ICD-9 codes had their self-reported status reset to negative.

6. Analysis Plan

Several groups of patients were collected for comparison from both the subset of patients with self-reported diabetes status and general patient population. These groups are the patients identified by the eMERGE DM2 Case algorithm (eMERGE Case: Case = 1), the pool of potential cases meeting any of the diagnostic, medication, or lab value criteria (Case Pool: Diagnosis OR Medication OR Lab), and those patients meeting none of the criteria (Excluded: Not Diagnosis AND Not Medication AND Not Lab AND Not Control). For patients with self-reported status, patients responding "Yes" and "No" were also separated for analysis. The number of patients, fraction of patients who are female, average and standard deviation for age, number of visits, and time between the first and last recorded visit for each group were reported. For groups of patients with self-reported status, the number of patients identifying as diabetic was also reported. Summary values for each group were quantitatively described and compared.

Sensitivity, specificity, and positive predictive value against all patient self-reported statuses were calculated for the eMERGE DM2 Case algorithm, the component criteria individually (Diagnosis, Medication, Lab), the group of patients meeting all the criteria (Diagnosis AND Medication AND Lab), and patients meeting any of the criteria (Diagnosis OR Medication OR Lab). Sensitivity, specificity, and positive predictive value were also calculated for the eMERGE DM2 Case group using just the individuals identified by the paired Control algorithm.

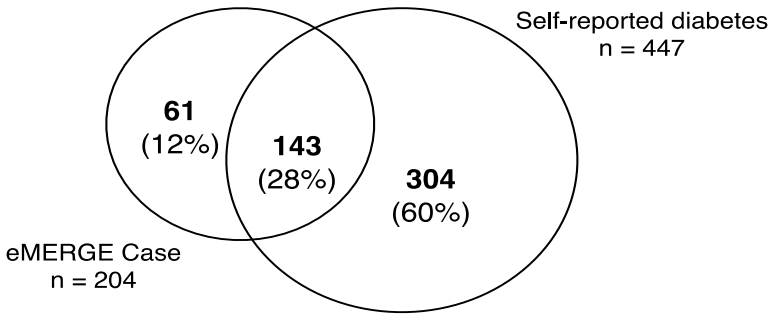
The eMERGE DM2 Case algorithm was expected to identify patients who do not report having diabetes, and not all patients reporting diabetes were expected to be identified by the algorithm. To investigate whether identification by the DM2 Case algorithm was a result of different subtypes of diabetes, with different patterns of comorbidities, all ICD-9 codes were pulled for each patient. ICD-9 codes were truncated at the root code level, or the whole number component of the code, and the frequencies of codes for each group were reported.

Results

We report our results in Tables 3-6, which includes summary statistics and demographics on specified patient groups, as well as validation statistics against all patient self-reported diabetes statuses and only those identified by the Control algorithm. See Figure 1 for a Venn diagram displaying the overlap between the patients identified by the eMERGE DM2 Case algorithm and those patients self-reporting positive diabetes status.

There were 2,249 WICER Survey participants with self-reported diabetes status who had at least one visit recorded at our institution within the last five years. Table 3 presents summary statistics and demography for patients reporting diabete s and no diabetes. The patients identified by the eMERGE DM2 Case algorithm (eMERGE Case), the pool of potential cases meeting any of the diagnostic, medication, or lab value criteria (Case Pool), and those patients meeting none of the criteria (Excluded) are presented for both the patients with reported diabetes status and the general population. In

Figure 1: Venn diagram of overlap between patients identified by the eMERGE DM2 Case algorithm and patients self-reporting positive diabetes status.



patients with self-reported status, eMERGE Cases and patients in the Case Pool are, on average, more than 15 years older than the Excluded group, and have twice as many recorded visits. The same difference is more than 24 years in the general patient population, with three times as many recorded visits. Patients with reported status are more likely to be female, as expected, but follow the same trend with regard to age and visits, albeit with 1.8-3.5x as many visits. While patients with reported status do tend to be older than the general population in general (46.1 vs. 36.4), those in the respective Case Pools are approximately the same age (61.8 vs. 61.0).

Table 4 shows the validation statistics against self-reported status. Sensitivity and specificity for the eMERGE phenotyping algorithm were .32 and .97, respectively, while positive predictive value was .70. The highest positive predictive value (.85) was achieved by requiring all criteria (Diagnosis AND Medication AND Lab). This combination also has the highest specificity (.98). While the highest sensitivity (.87) was achieved by the least restrictive combination (Diagnosis OR Medication OR Lab), the sensitivity of the combination requiring all criteria (.55) was still higher than that of the eMERGE algorithm.

Validation statistics were also computed for the eMERGE DM2 Case algorithm using only the eMERGE DM2 Control patients for comparison. These results are presented in Table 5. As a pair the DM2 Case and Control algorithms excluded 1,449 patients, reducing the pool of analyzable patients to 800. The majority of self-identified diabetes patients fell into the excluded group, which raised the apparent sensitivity of the eMERGE DM2 Case

Table 3: Cohort Demography and Characteristics. Groups are patients identified by the eMERGE DM2 Case algorithm (eMERGE Case), pool of potential cases meeting any of the diagnostic, medication, or lab value criteria (Case Pool), and those patients meeting none of the criteria (Excluded). Patients answering "Yes" or "No" to diabetes status are also presented.

Cohort	Group	N	Patient-reported Diabetes Count	Fraction Female	Average Age (SD Age)	Average Visits (SD Visits)	Average Time between First and Last Visit (SD Time)
Patient-reported Diabetes Status	**Yes**	447	447	0.76	62.0 (12.1)	40.3 (45.4)	1223.6 (665.3)
	No	1,802	0	0.79	48.0 (16.9)	24.4 (33.6)	1052.2 (654.6)
	eMERGE Case	204	143	0.72	62.4 (12.3)	34.8 (36.7)	1293.6 (568.5)
	Case Pool	670	387	0.76	61.8 (13.0)	43.3 (45.9)	1285.1 (520.1)
	Excluded + Control	1,579	60	0.79	46.1 (16.3)	20.9 (29.7)	1159.0 (564.9)
General Patient Population	**eMERGE Case**	25,310	n/a	0.50	65.8 (15.2)	18.7 (29.6)	902.1 (641.0)
	Case Pool	106,569	n/a	0.50	61.0 (21.3)	19.0 (32.2)	848.4 (649.7)
	Excluded + Control	680,324	n/a	0.58	36.4 (22.8)	5.8 (11.6)	677.3 (589.2)

algorithm to .93. However, the apparent specificity fell to .91.

The 15 most frequent ICD-9 codes for the intersections between the patients satisfying the eMERGE DM2 Case algorithm and the patients with positive self-identified diabetes status (+eMERGE +Self) are presented in Table 6. Codes for groups where the two methods disagreed (+eMERGE –Self, -eMERGE +Self) are presented in the same table as well as codes for the group of patients with no identification for diabetes (-eMERGE –Self). Note that DM1 and DM2 share the same root code (250) and no steps were taken to distinguish between types in this analysis. In general, the rank order of codes by frequency, as well as their general prevalence, is the same for the three diabetic groups regardless of how they were identified. The prevalence for diabetes ICD-9 codes is notably high in these groups. Prevalence for many of these codes is very different from patients without any indication of diabetes. Other comorbidities which are at least twice as prevalent in a diabetes group as in the non-diabetes group are hypertension, high cholesterol, diseases of the esophagus, and obesity. Patients with some identification for diabetes resemble the non-diabetic, general patient population in the prevalence of codes for follow-up examination, special investigations

Table 4: Positive predictive value, sensitivity, and specificity for the eMERGE DM2 Case algorithm using only the patients identified by the eMERGE DM2 Control algorithm.

Set	N	Patient-reported Diabetes Count	Positive Predictive Value	Sensitivity	Specificity
eMERGE Case	204	143	0.70	0.32	0.97
Diagnosis	517	369	0.71	0.83	0.92
Medication	320	260	0.81	0.58	0.97
Labs	549	330	0.60	0.74	0.88
Diagnosis AND Medication AND Lab	291	246	0.85	0.55	0.98
Diagnosis OR Medication OR Lab	670	387	0.58	0.87	0.84

Table 5: Positive predictive value, sensitivity, and specificity for the eMERGE DM2 Case algorithm, the component criteria individually (Diagnosis, Medication, Lab), the group of patients meeting all criteria (Diagnosis AND Medication AND Lab), and patients meeting any of the criteria (Diagnosis OR Medication OR Lab). All statistics were calculated against patient-reported diabetes status.

Set	N	Patient-reported Diabetes Count	Positive Predictive Value	Sensitivity	Specificity
eMERGE Case	204	143	0.70	0.93	0.91
eMERGE Control	596	11	n/a	n/a	n/a
Excluded	1449	293	n/a	n/a	n/a

Table 6: Prevalence of comorbidities for group of patients identified by the eMERGE DM2 Case algorithm (+eMERGE +Self), groups of patients where the two methods disagree (+eMERGE -Self, -eMERGE +Self), and the group of patients with no identification for diabetes (-eMERGE –Self).

ICD9 Root Code	Root Code Description	+eMERGE +Self (n= 143)	+eMERGE -Self (n = 61)	-eMERGE +Self (n = 304)	-eMERGE -Self (n = 1,275)
250	Diabetes mellitus	0.99	0.93	0.74	0.05
401	Essential hypertension	0.86	0.85	0.79	0.34
272	Disorders of lipid metabolism	0.65	0.67	0.63	0.21
786	Symptoms involving respiratory system	0.48	0.47	0.44	0.31
V67	Follow-up examination	0.46	0.44	0.48	0.44
V76	Special screening for malignant neoplasms	0.46	0.43	0.54	0.30
724	Other and unspecified disorders of the back	0.41	0.33	0.37	0.28
V72	Special investigations and examinations	0.39	0.39	0.49	0.47
789	Abdominal pain	0.38	0.43	0.39	0.34
780	General Symptoms	0.36	0.43	0.40	0.28
719	Other and unspecified disorders of joint	0.35	0.43	0.39	0.27
530	Diseases of the esophagus	0.35	0.36	0.29	0.16
729	Disorders of the soft tissue	0.34	0.33	0.39	0.23
278	Obesity	0.33	0.43	0.40	0.21
V04	Need for prophylactic vaccination and inoculation against single disease	0.31	0.49	0.48	0.25

or examinations.

Discussion

The results of the eMERGE DM2 Case algorithm, as well as its component criteria, was validated against all patients with self-reported diabetes status, prompting several points for consideration. We will discuss issues surrounding the generalizability of the patients with self-reported diabetes status to the general patient population, discrepancies between identification from the eMERGE DM2 Case algorithm and the self-reported statuses, and the potential contributions of patient self-reported data to EHR phenotyping.

Patient Comparison and Generalizability

One concern with this dataset is the patients with self-reported diabetes status, those who participated in the WICER Community Survey, are known to differ from the general population in several ways.

The group is older, containing more women, and is mostly Hispanic. However, the portion of these patients with positive indications for diabetes do resemble their counterparts in the general patient population in terms of age, and

the relatively increased number of recorded visits, as shown in Table 3. These findings suggest that the characteristics of patients with diabetes do not depend on the population from which they are drawn.

In Table 6, ICD-9 codes for diabetes are the most frequently represented in patients with some identification, either by the eMERGE DM2 Case algorithm or self-report, for diabetes, as expected. However, there are some discrepancies. The relatively lower prevalence of diabetes ICD-9 codes in the portion of self-reporting patients not identified by the eMERGE DM2 Case algorithm may indicate self-report inaccuracies or the effect of missing data in this group. The 5% prevalence of diabetes ICD-9 codes in the group with no identification for diabetes (-eMERGE –Self) may be a result of codes specific for DM1 which were filtered out by the DM2 case algorithm and not in that analysis.

Discrepancies in Identifying Diabetes

The eMERGE DM2 Case algorithm is known to perform well against case review and does achieve very high specificity in this evaluation. The algorithm performs less well in selecting all of the individuals who self-report having diabetes, and this may be for many reasons. First, the case algorithm is restrictive in order to limit the inclusion of DM1 patients. While steps were taken to exclude any patients who obviously had DM1, some of the patients who remain in the pool of potential cases may be rightfully excluded for this reason. Second, the non-selected patients may be incorrect about their diabetes status, though this is probably unlikely as this group of patients resembles the selected patients in patterns of visits and other demographics as well as the presence and frequency of comorbitidies. Moreover, if a large number of patients were in fact incorrect about their diabetes status, we would expect to see more discovered by the control selection algorithm. Lastly, and suggested by Wei, et al., the non-selected patients may be the product of data fragmentation, which is to say they do not have enough of their healthcare data consolidated in our system to allow identification by the eMERGE DM2 Case algorithm. For example, 83% of the self-reporting diabetic patients have at least a ICD-9 code for DM2 in our data warehouse, but at least 60% of those fail to be identified by the eMERGE DM2 Case algorithm for lack of sufficient clinical evidence for that diagnosis.

The more interesting group may be those patients selected by the eMERGE DM2 Case algorithm who do not self-identify as having diabetes. They have met the algorithm's stringent inclusion criteria, have visit patterns, other demographics, and comorbidities in common with the self-identifying diabetic patients, suggesting by very objective measures that they do have diabetes. That these patients seem to not be aware they have diabetes may have large implications to their treatment, adherence to that treatment, and their engagement with any treatment. Pacheco reported that only approximately half of the patients identified by the eMERGE DM2 algorithm at Northwestern had diabetes as part of the patient’s problem list, further suggesting that this effect is not confined to the patient[14].

Contribution of Patient Self-reported Data

There are pros and cons to both EHR data and patient self-reported data (Table 7) which point to how the two data sources might complement each other. EHR data is very heterogenous, with many data types, but that data may have issue such as missingness and inaccuracies that limit their secondary use for research. The more common elements have successfully been used for patient phenotyping algorithms, but that does not necessarily imply the algorithms have high sensitivity. In contrast, patient self-reported data reflects the patient's perception of their health status and may imply higher patient engagement in treatment, but may also be inaccurate and does not imply there is a useful quantity of clinical data at any one institution.

The best use of patient self-reported status may be augmenting EHR-based phenotyping algorithms. Phenotyping algorithms like the eMERGE DM2 algorithm typically require multiple criteria for successful identification of a disease and in our study the majority of patients who self-reported positive diabetes status did not have enough data in our system to be selected by the DM2 Case algorithm. Yet, 87% of them did have at least one ICD-9 code, medication order, or lab result to support a diagnosis of diabetes. If patient self-reported status could be standardized and used in addition to commonly captured EHR data elements for phenoyping algorithms, our study suggests the number of patients identified by such algorithms could be greatly increased.

This recommendation comes with two caveats, however. First, the contribution of patient self-reported status to phenotyping algorithms for research will depend on the needs of that research. If clinical data are important, as in a retrospective observational study, then patients who cannot be identified from their data alone may not be useful. Approaches such as the eMERGE DM2 Case algorithm would therefore be the best way to identify meaningful cases within a data source. On the other, if the goal is to simply identify as many patients with a disease or status as possible, for a potential prospective study or a GWAS, then self-reported data would be a valuable addition.

Table 7: Pros and cons of EHR and Patient self-report data sources.

	Data Source	
	EHR	Patient self-report
Pro	Heterogenous data types support high specificity. Common, standardized elements support portability.	Reflects patient perception. Might imply higher patient engagement.
Con	High rate of missingness. May only reflect encounter with a single institution	Does not imply useful quantity of clinical data. Patient perception may not be clinically accurate.

The second caveat is the issue of standardization. The portability of phenotyping algorithms relies on the use of common and standardized EHR data elements, such as ICD-9 codes. If the source of patient self-reported disease status is not standardized down to the exact wording of the question being answered, then the results may not be comparable and the resulting algorithm may not be portable. For example, the source of patient self-reported diabetes status in our study did not distinguish between DM1 and DM2. While steps were taken to address this limitation, the exact results of this study would probably be different if the survey question had specifically addressed DM2 alone. Therefore, any potential phenotyping algorithm built using our data might not perform the same on a data source with a patient self-reported data source specific to DM2.

Limitations

This study has several limitations. First, relatively few people were surveyed compared to the size of the large volume of patients in the EHRs. While the patients with self-reported status do appear to resemble identified cases from the general patient population, the population taking the WICER Community Survey is known to be older, and contain a higher proportion of women and Hispanic individuals. Weighting approaches exist which could be used to approximate the expected census distribution. These approaches were not used for two reasons. First, our prior research suggests the differences in measured variables are not large[15]. Second, the purpose of this paper was to explore the performance of the algorithm and we wished to leave its operation as transparent as possible. An additional limitation is that the WICER Community Survey does not distinguish between DM1 and DM2. While obvious DM1 cases were removed from the dataset, it is unknown what percentage of the remaining patients may have DM1.

Conclusions

There are pros and cons in both EHR data and patient self-reported health data. Phenotyping algorithms typically require multiple criteria for successful disease identification and may miss many patients with the disease in question. Likewise, self-reported health data does not imply sufficient EHR data to support a clinical diagnosis. If patient self-reported status could be used in addition to commonly captured EHR data elements for phenotyping algorithms, our study suggests the number of patients identified by such algorithms could be greatly increased.

References

1 *A First Look at the Volume and Cost of Comparative Effectiveness Research in the United States*. (2009). Academy Health.

2 McCarty, C. A., Chisholm, R. L., Chute, C. G., Kullo, I. J., Jarvik, G. P., Larson, E. B., Li, R., Masys, D. R., Ritchie, M. D., Roden, D. M., Struewing, J. P., Wolf, W. A., & e, Merge Team. (2011). The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med Genomics, 4*, 13. doi: 10.1186/1755-8794-4-13

3 Kho, A. N., Hayes, M. G., Rasmussen-Torvik, L., Pacheco, J. A., Thompson, W. K., Armstrong, L. L., Denny, J. C., Peissig, P. L., Miller, A. W., Wei, W. Q., Bielinski, S. J., Chute, C. G., Leibson, C. L., Jarvik, G. P., Crosslin, D. R., Carlson, C. S., Newton, K. M., Wolf, W. A., Chisholm, R. L., & Lowe, W. L. (2012). Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. *J Am Med Inform Assoc, 19*(2), 212-218. doi: 10.1136/amiajnl-2011-000439

4 Pacheco, J; Thompson, W. (2012). Type 2 Diabetes Mellitus. from http://phenotype.mc.vanderbilt.edu/phenotype/type-2-diabetes-mellitus

5 Wei, W. Q., Leibson, C. L., Ransom, J. E., Kho, A. N., Caraballo, P. J., Chai, H. S., Yawn, B. P., Pacheco, J. A., & Chute, C. G. (2012). Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. *J Am Med Inform Assoc, 19*(2), 219-224. doi: 10.1136/amiajnl-2011-000597

6 Richesson, R. L., Rusincovitch, S. A., Wixted, D., Batch, B. C., Feinglos, M. N., Miranda, M. L., Hammond, W. E., Califf, R. M., & Spratt, S. E. (2013). A comparison of phenotype definitions for diabetes mellitus. *J Am Med Inform Assoc, 20*(e2), e319-326. doi: 10.1136/amiajnl-2013-001952

7 Palepu, Prakruthi R., Brown, Catherine, Joshi, Gautam, Epsin-Garcia, Osvaldo, Eng, Lawson, Ramanna, Jayalakshmi, Hon, Henrique, Momin, Salma, Pringle, Dan, Cuffe, Sinead, Waddell, Thomas K., Keshavjee, Shaf, Darling, Gail Elizabeth, Yasufuku, Kazuhiro, de Perrot, Marc, Pierre, Andrew, Cypel, Marcelo, Xu, Wei, & Liu, Geoffrey. (2013). Assessment of accuracy of data obtained from patient-reported questionnaire (PRQ) compared to electronic patient records (EPR) in patients with lung cancer. *ASCO Meeting Abstracts, 31*(31_suppl), 40.

8 Kriegsman, D. M., Penninx, B. W., van Eijk, J. T., Boeke, A. J., & Deeg, D. J. (1996). Self-reports and general practitioner information on the presence of chronic diseases in community dwelling elderly. A study on the accuracy of patients' self-reports and on determinants of inaccuracy. *J Clin Epidemiol, 49*(12), 1407-1417.

9 Martin, L. M., Leff, M., Calonge, N., Garrett, C., & Nelson, D. E. (2000). Validation of self-reported chronic conditions and health services in a managed care population. *Am J Prev Med, 18*(3), 215-218.

10 Okura, Y., Urban, L. H., Mahoney, D. W., Jacobsen, S. J., & Rodeheffer, R. J. (2004). Agreement between self-report questionnaires and medical record data was substantial for diabetes, hypertension, myocardial infarction and stroke but not for heart failure. *J Clin Epidemiol, 57*(10), 1096-1103. doi: 10.1016/j.jclinepi.2004.04.005

11 Pacheco, J. A., Avila, P. C., Thompson, J. A., Law, M., Quraishi, J. A., Greiman, A. K., Just, E. M., & Kho, A. (2009). A highly specific algorithm for identifying asthma cases and controls for genome-wide association studies. *AMIA Annu Symp Proc, 2009*, 497-501.

12 Sciortino S, Walter L, Ranatunga D, Ludwig D, Schaefer C, Kay J, Jorgenson E. (2013). PS3-14: CREX: Utility of a Computerized Methodology to Identify Health Conditions Using EMR for GWAS, in the Kaiser Permanente Research Program on Genes, Environment, and Health. *Clinical Medicine and Research, 11*(3), 149.

13 WICER: Washington Heights/Inwood Informatics Infrastructure for Community-Centerede Comparative Effectiveness Research. from http://www.wicer.org

14 Pacheco, J. A., Thompson, W., & Kho, A. (2011). Automatically detecting problem list omissions of type 2 diabetes cases using electronic medical records. *AMIA Annu Symp Proc, 2011*, 1062-1069.

15 Fort D, Weng C, Bakken S, Wilcox A. (2014). Considerations for Using Research Data to Verify Clinical Data Accuracy. *Proceedings of the 2014 Summit on Translational Bioinformatics*.

Risk Prediction for Acute Hypotensive Patients by Using Gap Constrained Sequential Contrast Patterns

Shameek Ghosh[1], Mengling Feng, PhD[2, 3], Hung Nguyen, PhD[4], Jinyan Li, PhD[1,4*]
[1]Advanced Analytics Institute, University of Technology, Sydney, NSW, Australia;
[2]Massachusetts Institute of Technology, Cambridge, MA;
[3]Institute for Infocomm Research, Singapore;
[4]Centre for Health Technologies, University of Technology, Sydney, NSW, Australia

Abstract

The development of acute hypotension in a critical care patient causes decreased tissue perfusion, which can lead to multiple organ failures. Existing systems that employ population level prognostic scores to stratify the risks of critical care patients based on hypotensive episodes are suboptimal in predicting impending critical conditions, or in directing an effective goal-oriented therapy. In this work, we propose a sequential pattern mining approach which target novel and informative sequential contrast patterns for the detection of hypotension episodes. Our results demonstrate the competitiveness of the approach, in terms of both prediction performance as well as knowledge interpretability. Hence, sequential patterns-based computational biomarkers can help comprehend unusual episodes in critical care patients ahead of time for early warning systems. Sequential patterns can thus aid in the development of a powerful critical care knowledge discovery framework for facilitating novel patient treatment plans.

Introduction

Critical care patients in an intensive care unit (ICU) may undergo dynamic and rapid physiological changes, subject to various biological conditions. There are multiple symptoms which require immediate attention in an ICU. Among these, acute hypotensive events (AHE) are of great importance[1]. An AHE is defined as a drastic minimization of patient blood pressure for an extended period of time. It can be caused by shock, and it may lead to multiple organ failures. As a result, changes in hemodynamic conditions need to be detected as early as possible, so that effective medical interventions can be staged. Subsequently, the effectiveness of a medical intervention in an ICU can be assessed by the associated risk of mortality and the medical costs involved[1, 2]. Both these factors tend to rise with the passage of critical time. Hence, a medical intervention could be termed most effective if it has been staged pro-actively to prevent a shock, based on an early warning system. A pro-active intervention is contingent to obtaining clinical evidence of an impending event in a patient. Good clinical evidence may have two important properties viz. 1) predictive capability and 2) interpretability i.e. intelligible to a clinician, so as to make a quick decision. Interpretable evidence is probably the first step that helps a clinician decide the next course of action. The path to a good treatment plan is thus largely dependent on a clinician's strong understanding of the patient's physiological state, from the time the treatment was initiated. Discovery of such representatively simple yet effective clinical evidence requires the development of a powerful knowledge discovery framework which possesses the above properties and can process readily available streaming critical care data. Stream bed monitors, which continuously monitor physiological variables viz. arterial blood pressure, heart rate, pulse and blood temperature, generally have embedded rules to predict critical events, but they are also known to generate a lot of false alarms[3]. Such devices seldom take into account the sequences of micro physiological events and numerous associations among physiological variables in the course of the patient's ICU stay. Thus, a system that is capable of deriving events that are based on temporal relationships to predict future hemodynamic behavior can be highly beneficial to clinicians in various ways such as - 1) reduction in ICU operational costs and increase in efficiency, 2) in the development of novel goal directed therapies and 3) reporting of a patient's state for scheduling additional services[2].

Generally, multivariate analyses of existing data are very useful for ascertaining prognosis at population cohort levels and to holistically improve the efficiency of resource allocation in hospitals, but may not be of much help, in the context of a specific ICU patient with a rapidly developing critical condition[2]. In contrast, individual goal directed therapies could be administered to an ICU patient, based on dynamic analyses of streaming physiological data to help counter critical conditions in a shorter span of time[4, 5]. Typically, an array of machine learning methods has been employed in the past to make individualized patient predictions. Although, the theoretical novelties of

* Address correspondence to Jinyan.Li@uts.edu.au

these methods are substantial, yet their black-box nature impedes the model interpretability[6]. As a result, significant results in machine learning may not be transformed into interpretable solutions, which could be analyzed by clinicians towards developing evidence-based therapies. In contrast, identifying interesting structures from continuous streams of physiological data may be useful in providing a dynamic view of the patient's hemodynamic state[7] and help develop a testable hypothesis towards initiating a medical intervention.

In this study, we propose to mine interesting minimal sequential contrast patterns from blood pressure time series, which can be used as novel computational biomarkers in a predictive model to stratify the risks of patients for the onset of a future AHE. These informative sequential patterns are extracted from a given patient population, who had historical occurrences of AHE and/or had been administered pressor medications to stabilize blood pressure levels.

The importance of this work lies in the fact that long-term physiological time series data in ICUs that register episodes of sharp rises and falls in various sequences are extremely hard to be explored by general machine learning models. In contrast, sequential pattern mining methods can extract unique episodes of arbitrary length which may be over or under-represented in medical time series data. These episodes may not just be employed to determine the onset of an AHE, but can also gauge a clinical understanding of the types of episodes that are characteristic of specific critical conditions like AHE. . The motivation towards this study is thus driven by a requirement to generate novel medical insights in critical care, which is possible by the application of promising sequential pattern mining approaches to discover clinically relevant episodes connected by temporal relationships.

Our contributions in this work mainly involve - 1) the advancement of existing work in acute hypotension prediction using a sequential pattern mining approach, 2) a demonstration of the feasibility of extracting easily understandable distinguishing sequential patterns from physiological time series that could be effective in predicting AHE ahead of time, and 3) validation of sequential patterns using a large scale critical care database like MIMIC II[9], with the intention of developing novel pattern mining methods for making predictions in ICUs.

The Blood Pressure (BP) Prediction Problem

The objective of the BP prediction problem is to predict the arterial blood pressure (ABP) of patients in an ICU. Knowledge of the ABP state within normal function or hypertensive (high BP) or hypotension (low BP) regimes, in a future time window may turn out to be critical information. Typically, the ABP is a variable that registers strong correlation with the heart beat frequency within normal physiological limits and is measured in mmHg[8]. A derived variable known as the mean arterial pressure (MAP) is often used in medicine as the popular measure of blood pressure which can be defined as

$$MAP = \frac{2(diasPress) + sysPress}{3} \qquad (1)$$

In equation (1), *sysPress* (the systolic blood pressure) denotes the arterial blood pressure when the heart beats while pumping blood, whereas *diasPress* refers to blood pressure when the heart is at rest between beats.

For learning BP behaviour, as shown in the MAP time series in Figure 1, the learning window is shown as beginning from the ICU admission to T_0. Our aim is to predict the occurrence of an AHE in the prediction window demarcated by the period T_0 *to* T_0*+1 hour*.

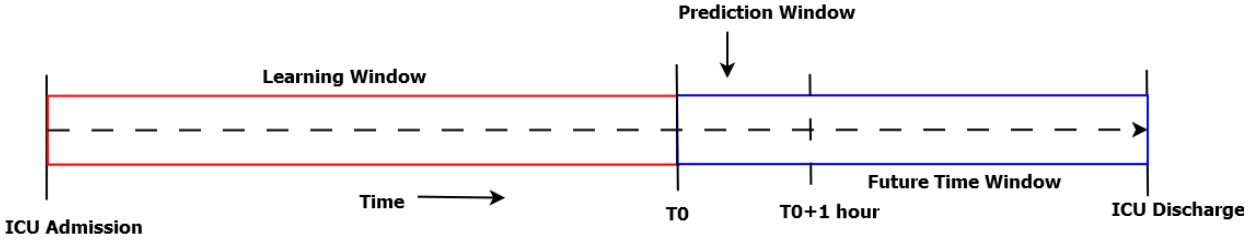


Figure 1: Mean Arterial Pressure time series with learning and prediction windows. T_0 indicates the time point following which, an AHE prediction is to be made. Typically the window after T_0, up to the patient's discharge, is part of a future time window. For purposes of prediction, one may choose to predict an event (an AHE) in a prescribed prediction window (i.e. from T_0 *to* T_0*+1 hour*).

Towards this objective, a dataset for a large number of patients has been made available as part of MIMIC II - a multi-parameter intensive care monitoring database[9]. MIMIC II consists of approximately 30,000 patients with ABP waveforms in various contiguous segments ranging in minutes to hours to even days that were recorded for patient

stays in the ICU. In the context of BP prediction, a subset of patient data with occurrences of acute hypotension was extracted from MIMIC-II and distributed as part of a set of challenges organized in 2009 by Physionet[8]. The details of the challenge datasets are provided in the following section.

Data Description

Acute Hypotensive Episode (AHE): Given a MAP time series, an acute hypotensive episode (AHE) may be defined as a period **of 30 minutes or more when 90%** or more of the MAP measurements are below the 60 mmHg regime. For training purposes, MIMIC II waveform signals were divided into two major groups viz. H and C, which indicated an occurrence of AHE in the forecast window and no occurrences of AHE in the forecast window. The groups H and C were further subdivided into H1, H2 and C1, C2. The definitions for each sub-group are as follows.

- ***H1***: Patients receiving pressor medication.
- ***H2***: Patients not receiving pressor medication.
- ***C1***: Patients with no acute hypotensive episodes during entire hospital stay.
- ***C2***: Patients having AHE before or after the forecast window.

Vital signs time series variables like heart rate (HR) and mean, systolic, and diastolic ABP were available for all the patient records. Some of the records often included respiration rate and SpO2. Although for the purposes of this study, we only consider the MAP time series (mean ABP) for each patient, as done in earlier studies. Accordingly, the prediction tasks consisted of the following two events -

- *Event I*: Patient risk classification between H1 and C1
- *Event II*: Patient risk classification between H and C

Dataset specifications are as given in Figure 2.

Number of Training samples for H and C	
H	***C***
H1 = 15	C1 = 15
H2 = 15	C2 = 15

Number of Test Samples for Event I and II	
H1 = 5, C1 = 5	*Event I*
H = 14, C = 26	*Event II*

Figure 2: Dataset Specifications for the AHE prediction problem

Especially in the case of *Event I*, the objective of the classification is to distinguish between two groups of patients, in which both received pressor medications. According to Moody and Lehman (2009), these two sets (H1 and C1) represented the extremes of AHE-associated risks.

Related Work

In recent years, a number of methods and scores have been suggested for understanding and predicting patient hemodynamic behaviour[7]. The Physionet 2009 AHE detection challenge served to advance such studies further by providing a test-bench for the use of neural-network based multi-models, static rules, support vector machines, histograms and statistical indices[10,11,12,13,14]. Although, these models extract informative features to construct off-line predictive models, they may be limited in their scope when analyzing real-time longitudinal medical data. Moreover, complete utilization of long range time series data may not be possible without powerful dimensionality reduction techniques. Typically, sequences of unusual spikes or falls in the physiological variables if detected may go a long way in the understanding of critical blood pressure behaviour that tends to occur before the onset of critical conditions. In a related context, Wang et al.[15, 16] indicated that pattern extraction from medical data is particularly challenging due to their sparse and longitudinal nature. Towards this aspect, the authors proposed a geometric image based mapping framework to mine temporal event signatures from large scale heterogeneous records of hospital patients. Moreover, popular methods like motif mining in bioinformatics have also been applied to cardiovascular time series for classifying medical events[17]. Accordingly, recent attention on data mining in healthcare indicate the shifts from traditional pattern mining problems to handling specific cases towards medical care improvements. Mining complex medical patterns could thus be more predictive of immediate outcomes, and may be able to report episodes soon after ICU admissions. From a clinical point of view, identifying such patterns are paramount, to detect

and understand the variability of a critical care patient's physiological response to an event like hypotension and thus help in the determination of the most optimal treatment combinations.

Methodology

In the following sections we provide an overview of sequential pattern mining and subsequently describe the processes involved in the discovery of sequential contrast patterns that were over-represented in the positive samples and under-represented in the negative samples. Sequential patterns are generally described in the form of short sub-sequences. Depending on the constraints of the formulated problem, representative sequences of certain lengths may be grown, which may demonstrate extensive repetitive behavior in the concerned datasets. In this context, sequential pattern mining techniques can provide a flexible and fast way to deal with streaming physiological data to derive interesting patterns which may provide important insights in the form of a chain of episodes, which could be of concern.

Symbolic Data Transformation

With the explosion of streaming real-valued physiological time series data generated from stream bed monitors in ICUs, it becomes essential to transform such data into symbolic representations for large scale data mining operations. Symbolic sequences can subsequently help avail a variety of existing pattern mining algorithms for efficient manipulation of such representations. Pattern mining algorithms typically employ techniques related to hashing, markov models, suffix trees, decision trees etc., which may be applied to symbolic sequences unlike real-valued continuous representations. In the last decade, the symbolic aggregate approximation (SAX) method has emerged as a popular technique which achieves efficient informative symbolization of large-scale time series data (for e.g. more than a billion time points) [18]. SAX first transforms the real-valued time series into a piecewise aggregate approximation (PAA) representation[19] and then converts the PAA series to a symbolic string. As claimed by the authors, the advantages of SAX involve that of dimensionality reduction and lower bounding. As a result, due to the nature of physiological time series generated over a number of days, and their importance in determining critical conditions, SAX provides a proper platform to create efficient indexing and pattern mining algorithms for medical purposes. Figure 3 illustrates a real-valued time series being converted to a sequence like *PQRQQPQ*, where *P, Q* and *R* are the three regions demarcated by the X and Y cut-points on the vertical axis.

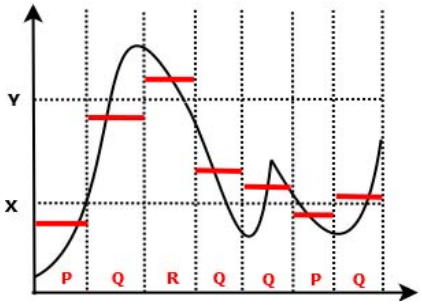


Figure 3: SAX Approximation of a time series. The given time series is symbolically represented as PQRQQPQ, using P, Q and R which denote three equiprobable regions.

Each MAP time series patient record before being discretized to a symbolic form goes through certain data transformation steps based on the SAX algorithm. This involves normalization of the time series to have a mean as 0 and variance as 1. Since normalized time series follow a Gaussian distribution, breakpoints are selected such that discrete symbols are equiprobable in the time series. For a normalized time series, this means that with five symbols, discrete regions are given by [-inf, -0.84, -0.25, +0.25, +0.84, +inf]. SAX thus adopts a symbolic representation which characterizes the inherent properties of the time series data. Hence SAX ensures that there exists an equiprobable distribution of symbols in the given time series.

Contrast Patterns

Mining contrasting patterns between groups of data under various labels was initially introduced as Emerging patterns by Dong & Li[20]. It is a sub-field of pattern mining, which typically aims to discover statistically significant patterns based on principles of frequency support, in various kinds of data viz. transactional, sequences, time series etc. Given that sequences are an important representation for data, minimal distinguishing subsequences (MDS) were later proposed as sequential contrast patterns[21]. Typically, an MDS is a sequential pattern that does not have a

subsequence, which itself is a sequential pattern satisfying algorithmic preconditions of minimum frequency support in the given data. In the following sections, we briefly describe the terminologies and processes associated with extracting MDS patterns.

Definitions and Terminologies

Subsequences: Let us define a set of items as $\boldsymbol{I} = \{i_1, i_2, i_3 \ldots, i_n\}$, also known as the *alphabet*. A sequence $\boldsymbol{S} = e_1\text{-}e_2\text{-}\ldots\text{-}e_n$ can be defined as an ordered collection of items belonging to $\boldsymbol{I}$. A sequence $\boldsymbol{A} = e_{k1}\text{-}e_{k2}\text{-}\ldots\text{-}e_{km}$ is said to be contained in $\boldsymbol{B} = e_1\text{-}e_2\text{-}\ldots\ \text{-}e_n$, such that $1<k_1<k_2<\ldots<k_m<n$. This means that the order of sequence $\boldsymbol{A}$ is maintained in $\boldsymbol{B}$, although the exact items may not be consecutive i.e. occurrences of A in B, may have gaps between the exact items, but the order will always be maintained. Thus if $\boldsymbol{A}$ is contained in $\boldsymbol{B}$, then $\boldsymbol{B}$ is a supersequence of $\boldsymbol{A}$. For example, given a set of symbols, $P=\{X, Y, Z\}$, XY is a subsequence of $XZYZ$, but not YX.

Suppose $\boldsymbol{D} = \{D_1, D_2 \ldots D_n\}$ be a set of sequences organized in a database. Now if there exists a sequential pattern $\boldsymbol{P}$, such that P is contained in '**q'** number of sequences in the database, then '**q'** is defined as the frequency support of $\boldsymbol{P}$, denoted as ***freqsup***.

Minimal Sequential Contrast Patterns: Given two sets of sequences $\boldsymbol{D}^+$ and $\boldsymbol{D}^-$, where $\boldsymbol{D}^+$ comprises of sequences with a positive class label and $\boldsymbol{D}^-$ is from a negative class label, we need to find the set $\boldsymbol{S}$ of all subsequences of a given length **L**, such that each subsequence $\boldsymbol{s_k} \in \boldsymbol{S}$ has the following properties:

i) $\boldsymbol{freqsup_{pos}}\ (\boldsymbol{s_k}) \geq \alpha$ (2)

ii) $\boldsymbol{freqsup_{neg}}\ (\boldsymbol{s_k}) < \beta$ (3)

iii) There is no subsequence of $\boldsymbol{s_k}$ *that satisfies (2) and (3) [**condition of minimality**]*

Here α is the minimum support required for $\mathbf{D}^+$ and β is the maximum support allowed for $\mathbf{D}^-$.

Thus given $\mathbf{D}^+$, $\mathbf{D}^-$, α and β, the MDS mining problem is concerned with finding all the minimal sequential patterns that are highly likely to occur in the positive set of samples but less likely to occur in the negative set.

Gap Constraint: While mining for a sequential pattern or a sequence, it may not be necessary for elements in a sequence, to occur consecutively. In such a scenario, defining a maximum gap constraint denoted as $\boldsymbol{g}$, allows the mining algorithm to search for a sequence, in which consecutive elements may have contiguous differences upto $\boldsymbol{g}$. For example, if $\boldsymbol{g}=2$, then XY is a subsequence of XZY but not $XZZZY$.

***Max-prefix*:** The max-prefix of a sequence **A** is the leading sequence of elements, without the final element of **A.** Thus the max-prefix of XZY is XZ.

Suppose, $\mathbf{D}^+ = \{XYZY, YZXY, YYZY\}$ and $\mathbf{D}^- = \{XZXZ, ZXXY\}$, are two sets of sequences. Let us consider ZY as a sequential pattern. Then for G=1, ZY has a frequency support of 3 in $\mathbf{D}^+$ and 0 in $\mathbf{D}^-$. In this context, if the positive (α) and negative (β) thresholds are set to 2 and 1, then ZY may be accepted as a sequential pattern.

Mining Minimal Contrast Sequences in Symbolic MAP Data

Towards the purpose of mining minimal contrast subsequences we employ the *ConSGapMiner*[21] algorithm, which is used to solve the MDS mining problem with gap constraints. The algorithm involves the application of a depth first search (DFS) technique to generate a set of distinguishing subsequences between two sets of sequences. Internally, the frequency support of each generated subsequence is computed for comparison with α and β, as per conditions (2) and (3) (the minimum positive and maximum negative support thresholds). After the set of all contrast subsequences are obtained, a post-processing step is employed to remove sequences that are non-minimal. These three stages of the algorithm are described next.

Generating Candidate Subsequences: Traditionally, every pattern mining algorithm begins with the candidate generation process. A candidate solution is one that can satisfy all the constraints posed by the concerned problem formulation. To generate candidate subsequences for MDS mining, DFS is performed to obtain a lexicographic sequence tree (LST) as shown in Figure 5. The DFS operation is a popular technique to grow rooted trees which have user-defined nodes. In the present context, each node in the LST (i.e. the DFS constructed tree) represents a sequence, along with its positive and negative frequency supports[21]. For example, in the LST in figure 5, node AAC(2,1) represents the sequence AAC with 2 as positive and 1 as negative supports. A child sequence may be grown by extending the parent sequence with a unique symbol from the alphabet, based on a certain lexicographic order. Thus, given the present LST, whose alphabet is defined as I = {A, B, C}, AAC may has three children nodes

as AACA, AACB and AACC. Subsequently each node's supports are computed from the positive ($\mathbf{D}^+$) and negative ($\mathbf{D}^-$) sample sets.

Pruning non-minimal distinguishing subsequences: After a sequence node is generated, if it satisfies the conditions (2) and (3), then the sequence is not extended any further. This is based on the fact that a supersequence of a distinguishing contrast sequence cannot be minimal[21]. Testing the minimality condition on generated sequences allows us to restrict the generation of redundant sequential patterns or superseqeunces. Handling minimality thus turns out to be extremely important since it can be a major factor in speeding up the mining process.

Pruning infrequent max-prefix extensions: If the current node's positive support is less than α, then we need not extend that node further. This is because, if a certain node fails to satisfy the condition in equation (1), its descendent nodes are also expected to be infrequent[21].

Checking Gap Constraints: Given a sequence XY, a gap constraint of g in a sequence P may be considered satisfied for XY if the sequential difference between the position of X and Y does not exceed **g**. For verifying if a generated candidate satisfies the gap constraint, a bitmap representation[22] is employed.

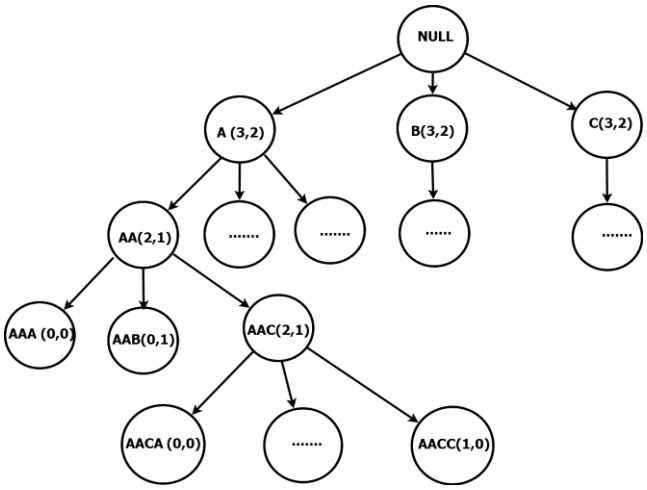


Figure 5: Part of a lexicographic sequence tree (LST), where the symbol alphabet consists of {A, B, C}.

Table 1. Checking gap constraint satisfaction of *XY* in *XZXZY*. The existence of a '1' in the last (7^{th}) row denoted by *AND* indicates that a maximum gap of 2 is satisfied by *XY*.

	X	**Z**	**X**	**Z**	**Y**
Index	*1*	*2*	*3*	*4*	*5*
$\mathbf{X_{index}}$	1	0	1	0	0
$\mathbf{1_X}$	0	1	1	1	0
$\mathbf{2_X}$	0	0	0	1	1
$\mathbf{1_X(OR)2_X}$	0	1	1	1	1
Y	0	0	0	0	1
AND	0	0	0	0	1

As shown in Table 1, we check for the gap satisfaction of *XY* in *XZXZY*, when maximum gap is set to 2. Initially, all the occurrences of X in the given sequence are set to 1(as shown in $\mathbf{X_{index}}$). In the present case, these are positions 1 and 3, in the given sequence. Later, (g+1) bits are set to 1 for each occurrences of X, separately as illustrated in rows 3(given as $\mathbf{1_X}$) and 4(given as $\mathbf{2_X}$). The bit vectors in rows 3 & 4 are then processed by the logical OR operator in row 5. Finally, a logical AND operation is carried out on the bit vectors in row 5 and for the occurrence of Y in row 6 to obtain a final sequence of bits, as in row 7. An occurrence of 1 anywhere in the final bit vector indicates that the gap constraint of 2 has been satisfied.

Post Processing: A post-processing step is finally applied to remove any sequence, which turns out to be a supersequence of at least another shorter subsequence in the resultant MDS set.

A flowchart illustrating the data flow architecture of the system is given in Figure 6.

Results and Discussion

The above described methodology was employed to mine sequential contrast patterns for the Physionet 2009 AHE prediction challenge datasets. Towards this purpose, we employed the SAX converted truncated MAP time series data from the given observation windows, for each patient in the given AHE training datasets. For event I, the training set comprised of H1 (for positive) and C1 (for negative), whereas for event II the whole dataset of size 60 was considered for training purposes.

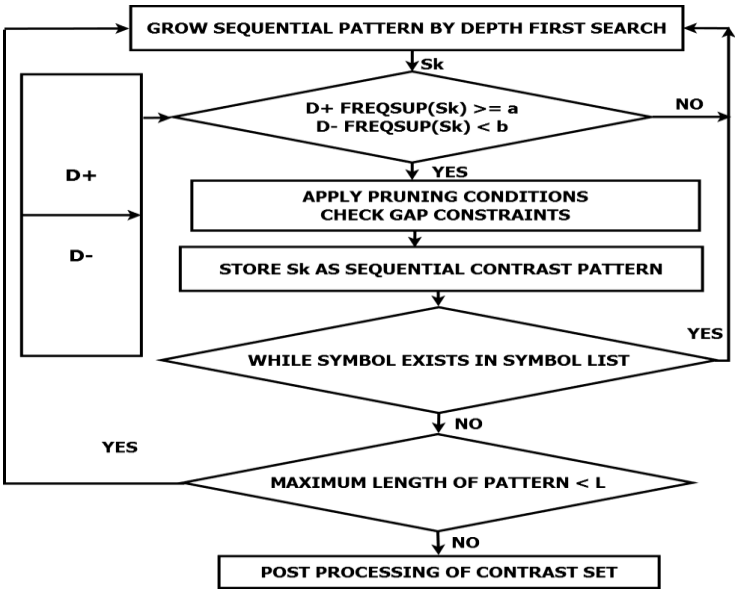


Figure 6: The Data Flow Architecture for mining sequential contrast sequences using a depth first search method (DFS).

In this context, an example training record like *a40439* consisted of T_0 indicated as 18.30 on 04/09/2008 (T_0 was provided with each record). The time series data prior to T_0 was thus used for training purposes (treated as the observation window). The extracted contrast sets for the two events were then applied to the test sets (as given in Figure 2). Each test segment comprised of an unlabeled MAP time series segment. If a test sample matched any one of the sequences in the contrast set, it was predicted as positive (H1 for event I and H for event II). A number of simulations were carried out with various parameters viz. subsequence length (L), alphabet size (S) and maximum gap (G). A 2 –fold cross validation (CV) was also performed using the larger training dataset consisting of 60 samples (30 – H and 30 - C), which reported a CV accuracy of 94.9%. In Table 2, the AHE test prediction accuracies have been given for event I and II, when the gap constraint was set to 3. As seen, the best performances were achieved using a maximum gap of 3, subsequence length of 10 and alphabet of cardinality 5.

Table 2. AHE Test Prediction Classification Accuracies while varying S and L for events I and II. Best performances are recorded for G=3, S=5, L=10/11 for event I and II.

G=3	**Event I**			**Event II**		
	S=3	*S=4*	*S=5*	*S=3*	*S=4*	*S=5*
L=5	5/10	5/10	7/10	19/40	19/40	32/40
L=6	5/10	5/10	7/10	19/40	19/40	32/40
L=7	5/10	5/10	7/10	19/40	19/40	32/40
L=8	5/10	7/10	7/10	23/40	23/40	32/40
L=9	5/10	7/10	9/10	23/40	25/40	33/40
L=10	5/10	7/10	10/10	25/40	32/40	36/40
L=11	5/10	7/10	10/10	25/40	32/40	36/40

Table 3: A Comparison of classification methods employed for the AHE prediction problem. Sequential patterns report comparable accuracies against existing methods.

Method	**Event I**	**Event II**
GRNN[11]	10/10	37/40
5-min average of diastolic ABP[10]	10/10	37/40
MAP averaging Rule [12]	10/10	36/40
5-min average of ABP[10]	10/10	36/40
Linear Regression[10]	10/10	36/40
Median of MAP[10]	10/10	34/40
NN with feature selection[12]	9/10	32/40
SVM[13]	10/10	30/40
RPS-NN[12]	2/10	25/40
Sequential Contrast Patterns	10/10	36/40

In Table 3, we provide a comparison of our results with the reported results from the Physionet 2009 challenge. As seen, models employing neural networks (GRNN, RPS-NN) and kernel methods like SVM are heavily dependent on several parameters, and can have performances over wide ranges[11, 12, 23]. Most of the other methods employed rules based on simple averaging measures and still performed fairly[10, 13]. Moreover, hidden markov models (HMM) for hypotension had reported a cross-validation accuracy close to 97%[24], which compares well with our cross-validation results too. Given this context, our results are comparable to the physionet 2009 challenge results. A general trend may be observed, where informative sequences could be extracted if the maximum gap constraint is iteratively increased. This is demonstrated by the Figure 7. The given results indicate that G=3 provides a good coverage to the length of the sequence.

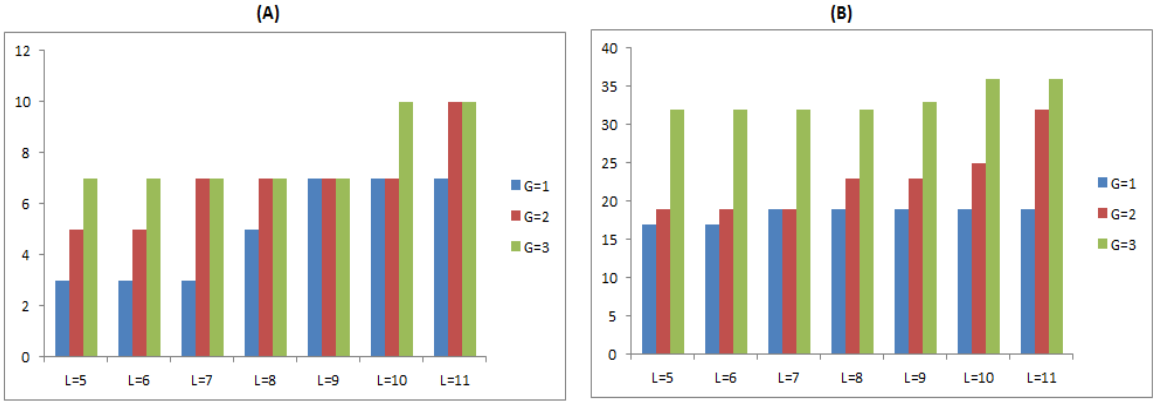


Figure 7: Effect of parameters L and G on the performance (A) For Event I, (B) For Event II.

It may be observed that performances tend to improve with larger values for gap sizes. At the same time keeping G too large, also means that two consecutive symbols in a sequential pattern may have occurred over a wide range, whose size was G. Extremely large gap sizes can thus impede a proper interpretation of contiguous events of importance. Finding an optimal value of G is therefore important to obtain meaningful predictions. Thus finding interesting sequences is highly dependent on the use of various parameters like the number of symbols, subsequence length and gap sizes. Based on the results, for detecting differential blood pressure patterns, obtaining optimal gap sizes may be more effective in reporting important episodes. For larger cohorts, finding out the optimal gap would thus be very important, given their dependency on the resolution of the time series (i.e. the sampling frequency). In addition, increasing S provides more number of discrete cut points for MAP, and enables the algorithm to capture patterns which characterize significant fluctuations in the BP. Thus, for cases with S=5, the algorithm is able to find a more expressive pattern, than for S=3. This also contributes to making improved predictions. Hence, selecting an alphabet size of 5 turned out to be an optimal choice, both in terms of the discretization of blood pressure range as well as keeping the algorithmic running costs within limits.

In contrast to our method, the 5 minute averaging measures are statistical features obtained from a 5 minutes window prior to the immediate occurrence of an AHE. Thus, a major difference lies in the fact that our method considers a wider window prior to the onset of AHE in comparison to just a 5 minute window using averaging measures[10]. This also indicates that a method which is effective in predictions within a 1 hour window may be more suitable in a real time scenario, in comparison to statistical measures obtained from a 5 minutes time window (prior to AHE). In this context, better results from 5 minutes prior to an AHE, may be due to temporal proximity to the onset of an AHE. For methods based on neural networks, both GRNN and RPS-NN report 10/10, 2/10 (for Event 1) and 37/40, 25/40 (for Event II). These methods tend to be strongly dependent on parameter tuning, as was also discussed by the authors[11]. The contrast mining method, on the other hand, helps to extract discretized sequential representations of the MAP time series, which provide the maximum support towards the occurrence of an AHE. These patterns are later useful, to not only predict an AHE for an unknown record, but may also be employed for further clinical interpretation by domain experts.

Our results indicate that sequential contrast patterns are capable of extracting informative symbolic episodes, which may be employed for both AHE risk prediction and understanding of hemodynamic behaviour towards effective analyses of sequential episodes that may be indicative of medical symptoms.

Clinical Significance of Sequential Contrast Patterns

In contrast to traditional machine learning models, the purpose of pattern mining is to extract hidden and interpretable knowledge from large amounts of data. In a typical medical care environment, the consumers of an important medical insight are both clinicians and computational systems. Machine learning methods like neural networks, SVM etc. which are heavily dependent on parameters may use extracted patterns as inputs/features to fine tune models, but do not crawl large-scale medical databases towards clinical knowledge discovery. Some of the sequential contrast patterns from the contrast sets extracted by the MDS algorithm are given in Table 4 for each of the two events.

Table 4: Examples of sequential contrast patterns for AHE prediction events

Event I	*DEDEDABCDC, DCEDCBCDCD, BCDCACDEDC, DCECACDEDE*
Event II	*CABAEECBCD, ABAEDBBBCA, ECBABABCD, ABBDECBCD*

Our approach was accordingly able to extract a sequential pattern like *ABAEDBBBCA,* which was prominent in acute hypotensive patients. Since the mean arterial pressure area was divided into 5 equiprobable regions (given by A , B, C, D, E), the above pattern indicates that the blood pressure signal follows a situation where majority of the AHE patients record an episode of events represented by the MAP value in the following order of blood pressure regimes - $A < B < A < E < D < B < B < B < B < C < A$. Extracting a pattern of this nature which may be regularly occurring in AHE patients, in contrast to patients with no occurrences of AHE , provides a sound basis towards carrying out a medical analysis of the given train of events. If later, we could derive the exact symptomatic signs displayed by a patient corresponding to this chain of events, we can establish a potential combination of observable physical indicators that precede AHE. Traditional machine learning models like SVM, neural networks are typically unable to extract such patterns that may lead to a significant discovery of a sequence of important stages or events which possibly lead to critical condition. Thus a higher frequency of the occurrence of complex contrast sequences while comparing hypotensive and normotensive patient groups may be beneficial to a clinician to develop a clinical hypothesis relating a succession of clinical events leading to an AHE. Thus, extracting sequential patterns from hypotensive patient groups can inform decision-making towards the diagnosis and investigation of AHEs. Moreover, the MDS method is flexible enough to accommodate clinician-defined constraints to detect specific types of patterns in medical care databases. Future work in this area, would concentrate on mining contrast patterns for larger populations with AHE. Although, earlier studies have reported good results using MAP, yet the development of real-time robust predictors using multivariate physiological data has still remained an open area. Additionally, a significant problem in predicting hypotensive events arises when the prediction window is placed further away from the observation window.

Conclusion

In this study, we present a novel application for mining contrast sequential patterns to predict acute hypotension in critical care patients. Mining sequential patterns has typically been a difficult problem for streaming long sequences. Moreover, this study aimed to introduce the extraction of contrast sequences for predicting critical conditions in ICUs. The importance of this work is associated with the introduction of a powerful sequential pattern mining based knowledge discovery process for analyzing time series data towards critical care predictions. In the future, we intend to enhance the existing framework for real time analyses while considering temporal associations and computational speed, in the context of specific issues in intensive care units.

Acknowledgements

Shameek Ghosh would like to thank Dr. Qian Liu, Jing Ren and Renhua Song from the Advanced Analytics Institute, the University of Technology, Sydney for their comments about the manuscript.

References

1. Roth RN, Idris AH, Fowler. Hypotension and shock. In Prehospital systems and medical oversight, 3rd ed., AE Kuehl, Ed. Dubuque: Kendall/Hunt, 2002.
2. Lilly CM, Cody S, Zhao H, Landry K, Baker SP, McIlwaine J et al. Hospital mortality, length of stay, and preventable complications among critically ill patients before and after tele-ICU reengineering of critical care processes. JAMA 2011; 305(21); 2175-2183.
3. Pinsky MR. Hemodynamic evaluation and monitoring in the ICU. CHEST Journal 2007; 132(6); 2020-2029.
4. Mayaud L, Lai PS, Clifford GD, Tarassenko L, Celi LA, Annane D. Dynamic Data During Hypotensive Episode Improves Mortality Predictions Among Patients With Sepsis and Hypotension*. Critical care medicine 2013; 41(4); 954-962.
5. Ehlenbach WJ, Cooke CR. Making ICU Prognostication Patient Centered: Is There a Role for Dynamic Information?*. Critical care medicine 2013; 41(4); 1136-1138.
6. Hart A, Wyatt J. Evaluating black-boxes as medical decision aids: issues arising from a study of neural networks. Informatics for Health and Social Care 1990; 15(3); 229-236.
7. Truijen J, van Lieshout JJ, Wesselink WA, Westerhof BE. Noninvasive continuous hemodynamic monitoring. Journal of clinical monitoring and computing 2012; 26(4); 267-278.
8. Moody GB, Lehman LH. Predicting acute hypotensive episodes: The 10th annual physioNet/computers in cardiology challenge. In Computers in Cardiology, 2009; 541-544.
9. Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman LW, Moody G, Heldt T, Kyaw TH, Moody B, Mark RG. Multiparameter Intelligent Monitoring in Intensive Care II: a public-access intensive care unit database. Crit Care Med 2011;39:952–960.
10. Chen X, Xu D, Zhang G, Mukkamala R. Forecasting acute hypotensive episodes in intensive care patients based on a peripheral arterial blood pressure waveform. In Computers in Cardiology 2009; 545-548.
11. Henriques J, Rocha TR. Prediction of acute hypotensive episodes using neural network multi-models. In Computers in Cardiology 2009; 549-552.
12. Mneimneh MA, Povinelli RJ. A rule-based approach for the prediction of acute hypotensive episodes. In Computers in Cardiology 2009; 557-560.
13. Fournier PA, Roy JF. Acute hypotension episode prediction using information divergence for feature selection, and non-parametric methods for classification. In Computers in Cardiology, 2009 ; 625-628.
14. Ho TCT, Chen X. Utilizing histogram to identify patients using pressors for acute hypotension. In Computers in Cardiology 2009; 797-800.
15. Wang F, Lee N, Hu J, Sun J, Ebadollahi S, Laine AF. A Framework for Mining Signatures from Event Sequences and Its Applications in Healthcare Data. IEEE Trans Pattern Anal Mach Intell 2013; 35(2); 272-285.
16. Wang F, Lee N, Hu J, Sun J, Ebadollahi S. Towards heterogeneous temporal clinical event pattern discovery: a convolutional approach. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining 2012; 453-461.
17. Syed Z, Stultz C, Kellis M, Indyk P, Guttag J. Motif discovery in physiological datasets: a methodology for inferring predictive elements. ACM Trans Knowl Discov Data 2010; 4(1); 2.
18. Lin J, Keogh E, Lonardi S, Chiu B. A symbolic representation of time series, with implications for streaming algorithms. In Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery 2003; 2-11.
19. Keogh E, Chakrabarti K, Pazzani M, Mehrotra S. Dimensionality reduction for fast similarity search in large time series databases. Knowledge and information Systems 2001; 3(3); 263-286.
20. Dong G, Li J. Efficient mining of emerging patterns: Discovering trends and differences. In Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining 1999; 43-52.
21. Ji X, Bailey J, Dong G. Mining minimal distinguishing subsequence patterns with gap constraints. Knowledge and Information Systems 2007; 11(3); 259-286.
22. Ayres J, Flannick J, Gehrke J, Yiu T. Sequential pattern mining using a bitmap representation. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining 2002.429-435.
23. Jousset F, Lemay M, Vesin J M. Computers in cardiology/physioNet challenge 2009: Predicting acute hypotensive episodes. Computers in Cardiology 2009; 36; 637-640
24. Singh A, Tamminedi, T, Yosiphon G, Ganguli A, Yadegar J. Hidden Markov Models for modeling blood pressure data to predict acute hypotension. In IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), 2010, (pp. 550-553). IEEE.

A Novel Method to Assess Incompleteness of Mammography Reports

Francisco J. Gimenez, BS[1], Yirong Wu, PhD[2], Elizabeth S. Burnside, MD, MPH[2], Daniel L. Rubin, MD, MS[3]

[1]Biomedical Informatics Training Program, Stanford, CA; [2]Department of Radiology, School of Medicine and Public Health, University of Wisconsin, Madison, WI; [3]Department of Radiology and Medicine (Biomedical Informatics Research), Stanford University, Stanford, CA

Abstract

Mammography has been shown to improve outcomes of women with breast cancer, but it is subject to inter-reader variability. One well-documented source of such variability is in the content of mammography reports. The mammography report is of crucial importance, since it documents the radiologist's imaging observations, interpretation of those observations in terms of likelihood of malignancy, and suggested patient management. In this paper, we define an incompleteness score to measure how incomplete the information content is in the mammography report and provide an algorithm to calculate this metric. We then show that the incompleteness score can be used to predict errors in interpretation. This method has 82.6% accuracy at predicting errors in interpretation and can possibly reduce total diagnostic errors by up to 21.7%. Such a method can easily be modified to suit other domains that depend on quality reporting.

Introduction

Breast cancer affects 1 in 8 women in the United States. It is the second leading cause of cancer deaths amongst women. Mammography has shown to be beneficial for early detection of breast cancer[1]. Currently, the American Cancer Society recommends that women over 40 with no specific risk for breast cancer get yearly screening mammograms to detect malignant findings early[2]. However, a major issue with mammography for breast cancer detection and management is the inconsistency and variability in practice, particularly in terms of variations in sensitivity and specificity of diagnosing malignancy[1,3-7]. Such variability is not limited to diagnosis. It has been shown that variability extends to report findings[2,8-10]. Variability in diagnosis and reporting hamper the utility of mammography: false negatives result in delayed treatment at the expense of patient health while false positives cause excessive additional invasive testing (e.g., biopsy), rising healthcare costs, and long-term psychosocial harm for women [11,12].

Decision-support systems have been developed to improve upon mammography interpretation and diagnosis[13-16], however most of these systems follow the *Greek Oracle* model of decision-support: they simply give an answer to the diagnostic task rather then assisting the radiologist to improve their own decision[17,18]. Additionally, such systems interrupt the traditional radiological workflow[19]. We posit that *improving the radiologist's report during reporting time* mitigates both of these issues and is the ideal time to deliver effective decision-support.

The mammography report is of crucial importance since it documents the radiologist's imaging observations, interpretation of those observations in terms of likelihood of malignancy, and suggested patient management, such as follow-up imaging or biopsy. Studies have shown the importance of good reporting practices and identified several key traits of good reports: correctness of findings, completeness of the description of significant clinical findings, consistency of report language and findings, and timeliness of the report's completion[20-22]. There are numerous efforts to improve the mammography report with respect to these traits. The Breast Imaging-Reporting and Data System (BI-RADS) provides a standard lexicon of descriptors and interpretation guidelines to improve consistency in language and correctness of findings[3]. Furthermore, structured reporting systems have been designed to improve clarity of presentation and reduce variability of reports between readers[23]. Despite these benefits, structured reporting is generally more time-intensive and can impair the traditional radiological workflow, directly interfering with timeliness[20]. Moreover, current approaches aim mainly to improve upon reporting language and clarity rather than report content and decision-making.

In this study, we propose a system that evaluates the content of the report and links it to errors in diagnosis. We do this by quantifying and measuring the incompleteness of the report findings with respect to abnormalities seen in images. We define incompleteness to be the sensitivity of the radiologist's decision to new information. Should gathering more data about a mammographic abnormality potentially change the radiologist's decision about clinical

management, a report is considered incomplete, and the radiologist can be alerted to provide more information to disambiguate report elements found to be inconsistent..

Mammography Diagnosis Problem

Radiologists presented with mammograms are tasked with two problems: detection and interpretation. Detection is the task of visually inspecting the mammogram and identifying abnormalities. Interpretation is evaluating whether detected abnormalities are suspicious for breast cancer. We will focus on the interpretation problem in this paper.

Formally, the interpretation problem is defined as follows: A radiologist is presented with a lesion in a mammogram, patient history and demographics, and possibly prior mammograms. The radiologist must decide whether this lesion warrants no action or follow-up (either imaging or biopsy) based on their suspicion of malignancy. This suspicion of malignancy is quantified as the BI-RADS assessment category, which is an ordinal value ranging from 1 to 6. An additional assessment category of 0 is used to indicate there is not enough information in the mammogram to make a decision. These assessment categories were designed to have probabilistic interpretations, where each value has a range of posterior probabilities of malignancy as shown in Table 1. A BI-RADS assessment of 1, 2, or 3 indicates the recommendation is no immediate follow-up (a negative assessment). A BI-RADS assessment of 4 or 5 indicates a recommendation for follow-up imaging or biopsy should be considered (a positive assessment). An assessment of 0 *should* not count as either positive or negative, but the fact that it necessitates immediate follow-up imaging means that it is treated as a positive finding [24]. BI-RADS 6 is a non-diagnositc category used to indicate that the images reflect a known cancer diagnosis being evaluated for treatment planning. These assessment categories implicitly mean that any lesion with a posterior probability of greater than 2% should be considered as a positive finding. Recent work has shown that this 2% threshold rule is justified via epidemiological risk analysis[25]. In addition to providing an assessment, radiologists must provide a report that justifies their decision. This report has a set of categorical descriptors standardized by BI-RADS, which can be interpreted as evidence for their decision.

BI-RADS Assessment	Probability of Malignancy	Description
0	N/A	Additional Imaging Needed
1	0%	No Abnormality
2	0%	Benign Finding
3	< 2%	Probably Benign Finding
4	2-95%	Suspicious Abnormality
5	> 95%	Highly Suggestive of Malignancy
6	100%	Biopsy Proven

Table 1: The BI-RADS assessment categories and their probabilistic interpretations.

Though BI-RADS assessments have objective probabilistic underpinnings, mammography interpretation is inherently subjective. Modern practice traditionally does not include quantitative estimates of these probabilities. Rather, radiologists provide the assessment categories based on training and experience. The use of BI-RADS assessment categories allows us to evaluate radiological performance as if radiologists are binary classifiers. We can measure their true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) as well as all associated statistics (e.g. positive predictive value, sensitivity, specificity). Moreover, the use of categorical descriptors allows us to build joint models of their decision given evidence.

Measuring Incompleteness

Missing descriptors in mammography reports do not necessarily mean that the report does not have all the information to make a correct and justified diagnosis. Conversely, there are cases when most of the descriptors are

reported, but the report still reaches an ambiguous diagnosis. Incompleteness of reports needs to be sensitive to the context of the information already given as well as the effect of missing information on the diagnosis.

Given that the final result of a mammogram is a decision whether to follow-up on patients with mammographic lesions, this decision should be the primary driver of determining whether enough information has been provided in the report. Early approaches to measuring whether medical diagnostic tests were necessary involved calculating thresholds for posterior probabilities that would warrant more testing or treatment [26]. Such methods required that the practicing physician provide the posterior probability. The Pathfinder system used a value of information calculation to repeatedly request more information for diagnosis until there was only one possible diagnosis left [27]. This did not provide a flexible framework for stopping if there was more than one possible diagnosis outside of physician judgment. The STOP criteria provide a quantitative algorithm for when to stop requesting information and make a decision, but this is formulated only to measure whether the probability of an event exceeds a certain threshold [28].

In response to these shortcomings for decision support systems, the same-decision probability (SDP) has been proposed [29]. This is defined to be the probability that a diagnostician will make the same decision they are currently considering given the unobserved information in a system. This metric has a nice intuitive meaning; only collect more information if it will change a decision. The SDP is defined for systems that make binary decisions based upon a posterior probability of an event being above a threshold. Formally, given a system with a decision function *D*, observed variables ***O***, unobserved variables ***U***, and a decision threshold *T* the SDP is defined as:

$$SDP = \sum_{u \in U} \mathbb{I}\left[D(\mathbf{O}; T) = D(\mathbf{u},\mathbf{O}; T)\right] Pr(\mathbf{u}|\mathbf{O})$$

Where I[•] is an indicator function that outputs 1 if true and 0 if false.

In the context of mammography, *D*(•) = diagnosis, ***O***= report data, ***U*** = Unobserved descriptors, and *T* = BI-RADS 2% threshold.

Though this is a well-defined metric, it is intractable to compute [30]. The summation requires iterating over all possible combinations of missing data which is an exponentially large search space. Despite this challenge, there are algorithms to approximate it based on statistical bounds on its value [29]. The drawback here is that such bounds can be weak under a variety of non-trivial cases. There is also an exact algorithm that can take advantage of certain Bayesian network structures to compute it in tractable time [30], but this method may break down for extreme value thresholds and Bayesian networks that do not have highly independent sets of unobserved nodes.

Here we propose a new method to compute an approximation of the SDP based on monte-carlo simulations. The difference between our approximation method and previous ones[29] is that they compute and exact value for an approximate bound on the SDP whereas we compute an approximation to the exact value of the SDP. This follows the advice of John Tukey, “It is far better an approximate answer to the right question, which is often vague, than the exact anwer to the wrong questions, which can always be made precise.” [31] Moreover, our approximation can be made arbitrarily accurate given more monte-carlo sampling steps.

For sake of convenience, we compute the complement of the SDP which is simply 1-SDP. This is the probability that our decision will *change* given new information. We will refer to this value as the *incompleteness score*. In this context, a *lower* value of the incompleteness score means the report is *more* complete.

Algorithm 1 Compute incompleteness score in Bayesian network

Input:
B: a Bayesian network
D: a diagnosis node
T: a decision threshold
N: an integer number of samples to use
O: a set of observed nodes
U: a set of unobserved nodes
Output: an incompleteness score value
Main:
d ← Pr(D=Malignant | **O**) > T
s ← 0
for i = 1:N **do**
u ← junction_tree_sample(B,**U**)
d_{new} ← Pr(D=Malignant|**O,u**) > T
if d_{new} != d **then**
s = s + 1
end if
end for
incompleteness_score ← s/N

Where join_tree_sample is the standard algorithm for sampling from a Bayesian network that has been compiled into a join-tree.

Experimental Methods

Study Design

Data used for this project were de-identified prior to analysis, and our work was thus not considered human subjects research. We acquired mammography report data from two teaching hospitals. Both of these institutions captured mammography findings using a structured reporting system (Mammography Information System, versions 3.4.99–4.1.22; PenRad, Buffalo, MN). Five attending radiologists read the mammograms at Institution 1. They reviewed 52,943 findings, of which 421 were malignant. These data were collected between April 5, 1999 and February 9, 2004. Eight attending radiologists read the mammograms at Institution 2. They reviewed 59,490 findings, of which 793 were malignant. These data were collected between October 3, 2005 and July 30, 2010. These datasets only contain 1 radiologist in common (though no interpreting radiologists were identifiable due to anonymization), the remainder of the radiologists in the two practices were distinct.

Analysis was done at the "finding" level, where a finding is defined as a set of observations about an abnormality in a mammogram, or the record for a mammogram with no abnormalities. Each finding can include patient demographic risk factors, BI-RADS descriptors characterizing an abnormality, BI-RADS assessment category, and pathologic findings from biopsy. Pathological ground truth was determined via matching patients with state cancer registries. By comparing the radiologist assessment to the pathological ground truth, we assessed whether a finding was a false positive (FP), false negative (FN), true positive (TP), or true negative (TN).

The structured reporting system separates masses, calcifications, and general findings. There is ambiguity in assessing when these three types of findings are associated with each other. In general, model builders can ignore the possible correlations since they do not seem to hamper performance of computer-aided diagnostic systems [14,15,32]. Unfortunately, we cannot make such relaxations of the model since we require that all descriptors provide meaningful information. As an example, descriptors specific to calcifications would spuriously affect incompleteness scores on mass findings. We chose to focus on analyzing mass findings to mitigate this issue. The resulting data set had 24,645 mass findings, 672 of which were malignant.

Masses were randomly split into two sets, 85% training and 15% testing. Training and testing groups were stratified by malignancy and care was taken to ensure patients with multiple masses were not represented in both groups. The training set was used to learn a tree-augmented naïve bayes (TAN) model for mammography diagnosis as described

by Burnside[32,33]. The incompleteness score was calculated for all masses in the test set using 5,000 monte-carlo samples with a decision threshold of 2% in concordance with BI-RADS recommendations. All model learning and classification was done in Norsys Netica 5.14.

Statistical Analysis

We stratified resultant incompleteness scores by radiological predictive categories (FP, FN, TP, and TN) to assess how the incompleteness scores differentiated between correct and incorrect evaluations. We quantified this difference by comparing the scores for errors (FP, FN) to the scores for correctly diagnosed findings (TP, TN) with a one-tailed Mann-Whitney U Test (aka Wilcoxon rank-sum test) [34]. The test was performed using the `wilcox.test` function in R version 3.0.2 (2013-09-25) -- "Frisbee Sailing."

Results

We trained a Tree-Augmented Naïve Bayes network on 20,950 training cases and measured the incompleteness score on 3,695 test cases. The resulting incompleteness scores were heavily right-skewed distributions. 83% of the incompleteness scores were equal to zero, meaning no new information would have changed the follow-up decision. Hence, both the median and mode incompleteness scores were zero. The mean incompleteness score was 0.021, but this is not a good indicator of group tendency since the large tail distribution has a disproportionate effect on the mean.

In order to verify that the incompleteness score can be used to predict mammographic error, we plotted its histogram and density estimate stratified by radiological predictive categories: true negative (TN), false negative (FN), true positive (TP), and false positive (FP) [Figure 1]. The graphs show that there are a large number of false positive and false negative cases that have non-zero incompleteness scores. Intuitively, this shows that incomplete reports have a higher likelihood of containing errors. The difference between error (FP,FN) and non-error (TP,TN) incompleteness scores was statistically significant ($p < 2.2*10^{-16}$).

An issue with this data is that there are a small number of false negative findings compared to false positive findings. This could skew results since positive findings may contain descriptors more prone to noise in the model. To account for this, we compared false positive to true positive results since both groups would have similar descriptors. The analysis showed that they were still statistically significantly different ($p<0.0026$).

We then tested how well the incompleteness score could predict error in mammography reading. Figure 2 shows several performance metrics for different cutoffs of the incompleteness score. The maximal accuracy with respect to cutoffs was 0.826 at a cutoff of 0.018. This means that > 1.8% probability of changing decisions when given new information should warrant describing more observations. The precision associated with this cutoff was 0.713 meaning 71.3% of cases classified as errors via the incompleteness score with cutoff 0.018 will actually be errors. Finally, we measured the percentage reduction in total mammography error if each error marked for revision was corrected. Using the given cutoff, we saw a potential 21.7% decrease in total errors.

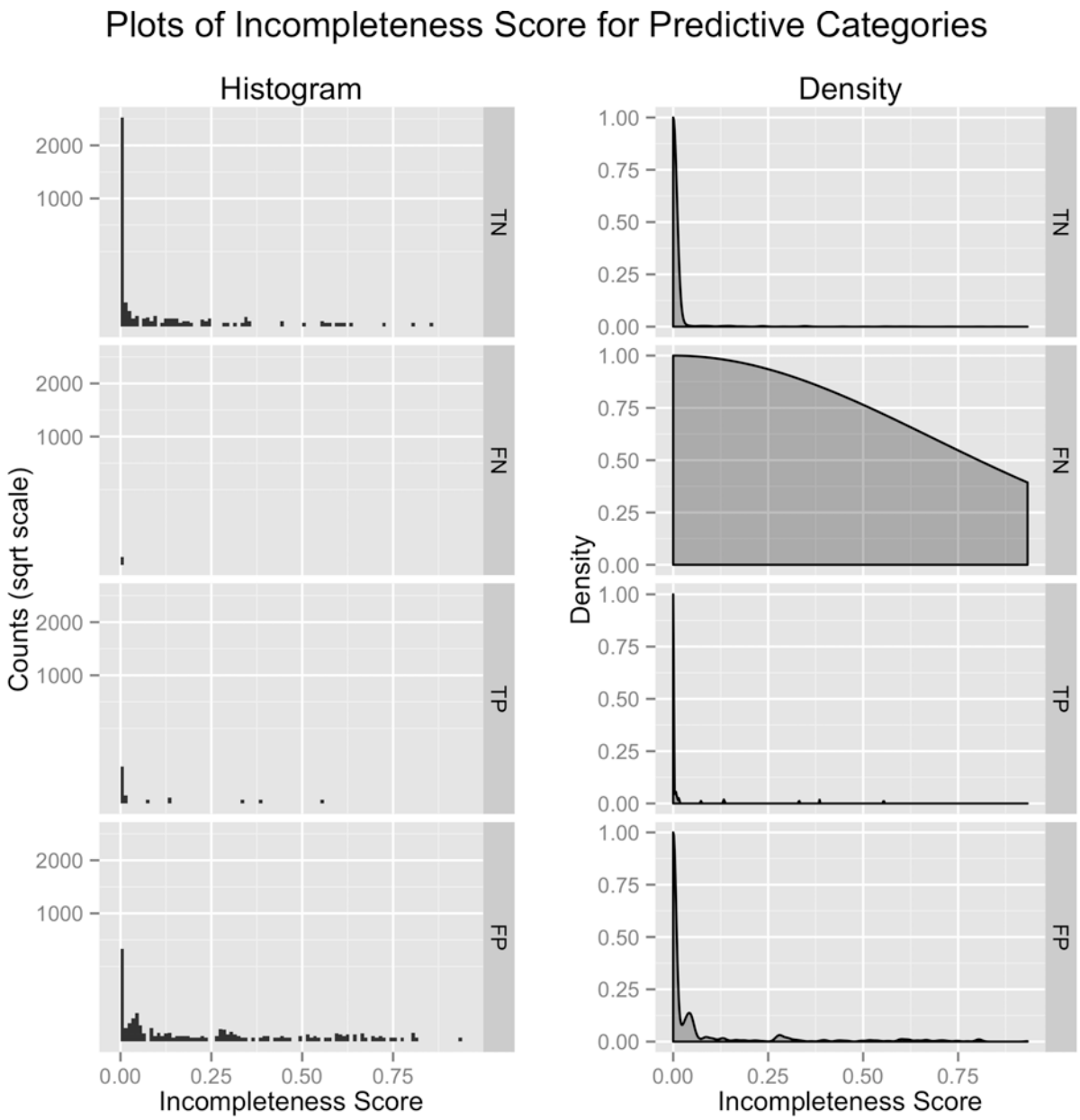


Figure 1: Plots of the histogram and density of incompleteness scores, stratified by radiologist performance on their respective cases. True Negative (TN) findings have the lowest incompleteness scores (indicating they are **most** complete) while false positive (FP) findings have higher incompleteness scores (indicating less completely reported findings). Joining (TN,TP) and (FP,FN), we can compare cases that were correctly assessed to cases with errors. Note that the false negative density graph has a nearly uniform distribution. This is an artifact due to the small amount of false negatives in the data set that skew density estimation.

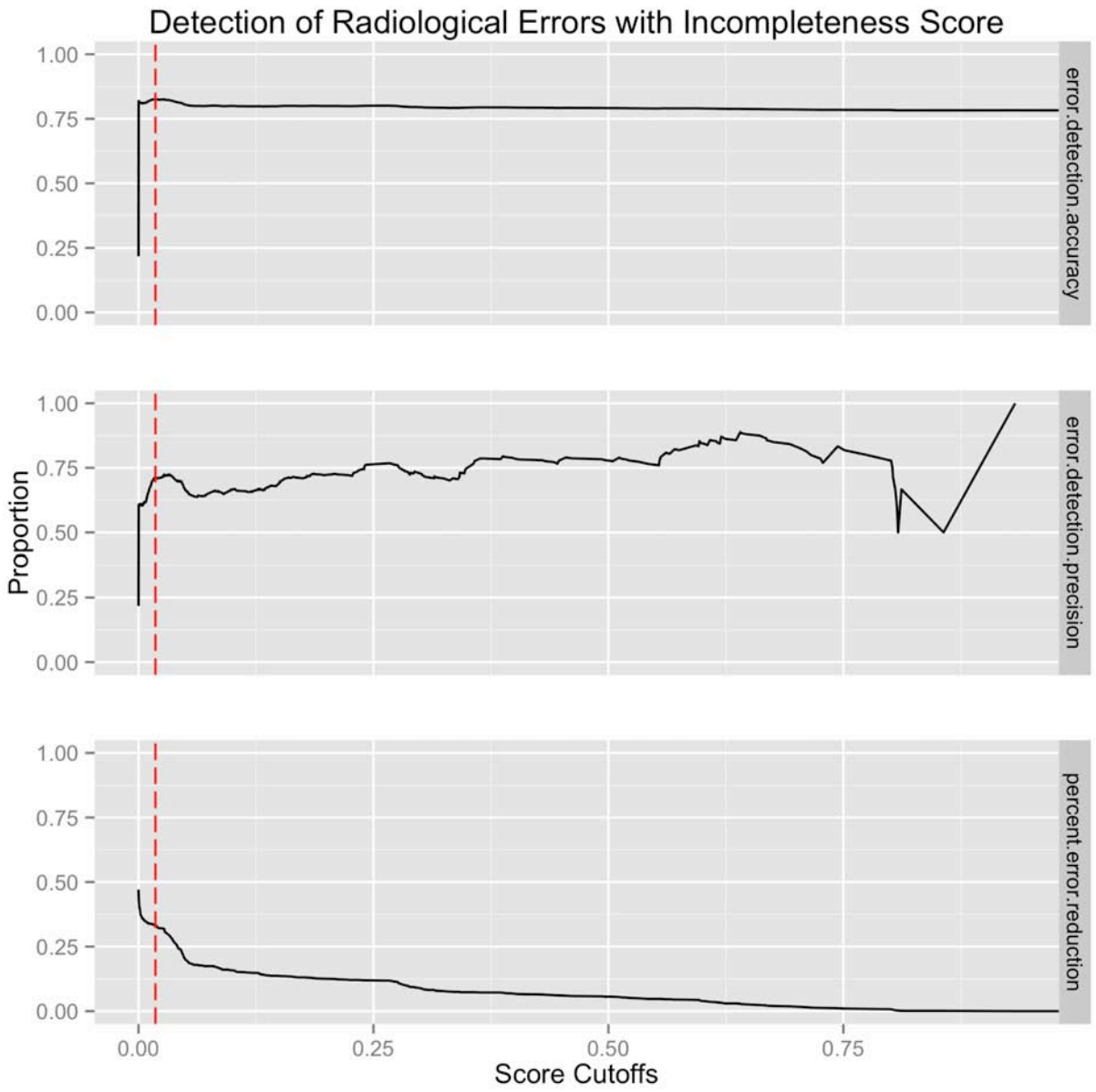


Figure 2: Improvement in radiological performance for difference incompleteness cutoffs. First row shows the incompleteness score accuracy in predicting radiological errors. Second row shows incompleteness score positive predictive value in predicting error. Third row shows the percent reduction in error if identifying error at the specified cutoff. The red-dotted line shows the cutoff point that maximizes error classification accuracy (first panel).

Discussion

We described a method to quantify how incomplete the content is in a mammographic report. We then presented an algorithm to measure this value in a computationally tractable manner. Finally, we showed that this *incompleteness score* is a strong indicator of errors in mammography interpretation. Implementation of this metric during mammography reporting time could provide a useful real-time feedback to radiologists to indicate possible errors.

Reporting in mammography is a labor intensive but critically important task for results communication. Though there is pressure for radiologists to expend their efforts efficiently, we show that poor report quality (measured by

our incompleteness score) is a marker for interpretation errors. This result might stem from a few different causes. The causal explanation is that poor interpretation leads to poor reports. In this case, a radiologist might have not seen a relevant descriptor in the image or neglected to highlight its importance. Another possibility is that incomplete reporting is a function of available time. When required to read large volumes of images, speed may decrease accuracy. In this case, a practitioner producing brief or incomplete reports may also be spending less time interpreting the image. A third possibility is that the process of reporting improves diagnosis by requiring radiologists to reason about their diagnosis. Thus, individuals who do not spend as much time on their reports do not go through the same formal thinking process. For future work, we will consult breast imaging radiologists on cases that were correctly classified as erroneous to see if humans can also identify when poor reporting leads to mis-diagnosis. If this is the case, we can begin to discover reporting practices that reduce error rates.

Though the system we present shows promising results with regards to predicting radiological errors, it does have some shortcomings. The use of an approximate algorithm to estimate the incompleteness score allows for some degree of error. We correct for this by using a large number of samples with respect to the number of hidden variables, but unfortunately, it is difficult to empirically evaluate our system as calculating the exact incompleteness score is prohibitively expensive with regards to computational time. For future work, we will evaluate alternative approaches for measuring incompleteness. Another issue with our system is that it does not actually correct the errors in interpretation or give any constructive feedback. So although the system can *potentially* reduce the amount of errors by ~20%, we have not shown which of these reports would actually be corrected. We plan to incorporate this into a clinical setting to measure the true impact of this decision-support system. Finally, this study was designed to be descriptive rather than predictive, so we did not measure classification results with an optimal cutoff in a third held-out test set. Thus, the results will be overly-optimistic in terms of error-prediction. In the future, we plan to implement our algorithm on faster cluster computers, which will allow us to perform a thorough cross-validation analysis to obtain better accuracy measurements.

Though we developed this system for mammography reporting, this methodology could be extend to any domain that uses expensive information to make threshold-based decisions. All this system requires is a generative model linking descriptors to diagnosis and a method to sample from this model. It is straightforward to implement this in any medical domain where testing can be a costly and/or risky task. Not only can this method improve diagnostic accuracy, but it inherently rewards good, thorough reporting practices. This is beneficial for patients and researchers alike.

Acknowledgements

Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under Award Numbers F31CA171789, U01CA142555, R01LM010921 and R01CA127379. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Nyström L, Andersson I, Bjurstam N, Frisell J, Nordenskjöld B, Rutqvist LE. Long-term effects of mammography screening: updated overview of the Swedish randomised trials. Lancet. 2002 Mar 16;359(9310):909–19.

2. Smith RA, Saslow D, Sawyer KA, Burke W, Costanza ME, Evans WP, et al. American Cancer Society Guidelines for Breast Cancer Screening: Update 2003. CA: A Cancer Journal for Clinicians. John Wiley & Sons, Ltd; 2003;53(3):141–69.

3. Baker JA, Kornguth PJ, Floyd CE. Breast imaging reporting and data system standardized mammography lexicon: observer variability in lesion description. AJR Am J Roentgenol. 1996 Apr;166(4):773–8.

4. Jackson SL, Taplin SH, Sickles EA, Abraham L, Barlow WE, Carney PA, et al. Variability of interpretive accuracy among diagnostic mammography facilities. J Natl Cancer Inst. 2009 Jun 3;101(11):814–27.

5. Beam CA, M LP, Sullivan DC. Variability in the Interpretation of Screening Mammograms by US Radiologists: Findings From a National Sample. Arch Intern Med. 1996 Jan 22;156(2):209–13.

6. Elmore JG, Miglioretti DL, Reisch LM, Barton MB, Kreuter W, Christiansen CL, et al. Screening

Mammograms by Community Radiologists: Variability in False-Positive Rates. J Natl Cancer Inst. 2002 Jan 18;94(18):1373–80.

7. Taplin S, Abraham L, Barlow WE, Fenton JJ, Berns EA, Carney PA, et al. Mammography Facility Characteristics Associated With Interpretive Accuracy of Screening Mammography. J Natl Cancer Inst. 2008 Jan 18;100(12):876–87.

8. Reiner B, Siegel E. Radiology Reporting: Returning to Our Image-Centric Roots. American Journal of Roentgenology. 2006 Jan 1;187(5):1151–5.

9. Hobby JL, Tom BD, Todd C, Bearcroft PW, Dixon AK. Communication of doubt and certainty in radiological reports. British Journal of Radiology. 2000 Jan 1;73(873):999–1001.

10. Robinson PJ. Radiology“s Achilles” heel: error and variation in the interpretation of the Röntgen image. British Journal of Radiology. 1997 Jan 1;70(839):1085–98.

11. Kerlikowske K, Zhu W, Hubbard RA, Geller B, Dittus K, Braithwaite D, et al. Outcomes of Screening Mammography by Frequency, Breast Density, and Postmenopausal Hormone Therapy. JAMA Intern Med. 2013 Mar 18;:1–10.

12. Salz T, Richman AR, Brewer NT. Meta-analyses of the effect of false-positive mammograms on generic and specific psychosocial outcomes. Psycho-Oncology. John Wiley & Sons, Ltd; 2010;19(10):1026–34.

13. Garg AX, Adhikari NKJ, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, et al. Effects of Computerized Clinical Decision Support Systems on Practitioner Performance and Patient Outcomes: A Systematic Review. JAMA: The Journal of the American Medical Association. American Medical Association; 2005 Mar 9;293(10):1223–38.

14. Burnside E, Rubin D, Shachter R. A Bayesian network for mammography. Proceedings of the AMIA Symposium. American Medical Informatics Association; 2000;:106–10.

15. Elizabeth S B. Bayesian networks: Computer-assisted diagnosis support in radiology1. Acad Radiol. 2005 Apr;12(4):422–30.

16. Rubin D, Burnside E, Shachter R. A Bayesian Network to Assist Mammography Interpretation. In: Brandeau ML, Sainfort F, Pierskalla WP, editors. International Series in Operations Research & Management Science. Boston: Springer US; 2005. pp. 695–720.

17. Miller RA, Masarie FE Jr. The demise of the “Greek Oracle” model for medical diagnostic systems. Methods Inf Med. 1990 Jan.

18. Friedman CP. A “Fundamental Theorem” of Biomedical Informatics. Journal of the American Medical Informatics Association. 2009 Mar 1;16(2):169–70.

19. Morgan MB, Branstetter BF IV, Clark C, House J, Baker D, Harnsberger HR. Just-in-Time Radiologist Decision Support: The Importance of PACS-Integrated Workflow. Journal of the American College of Radiology. 2011 Jul;8(7):497–500.

20. Weiss DL, Langlotz CP. Structured Reporting: Patient Care Enhancement or Productivity Nightmare? Radiology. 2008 Dec 1;249(3):739–47.

21. Johnson AJ, Ying J, Swan JS, Williams LS, Applegate KE, Littenberg B. Improving the quality of radiology reporting: A physician survey to define the target. Journal of the American College of Radiology. 2004 Jul;1(7):497–505.

22. Harald O S. Re: “Improving the quality of radiology reporting: A physician survey to define the target” (J Am Coll Radiol 2004;1:497–505). Journal of the American College of Radiology. 2004 Sep;1(9):700–1.

23. Reiner B. The Challenges, Opportunities, and Imperative of Structured Reporting in Medical Imaging. J Digit Imaging. Springer New York; 2009 Oct 9;22(6):562–8.

24. Barlow WE, Chi C, Carney PA, Taplin SH, D'Orsi C, Cutter G, et al. Accuracy of Screening Mammography Interpretation by Characteristics of Radiologists. J Natl Cancer Inst. 2004 Jan 15;96(24):1840–50.

25. Burnside ES, Chhatwal J, Alagoz O. What Is the Optimal Threshold at Which to Recommend Breast Biopsy? PLoS ONE. Public Library of Science; 2012 Nov 7;7(11):e48820.

26. Pauker SG, Kassirer JP. The Threshold Approach to Clinical Decision Making. N Engl J Med. Massachusetts Medical Society; 1980 May 15;302(20):1109–17.

27. Heckerman DE, Horvitz EJ, Nathwani BN. Toward normative expert systems: Part I. The Pathfinder project. Methods Inf Med. 1992 Jun;31(2):90–105.

28. Gaag L, Bodlaender H. On Stopping Evidence Gathering for Diagnostic Bayesian Networks. In: Liu W, editor. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg; 2011. pp. 170–181–181.

29. Choi A, Xue Y, Darwiche A. Same-decision probability: A confidence measure for threshold-based decisions. Fifth European Workshop on Probabilistic Graphical Models (PGM-2010). 2012 Dec 1;53(9):1415–28.

30. Chen S, Choi A, Darwiche A. An exact algorithm for computing the same-decision probability. Beijing, China: AAAI Press; 2013;:2525–31.

31. Tukey JW. The Future of Data Analysis. The Annals of Mathematical Statistics. Institute of Mathematical Statistics; 1962 Mar 1;33(1):1–67.

32. Burnside ES, Davis J, Chhatwal J, Alagoz O, Lindstrom MJ, Geller BM, et al. Probabilistic Computer Model Developed from Clinical Data in National Mammography Database Format to Classify Mammographic Findings1. Radiology. 2009 Jun 1;251(3):663–72.

33. Friedman N, Geiger D, Goldszmidt M. Bayesian Network Classifiers. Machine Learning. Springer Netherlands; 1997;29(2):131–63.

34. Wilcoxon F. Individual Comparisons by Ranking Methods. Biometrics Bulletin. American Statistical Association; 1945;1:80–3.

How do Interruptions Impact Nurses' Visual Scanning Patterns When Using Barcode Medication Administration Systems?

Ze He, MS[1], Jenna L. Marquard, PhD[1], Philip L. Henneman, MD[2,3]
[1]College of Engineering, University of Massachusetts, Amherst, MA; [2]Baystate Medical Center, Springfield, MA; [3]Tufts University School of Medicine, Boston, MA

Abstract

While barcode medication administration (BCMA) systems have the potential to reduce medication errors, they may introduce errors, side effects, and hazards into the medication administration process. Studies of BCMA systems should therefore consider the interrelated nature of health information technology (IT) use and sociotechnical systems. We aimed to understand how the introduction of interruptions into the BCMA process impacts nurses' visual scanning patterns, a proxy for one component of cognitive processing. We used an eye tracker to record nurses' visual scanning patterns while administering a medication using BCMA. Nurses either performed the BCMA process in a controlled setting with no interruptions (n=25) or in a real clinical setting with interruptions (n=21). By comparing the visual scanning patterns between the two groups, we found that nurses in the interruptive environment identified less task-related information in a given period of time, and engaged in more information searching than information processing.

Introduction

Medication errors are common and can adversely affect patients' health levels[1], with the medication administration process accounting for about one third of serous medication errors[2]. Health information technology (IT) has shown some promise in reducing medication errors[3,4,5]. Barcode technology, for instance, can be used during medication administration to help verify a patient's identity and the medication to be administered. Nurses can use the barcode medication administration system (BCMA) in conjunction with an electronic medication administration record system (eMAR) to automatically document the administration of medications. Because eMAR systems can retrieve records from either provider order entry or pharmacy systems, they may also reduce transcription errors[6], thereby improving patient safety.

Despite their benefits, BCMA systems can introduce errors, side effects, and hazards into the medication administration process[7,8,9], with previous studies identifying many contributing factors leading to undesirable outcomes[7-10]. As Harrison, Koppel, and Bar-Lev pointed out in their Interactive Sociotechnical Analysis framework, unintended consequences of health IT are often the result of interactions between health IT use and health care organizations' sociotechnical systems, including technical and physical infrastructure, social interaction, culture, and workflow[11]. For example, an organization overemphasizing the need for "complete" information may cause cognitive overload during health IT use[12].

Care providers' cognitive processes (for instance, visual scanning patterns) and external interruptions are important factors leading to medical errors[13,14]. However, the interactions between these factors during health IT use remain relatively underexplored. Grundgeiger and Sanderson (2009) reviewed research examining the nature of the relationship between interruptions and medical errors, and noted that much of the research in this area is descriptive. They suggest that future research use cognitive theories as a basis for empirical research[14]. Using an eye tracking device in an intensive care unit, Grundgeiger, Sanderson, MacDougall, and Venkatesh (2010) employed two cognitive theories – memory for goals and prospective memory theories – to empirically investigate how long it takes nurses to resume interrupted critical care tasks[15]. This work is important because it is one of the few studies to explicitly link nursing cognitive processes with an environmental factor (i.e. interruptions).

In our study, we similarly use data from an eye tracking device to study how interrupting factors impact nurses' cognitive processes, specifically their visual scanning patterns during medication administration. We use eye tracking data because visual scanning patterns are an important proxy for information intake and attention allocation. Additionally, Marquard et al. (2011) have shown that visual scanning patterns may be related to an individual's ability to identify patient identification errors[16].

In our study, we attempt to identify the impact of interruptions on nurses' visual scanning patterns. Our study builds off of a large body of work exploring how to measure and evaluate visual scanning patterns.

Researchers in this domain have developed a common set of definitions for visual scanning behaviors. An eye *fixation* occurs when the eye-in-head position is stable and focused on a specific reference point for at least 100-200ms. A *saccade* is the rapid shift of the eye from one point of fixation to another, typically lasting less than 50ms. Jacob & Karn (2003) provide a comprehensive review of tested methods to analyze visual scanning data[17].

Two studies analyzing individuals' visual attention to specific information in an environment are presented in Fitts, Jones, & Miton's (1950) pioneering work addressing pilot-aircraft interaction and in Goldberg and Kotval's (1999) work addressing human computer interaction[18,19]. Fitts et al. (1950) analyzed pilots' frequencies of eye fixations and lengths of fixations on each aircraft instrument. Goldberg and Kotval (1999) used a fixation/saccade ratio to capture individuals' time spent processing (fixations) computer interface components to their time spent searching (saccades) for the components. Higher fixation/saccade ratios indicated that individuals tended to do more information processing than searching for information when using the interface components.

In our study, we use Fitts et al.'s measures to capture which artifacts nurses fixate on *most frequently* while administering medications, and which artifacts nurses fixate on *longest* while administering medications. We use Goldberg and Kotval's measure to capture whether nurses have *different patterns of processing and searching for information* across artifacts during the medication administration process. With this knowledge, we may be able to design corresponding cognitive strategies to minimize the impact of interrupting factors, revise current health IT workflow protocols, and design simulated training experiences to better prepare nursing students for real clinical settings.

Methods

To study the impact of interrupting factors on visual scanning patterns, we compared nurses' visual scanning patterns in their typical interruptive environment (i.e., real clinical setting) with those patterns in a simulated setting, which was controlled to be free from interruptions (i.e., controlled setting). By comparing the visual scanning patterns of these two groups, we can start to distinguish how visual scanning patterns change in response to interruptions.

Settings and Participants

We conducted two experiments at a 600-bed, urban, level 1 trauma, pediatric and tertiary referral center in Western Massachusetts with an annual emergency department (ED) census > 100,000. During each experiment, we observed ED nurses as they completed the medication administration process while wearing an eye tracking device. Study A was conducted in a controlled setting[5], and Study B was conducted in a real clinical setting.

Participants in the studies volunteered to participate during one of their day or evening shifts. We told participants that the purpose of the study was to evaluate how expert nurses use visual cues to perform common patient care processes. The hospital's institutional review board approved the study. All participants read and signed an informed consent form, and had any questions answered by a study researcher.

Study A

Each nurse participant (n=25) performed a common patient care process – administering a medication – to a researcher acting as a patient. The patient only engaged in conversations with the participant when asked questions. Researchers asked each participant to perform the process the same way he or she does every day in the ED except for giving an actual medication. The patient had an ID Band on his right hand wrist, with the patient's name, date of birth (DOB), and medical record number (MRN) printed on the ID Band. A researcher led each participant to a computer with an electronic medication administration record system and a barcode scanner tied to the computer system. The user interface and functions of the system were the same as the hospital's actual system, but the computer did not connect to the hospital database and used local simulated patient data. The researcher also gave the participant an intravenous medication. The medication bag was labeled with the patient's name, DOB, MRN, medication name, medication dose, and a barcode.

Study B

Each nurse participant (n=21) performed the same patient care process as in Study A – administering a medication using barcode technology – but in a real clinical setting. All participants performed the processes on one real patient. Each participant was asked to perform the process in exactly the same way he or she did every day. The context for each process was different for each participant, yet in each case the participant had to perform some kind of medication administration process.

All participants performed the medication administration process using a standalone computer and either a portable tablet computer integrated with a barcode scanner (3/21) or a barcode scanner wirelessly connected to a computer (18/21). The medications each participant administered also varied, including intravenous medication, pills, and liquid medicines; most medications had a barcode on the bottle or bag.

Procedure

In both studies, participants wore an eye tracking device while completing the medication administration process. The ASL Mobile Eye (Figure 1; Applied Science Laboratories, Bedford, MA) is a tetherless eye tracking device that can be worn by participants who must move freely through a study environment. The eye tracking device weighs 76 grams, includes a scene camera, optics, and reflecting mirror all mounted on safety glasses. This eye tracking device can record both video and audio. To calibrate the eye tacking device for each participant, participants looked at twelve specific reference points in their field of view, with marks of their fixation adjusted to correspond to the reference points. After calibration, the software program for the eye tracking device overlays crosshairs (plus signs) at the exact locations in a scene where the individual is fixating throughout the scenario. With the head stable, the device is accurate to within 0.5 degrees of visual angle, with a resolution of 0.10 degrees of visual angle; the visual range of the eye tacking device is 50 degrees horizontally and 40 degrees vertically with respect to the head. The eye tracking device's scene camera records a video of the area in front of the wearer and uses pupil–corneal reflection to measure the position of the eye – sampled at 25 Hz.

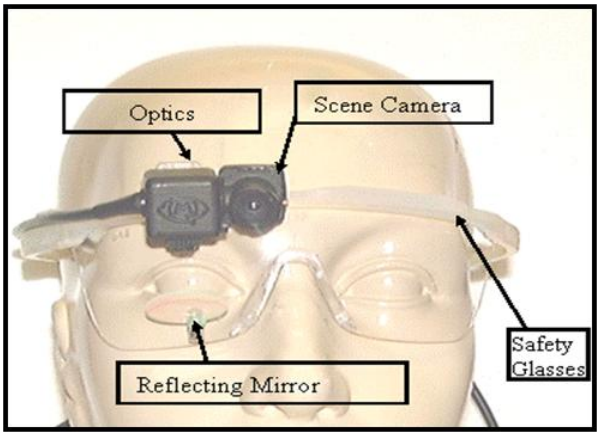


Figure 1. Eye tracking device

Two research team members independently reviewed all eye tracking videos, using standardized step names to code the steps each participant took to complete the medication administration process. The reviewers developed a standard coding policy and focused on the participants' medication administration related visual steps (i.e., fixations on artifacts including the ID band, medication labels, and computers) and interrupting events. The research team members determined a fixation occurred if the cross hairs on the video were within a one-centimeter squared box that covered a medication administration related information section on the pre-defined artifacts for 120 milliseconds. Another researcher resolved disagreements between the two initial coders. A shortened sample process execution is shown in Figure 2; a typical trial included about 20 to 30 steps.

We removed eye tracking data with insufficient quality to judge what artifacts the participant was looking at, e.g., the red crosshair did not show up to indicate where the fixations was located (n=3 in study A and n=2 in study B). We excluded data from one participant in study A who violated experimental protocols by changing the designed procedures of the experiment. We analyze his data separately because this change in the procedure may have affected his visual scanning process. We also removed two participants' data in study B from comparisons because they received a system warning message, but overrode the BCMA system; we analyze their data separately because we were specifically looking at the impact of external interruptions on the 'typical' process. Therefore, the numbers of participants included in our analysis are 21 in study A and 17 in study B.

Looked at Computer
Looked at ID Band
Looked at ID Band
Looked at Computer
Looked at ID Band
Looked at Computer
Looked at Computer
Looked at ID Band
Looked at Medication
Gave Medication to Patient

Figure 2. Example shortened process execution

Analysis of Visual Scanning Patterns

a) We measured how many medication administration-related fixations each participant completed per minute, calculated by their total number of fixations over the course of the trial divided by the duration of the trial in minutes. This measure approximates how much medication administration-related information a participant intakes in a given period of time.

b) Eye Fixation Distributions: Each nurse's eye fixation distribution is defined by the following formula:

$$\text{Eye Fixation Distribution on artifact A} = \frac{\text{\# of Eye Fixations on artifact A}}{\text{\# of Total Eye Fixations}}$$

This quantifier helps us to understand how nurses' visual attention is allocated across different artifacts (i.e., ID Band, Medication, Computer) throughout the process. This method is comparable to Fitts' frequency of eye fixation measure[18] and Jacob & Karn's number of fixations on each area of interest[17].

c) Eye fixation/saccade ratio: Each nurse's fixation/saccade ratio is defined by the following formula:

$$\text{Fixation/Saccade Ratio} = \frac{\text{\# of Total Eye Fixations}}{\text{\# of Saccades}}$$

This ratio quantifies how often a participant engages in information processing versus information searching between artifacts. If a participant engages in more fixations on one artifact before a saccade (a transition to another artifact), it implies the participant may do less searching for information across artifacts, instead dwelling on artifacts for longer periods of time for information processing. This method is comparable to Goldberg and Kotval's fixation/saccade ratio method[19], but in Goldberg and Kotval's work, they used durations of total fixations and saccades, while we used numbers of fixations and saccades.

d) Maximum Consecutive Eye Fixation: Each nurse's maximum consecutive eye fixation is calculated for each artifact. Similar to the eye fixation distribution measure, the artifact with the highest number of maximum consecutive eye fixations might be more heavily used, and thus may be a primary information source in the medication administration process. This method is comparable to Fitts' length of fixation[18].

Analysis of Interrupting Factors

We categorized interrupting events into three categories: a) completion of other tasks (e.g., transit between physical locations, other medication-related tasks such as ordering, transcribing, dispensing, etc.); b) unexpected artifact-related events (e.g., lost ID Band, barcode scanner failing to scan); and c) off-topic conversations with patients or colleagues (e.g., plans for the weekend). We measured the total durations of these interrupting events over the duration of the trial. Durations were recorded as the time from the start of a participant attending to the interrupting event to the time (s)he returned to the medication administration process.

Results

Figure 3 shows average number of fixations per minute completed by participants in the two settings, with the error bars representing the 95% confidence intervals for the values. We observed a significant difference between the two groups using a one tailed t test ($p<0.001$).

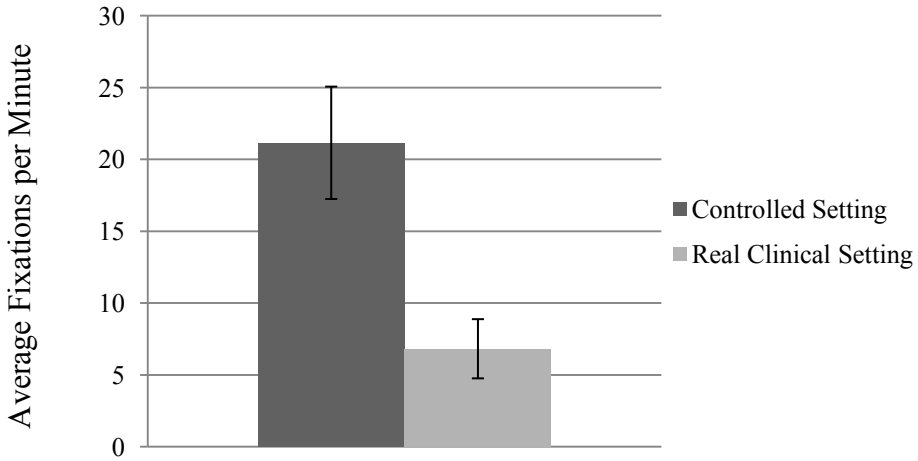


Figure 3. Average medication administration-related fixations per minute

Figure 4 displays the eye fixation distribution for each group. Both groups fixated on the computer most often, but nurses in the controlled setting fixated on the computer (77% of fixations) even more than nurses in the real clinical setting (64% of fixations) ($p=0.015$). Nurses in the controlled setting fixated on the ID band less (8% of fixations) than nurses in the real clinical setting (18% of fixations) ($p<0.001$). The two groups fixated on the medication label with similar frequency, with the percentages of fixations being of 15% for those in the controlled setting and 17% for those in the real clinical setting.

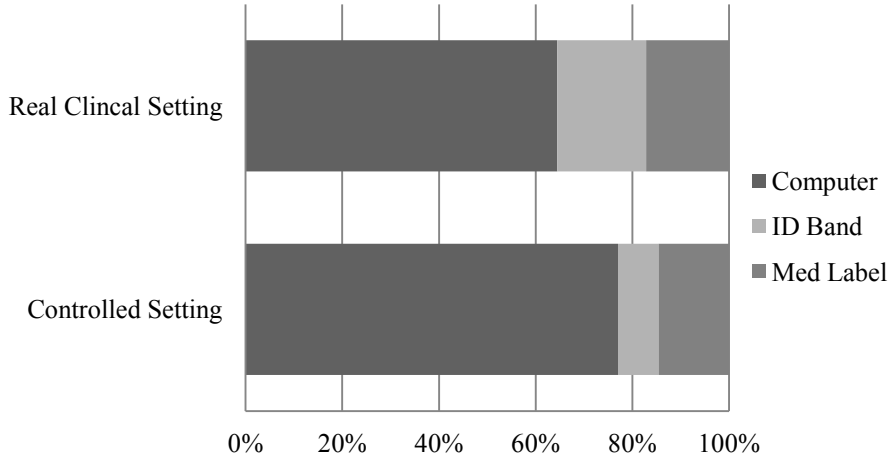


Figure 4. Eye fixation distribution for each group

Figure 5 shows the average number of fixations per saccade completed by the two groups, with the error bars representing the 95% confidence intervals for the values. There is a significant difference between the two groups using a one tailed t test (p=0.026).

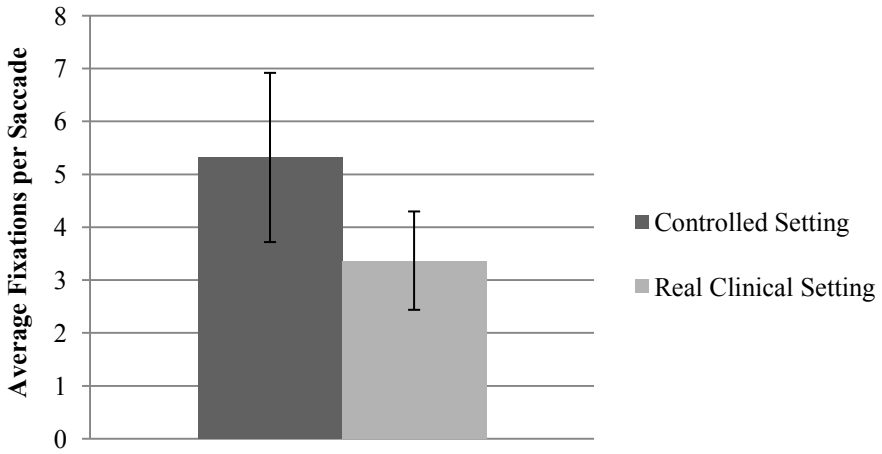


Figure 5. Average fixations per saccade for each group

Table 1 shows the average numbers of maximum consecutive fixations on each artifact for each group. Both groups had their longest average maximum consecutive fixation on the computer, but the nurses in the controlled setting had a much longer average maximum consecutive fixation on the computer (13.1 fixations) than the nurses in the clinical setting (5.2 fixations) ($p<0.001$). The nurses in the controlled setting also had a longer average maximum consecutive fixation on the medication label (1.8 fixations) than the nurses in the clinical setting (1.2 fixations) ($p = 0.04$). The average maximum consecutive fixations on the ID Band (1.4 fixations for nurses in the controlled setting; 1.6 fixations for nurses in the clinical setting) are relatively close for the two groups.

Table 1. Average maximum consecutive fixations on each artifact (confidence interval)

	Computer	ID Band	Medication Label
Controlled Setting	13.1 (9.0-17.1)	1.4 (0.9-1.9)	1.8 (1.4-2.2)
Real Clinical Setting	5.2 (3.6-6.8)	1.6 (1.2-2.0)	1.2 (0.7-1.7)

We identified 16 separate occurrences of interrupting events for 10 participants (59% of total participants) in the real clinical setting, summarized in Table 2. We recorded the durations of each interruption, but some interrupting events overlapped. For example, when completing another medication-related task (e.g., dispensing), a participant may also engage in off-topic conversations with his/her colleagues. In these cases, we count both types of interruptions but combined them into a single interruption time. The durations of each interruption ranged from 10 seconds to 164 seconds, with a mean value of 42.4 seconds (SD=34.5). In total, there were 12 times where nurses completed of other medication-related tasks, 2 times when unexpected events occurred with process artifacts (1 lost ID Band and 1 barcode scanner failure), and 7 off topic conversations with patients or colleagues.

Table 2. Summary of interrupting events in the real clinical setting

Interruption Category	Number of Nurses	Total Number of Occurrences	Description of Interruptions
Other tasks	8	12	• 8 nurses had to transit from the patients' bedside to the computer to complete the medication administration process, e.g., scanning barcode, documenting information • 4 nurses retrieved medications from the medication dispenser after they started the medication administration process
Unexpected artifact-related events	2	2	• One nurse had a medication that the scanner failed to scan, so she had to manually enter the medication administration information. • One nurse had a patient who lost their ID band, so she had to order a new one before she could continue to administer the medication.
Off-topic conversations	6	7	• 6 nurses engaged in off-topic conversations with their colleagues and/or patients.

Discussion

While previous studies made attempts to identify what contextual factors may be potentially interruptive, we used eye tracking video data to quantify how those factors may impact nurses' visual scanning patterns.

Participants in the real clinical setting tended to complete fewer fixations on medication administration-related artifacts in a given time period, therefore collecting less task-relevant information in the same time period. This may be attributed to the impact of various contextual factors that interrupt the participants' attention from attending to medication administration-related information. In our observation, we found that the medication process was non-linear in some cases, with overlap between steps. For example, we observed that some nurses engaged in medication dispensing during the medication administration process, which lengthened the primary task of medication administration. In addition, information sources were often more physically distant from one another in the real clinical setting as compared to the controlled setting, meaning nurses had to scan across larger distances in order the glean the same information.

Participants in the real clinical setting tended to allocate more fixations to the ID band and fewer to the computer. This may be due to the fact that in real clinical settings participants engaged in more direct interactions with patients. Yet, the bulk of fixations for participants in both groups (over 60%) were directed toward the computer. Therefore, special attention should be paid to the layout of the patient and medication information on this artifact, so that in real clinical settings, nurses can retrieve this information more efficiently. For instance, the patient identification information on the computer should be prominently displayed, whereas in many current systems they are in small print at the top of the computer screen.

Participants in the real clinical setting had fewer fixations on one artifact before making a saccade to another artifact than those in the controlled setting. Nurses in the controlled setting had significantly longer

maximum consecutive fixations on the computer, while both groups had similar consecutive fixations on the ID band and medication label. This may imply that the participants in the clinical setting engaged in more information search and less information processing when using the computer than those in the controlled setting, or that they were more likely to be distracted or interrupted while using the computer than at other times in the process.

This study suggests several possibilities for future research that could better determine mechanisms for process and technology redesign. For example, some nurses interrupted the medication administration process in a variety of ways (e.g., to dispense medications), making their visual scanning patterns more fragmented. Determining ways to reduce this type of fragmentation may allow nurses to more efficiently gather and use the information needed to complete their primary process. Technology-related changes may also significantly influence nurses' visual scanning patterns. Future studies could determine the effects of placing technologies physically close together so that nurses do not have to visually scan across large distances, or designing computer interfaces to better support nurses' visual scanning patterns. Finally, future studies may be able to identify especially effective visual scanning patterns, and train nurses to use those patterns.

Limitations

There are several limitations to this study. We could not fully control the variability in the real clinical setting (e.g., differing patients, differing medications) so cannot separate out the contribution of each factor on the differences in visual scanning patterns. Though our sample size was relatively small (n=21-25), there were statistically significant differences in our measures of interest. Additionally, this study was conducted at a single hospital, whereas participants at other institutions may exhibit different visual scanning patterns and follow different protocols. In this analysis, we did not account for verbal steps in the medication administration process, though auditory channels are also related to medication-related administration process. It is reasonable to think that if a nurse used a verbal method to verify a patient's identification, for instance, the nurse might conduct less visual information search. Finally, the coding of the eye tracking videos is somewhat subjective. To address this limitation, we had two independent evaluators review the videos, with another evaluator reconciling discrepancies between the initial evaluators' coding of the process steps. Yet, absolutely objective judgments are difficult to achieve.

Conclusions

This study showed that, on a variety of measures, nurses administering medications in a real clinical setting exhibited different visual scanning patterns than did nurses in a controlled setting. We assessed the nature and lengths of interruptions faced by nurses in the real clinical setting, to assess how these interruptions might be at least partially responsible for these differences.

Acknowledgements

This material is based upon work supported by NSF award 1150057.

References

1. Bond CA, Raehl CL, Franke T. Medication errors in united states hospitals. Pharmacotherapy. 2001;21(9):1023-1036.
2. Leape LL, Bates DW, Cullen DJ, et al. Systems analysis of adverse drug events. JAMA. 1995;274(1):35-43.
3. Hassink JJM, Jansen MMPM, Helmons PJ. Effects of bar code-assisted medication administration (BCMA) on frequency, type and severity of medication administration errors: a review of the literature. Eur J Hosp Pharm Sci Pract. 2012;19(5):489-494.
4. Bonkowski J, Carnes C, Melucci J, et al. Effect of barcode-assisted medication administration on emergency department medication errors. Acad Emerg Med. 2013;20(8):801–806.
5. Henneman PL, Marquard JL, Fisher DL, et al. Bar-code verification: reducing but not eliminating medication errors. JONA J Nurs Adm. 2012;42(12):562-566.
6. Poon EG, Keohane CA, Yoon CS, et al. Effect of bar-code technology on the safety of medication administration. N Engl J Med. 2010;362(18):1698-1707.

7. Cochran GL, Jones KJ, Brockman J, Skinner A, Hicks RW. Errors prevented by and associated with bar-code medication administration systems. Jt Comm J Qual Patient Saf Jt Comm Resour. 2007;33(5):293-301.
8. McDonald CJ. Computerization can create safety hazards: a bar-coding near miss. Ann Intern Med. 2006;144(7):510-516.
9. Patterson ES, Cook RI, Render ML. Improving patient safety by identifying side effects from introducing bar coding in medication administration. J Am Med Inform Assoc JAMIA. 2002;9(5):540-553.
10. Koppel R, Wetterneck T, Telles JL, Karsh B-T. Workarounds to barcode medication administration systems: their occurrences, causes, and threats to patient safety. J Am Med Inform Assoc JAMIA. 2008;15(4):408-423.
11. Harrison MI, Koppel R, Bar-Lev S. Unintended consequences of information technologies in health care—an interactive sociotechnical analysis. J Am Med Inform Assoc JAMIA. 2007;14(5):542-549.
12. Ash JS, Berg M, Coiera E. Some unintended consequences of information technology in health care: the nature of patient care information system-related errors. J Am Med Inform Assoc JAMIA. 2004;11(2):104-112.
13. Zhang J, Patel VL, Johnson TR, Shortliffe EH. A cognitive taxonomy of medical errors. J Biomed Inform. 2004;37(3):193-204.
14. Grundgeiger T, Sanderson P. Interruptions in healthcare: Theoretical views. Int J Med Inf. 2009;78(5):293-307.
15. Grundgeiger T, Sanderson P, MacDougall HG, Venkatesh B. Interruption management in the intensive care unit: Predicting resumption times and assessing distributed support. J Exp Psychol Appl. 2010;16(4):317-334.
16. Marquard JL, Henneman PL, He Z, Jo J, Fisher DL, Henneman EA. Nurses' behaviors and visual scanning patterns may reduce patient identification errors. J Exp Psychol Appl. 2011;17(3):247-256.
17. Jacob RJK, Karn KS. Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. Work. 2003;2(3):573-605.
18. Fitts PM, Jones RE, Milton JL. Eye movements of aircraft pilots during instrument-landing approaches. Aeronaut Eng Rev. 1950;9(2):24-29.
19. Goldberg JH, Kotval XP. Computer interface evaluation using eye movements: methods and constructs - Its psychological foundation and relevance to display design. Int J Ind Ergon. 1999;24(6):631-645.

A Method for Analyzing Commonalities in Clinical Trial Target Populations

Zhe He, PhD[1], Simona Carini, MA[2], Tianyong Hao, PhD[1],
Ida Sim, MD, PhD[2], Chunhua Weng, PhD[1]
[1]Department of Biomedical Informatics, Columbia University, New York, NY;
[2]Department of Medicine, University of California, San Francisco, San Francisco, CA

Abstract

ClinicalTrials.gov presents great opportunities for analyzing commonalities in clinical trial target populations to facilitate knowledge reuse when designing eligibility criteria of future trials or to reveal potential systematic biases in selecting population subgroups for clinical research. Towards this goal, this paper presents a novel data resource for enabling such analyses. Our method includes two parts: (1) parsing and indexing eligibility criteria text; and (2) mining common eligibility features and attributes of common numeric features (e.g., A1c). We designed and built a database called "Commonalities in Target Populations of Clinical Trials" (COMPACT), which stores structured eligibility criteria and trial metadata in a readily computable format. We illustrate its use in an example analytic module called CONECT using COMPACT as the backend. Type 2 diabetes is used as an example to analyze commonalities in the target populations of 4,493 clinical trials on this disease.

Introduction

An important area for clinical research standards development is participant selection for clinical studies. Many clinical studies are designed to emphasize internal validity, and some design decisions may compromise external validity, which is also referred to as generalizability. When a clinical trial has limited generalizability, the study results can be difficult to translate to the real-world population to which the study results are meant to apply. This is a concern of both the public and the clinical research community [1-3], and has significantly impaired the cost-benefit ratio of many clinical studies. Meanwhile, many clinical investigators prefer to reuse existing participant selection methods [4]. However, there is no well-accepted standard for participant selection.

To supplement the top-down model of traditional research standard development, we propose to use a data-driven approach to analyzing commonalities in participant selection for clinical trials. Leveraging the timely information on the Web, community-generated information has been shown to facilitate public health, e.g., real time prediction of epidemic disease [5]. We hypothesize that this method can identify frequent eligibility features that are in use in the clinical research community and represent *de facto* standards for reusable eligibility features. Understanding these practices can help the design and adoption of standards derived "from researchers and for researchers".

The official public registry of clinical trials and a valuable resource created by the National Library of Medicine of the United States [6], ClinicalTrials.gov presents a great opportunity for identifying frequently used criteria for research participant selection without expensive knowledge engineering by domain experts. Since September 2007, all clinical trials sponsored or conducted in the United States must be registered in ClinicalTrials.gov. As of 01/27/2014, 159,891 clinical trials with sites in more than 180 countries are registered, including 129,193 interventional studies and 29,061 observational studies. Strictly speaking, only interventional studies are called "clinical trials". In this paper, we use the term "clinical trials" to represent both interventional and observational studies. Trial summaries are semi-structured so that descriptive characteristics such as study title, sponsor, and target populations' ethnicity and gender are organized in structured fields, whereas eligibility criteria are written in free text, separated into inclusion criteria and exclusion criteria sections.

We have used ClinicalTrials.gov to identify common eligibility features for clinical research participant selection [7, 8]. As an extension to our previous work, this study integrated our previously developed parsing methods [9-11] to analyze commonalities in clinical trial participant selection. In particular, we designed and built a database called "Commonalities in Target Populations of Clinical Trials" (COMPACT) with searchable information for ClinicalTrials.gov's 159,891 entries as of 1/27/2014. COMPACT enables analysis modules, such as the CONECT (Commonalities in Target Populations in Eligibility Criteria) module described in this paper, to surface commonalities in clinical trial target populations from COMPACT. A use-case driven approach was employed to demonstrate how commonalities in clinical trial target populations can be derived from disease specific trials' inclusion and exclusion criteria and be leveraged to inform future clinical trial designs. This paper contributes to the research community (1) a novel method for analyzing commonalities in clinical trial target populations on the fly,

and (2) a new database to inform knowledge reuse for future clinical trial eligibility criteria designs. Both resources will be made open source in the near future.

Methods

In our design for COMPACT, we formulated the commonalities of target populations as common eligibility features, including numeric features, categorical features, and their attributes. Numeric features are eligibility criteria with a numeric value range requirement for participants, such as "HbA1c > 7%". Categorical features are eligibility criteria that accept one of a set of value options, such as "past history of stroke" (yes or no). We treat dichotomous (binary) criteria as a case of categorical criteria. To demonstrate the analytic utility of COMPACT, we developed an example analytic module called CONECT, which enables a user to mine contextual common eligibility features for trials on a certain disease from COMPACT. Next we will describe the details of both.

1. The Conceptual Design of COMPACT

The database Commonalities in Target Populations of Clinical Trials (COMPACT) includes four entities: *metadata of clinical trials*, *structured clinical trial eligibility criteria*, and *common eligibility features* (i.e., *numeric* and *categorical features*) *and their properties indexed by disease topics*.

Metadata of trials defines indexing characteristics of clinical trials provided by ClinicalTrials.gov, such as study type, intervention, medical condition, and study design (i.e., intervention model for interventional studies, allocation of participants to intervention group for interventional studies, and time perspective for observational studies).

Common eligibility features can be numeric or categorical. Each *common numeric feature* has five properties: numeric feature concept (e.g., HbA1c, BMI), the Unified Medical Language System (UMLS) semantic group [12] for the feature, the collective permissible value ranges derived from all the trials using this feature (e.g., [6.5%-7%] for HbA1c), the width of distinct mutually exclusive value intervals (e.g., 0.5 for [6.5%-7%] for HbA1c) and a salient modal boundary value if applicable (e.g., more diabetes trials use HbA1c of 7% as either a lower or an upper bound than any other threshold so that 7% is the modal boundary value).

Each *common categorical feature* has two properties: categorical feature concept and its UMLS semantic group. For example, the semantic group of "malignant neoplasm" is "Disorders". According to the NLM document [13], all 133 semantic types of the UMLS are grouped into 15 general semantic groups. "Disease or Syndrome" and "Neoplastic Process" along with 10 other semantic types are grouped into the semantic group "Disorders".

Figure 1 illustrates the information flow for a user, who can be a policy researcher, a clinical investigator, a trial sponsor, or others interested in such a system to interact with the COMPACT database. In this figure, the rectangle blocks represent data input and output when interacting with COMPACT. The round-corner blocks represent properties of input and output. The arrow lines represent processes when interacting with COMPACT.

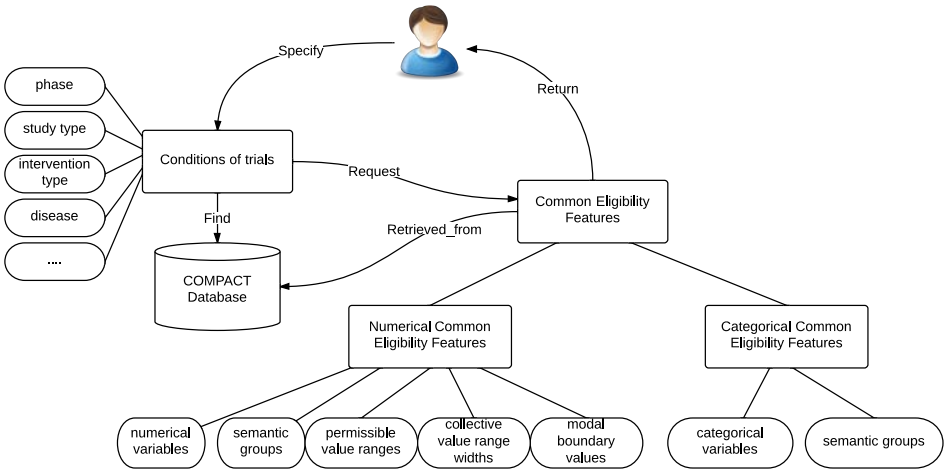


Figure 1. The information flow for a user to interact with the COMPACT database.

For example, when a user specifies certain queries of trials when interacting with the COMPACT database, e.g., *Type 2 diabetes trials recruiting patients with HbA1c >= 7.0 %*, he or she will retrieve the common eligibility

features with their attributes for these trials. Contextual attributes for common eligibility features present to a user how the clinical research community uses the feature HbA1c collectively in Type 2 diabetes trials. From COMPACT, three attributes for numeric features can be generated: collective value ranges, collective value range widths, and modal boundary values. Note that the data set for analyzing these three attributes comprises all the clinical trials of the given disease in the query.

Collective value ranges give the frequently used value ranges of a numeric eligibility feature for all the trials on a certain disease. We itemize all the permissible value ranges of a numeric feature in all the trials of a certain disease and count the number of trials that use a specific value range to get the collective value ranges. For example, in the inclusion criteria of the trial NCT00035984, the criterion "HbA1c value between 7.5% and 11%" gives the value range of HbA1c in this criterion: [7.5, 11.0], in which "[ ]" means being inclusive.

Collective value range widths present the distribution of value range widths of a numeric eligibility feature for all the trials on a certain disease. To get the collective value range widths, we counted the number of trials of a certain disease with the same value range width of a numeric feature. For example, the value range width of HbA1c derived from the same criterion "HbA1c value between 7.5% and 11%" is: 11 – 7.5 = 3.5.

Modal boundary values are defined as the "most-used boundary values" for a numeric feature for eligibility determination. For example, because more diabetes trials use HbA1c of 7% as either a lower or an upper bound, "7.0%" is the modal boundary value of HbA1c for Type 2 diabetes trials. Similarly, "140 mm Hg" is a modal boundary value for systolic blood pressure in the same data set. Note that for some numeric features, there might not be a modal boundary value.

2. COMPACT Database Construction

Figure 2 presents the workflow for constructing the COMPACT database. To enable agile discovery of common eligibility features, we downloaded all the trial summaries in ClinicalTrials.gov as of January 2014, excluding those with no or non-informative eligibility criteria text, such as "please contact site for information". The trial summaries were parsed and saved in the relational database (RDBMS) MySQL. For each trial, using previous developed parsing methods [9-11], we extracted, parsed, and stored the metadata of the trial (e.g., title, location, type, etc.), numeric features, and categorical features in three database tables "Metadata", "Numeric_features", and "Categorical_features", respectively. Ross *et al.* found that approximately 23% of eligibility criteria in ClinicalTrials.gov entries are numeric [14], such as for HbA1c, BMI, blood glucose, and creatinine. To unify these criteria, we negated the meaning of exclusion criteria and converted them to inclusion criteria without changing or losing meaning by replacing "<", "<=", ">", ">=" with ">=", ">", "<=", "<", respectively. The rationale of this conversion is that the complement of a numeric expression in an exclusion criterion indicates an inclusion criterion.

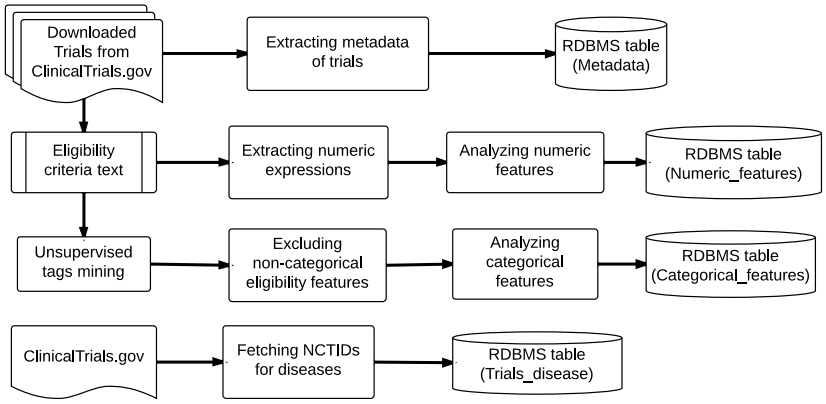


Figure 2. The pipeline for constructing the COMPACT database.

We applied unsupervised text mining on free-text eligibility criteria to automatically extract all frequent (i.e., appearing in at least 5% of all trials) eligibility features [9]. In this step, n-grams were generated for each criterion after noise reduction by a part-of speech (POS) tagger. The n-grams can be identified as eligibility features if they contain a substring that can be annotated using a UMLS concepts assigned one of 27 semantic types relevant to the clinical trial domain [15]. We excluded all the features whose UMLS semantic type was "Body Part, Organ, or

Organ Component". For example, in the inclusion criteria of trial NCT00934414, from the sentence "Smoking: BMI of = 40 kg/m2", two eligibility features "smoking" and "body mass index" were extracted and retained.

Valx, a numeric expression extraction and normalization tool [10, 11] developed in our lab, was employed to extract and parse complex numeric expressions in free-text eligibility criteria. For example, from the inclusion criterion "*type 2 diabetes mellitus diagnosed at least 3 months with fpg level <=240mg/dl and hba1c between 6.5% and 10% inclusive"*, Valx extracted two numeric features "glucose" ("fpg" is short for "fasting plasma glucose") and "HbA1c", and generated the expressions "[["Glucose", "<=", 240, "mg/dL"], [HbA1c, ">=", 6.5, "%"], [HbA1c, "<=", 10.0, "%"]]". For each numeric feature, the permissible value range, the width of the value range, and the boundary values were calculated to support common eligibility feature attributes mining. The value range of the numeric feature "HbA1c" in the criterion above is "[6.5, 10.0]" ("[ ]" means being inclusive, while "( )" means being non-inclusive). Its value range width is 10.0 – 6.5 = 3.5. Its boundary values are 6.5 and 10.0 since they both appear once in this criterion. Heterogeneous semantic representations for the same numeric feature were recognized. For example, HbA1c can be written as "hemoglobin A1c" or "A1c". Using synonyms in the UMLS and manually defined heuristics, we mapped these different representations to the same concept, HbA1c. Also, a rule-based algorithm was employed to recognize different representations for common comparison operators in eligibility criteria statements, such as ">" and "greater than", or ">=" and "greater or equal to". Various measurement units were harmonized. For example, blood glucose can be quantified by mg/dL or mmol/l.

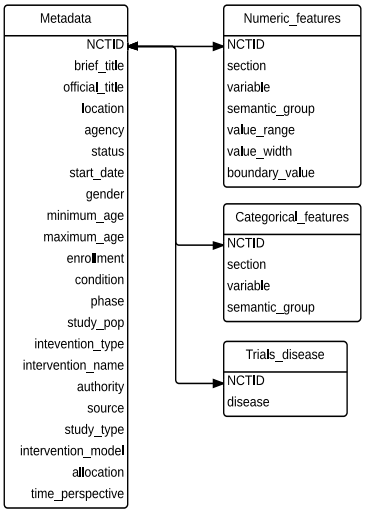


Figure 3. The database schema for COMPACT.

Figure 3 illustrates the database schema for COMPACT. For the table "trials_disease", we used all medical conditions defined by ClinicalTrials.gov (http://www.clinicaltrials.gov/ct2/search/ browse?brwse=cond_alpha_all) and created trial indexes for the selected conditions that were each studied by at least 50 clinical trials. The four tables in Figure 3 can be joined by a common attribute "NCTID", which is the unique identifier for a trial in the ClinicalTrials.gov. This knowledge base supports the mining of common eligibility features and their attributes in trials with various characteristics, e.g., a certain disease, a certain study design, recruiting patients of a certain gender, etc.

3. An Example Use of COMPACT

COMPACT enables flexible queries of sets of clinical trials for exploring the participant selection patterns in ClinicalTrials.gov. To illustrate such a use case, we developed an example common eligibility feature analytic module CONECT, leveraging RDBMS' capability of handling a wide range of queries to mine common eligibility features from COMPACT. **Figure 4** shows the workflow of CONECT using a concrete query example. The light blue rectangle boxes represent steps in the process. The ovals represent the output of each step.

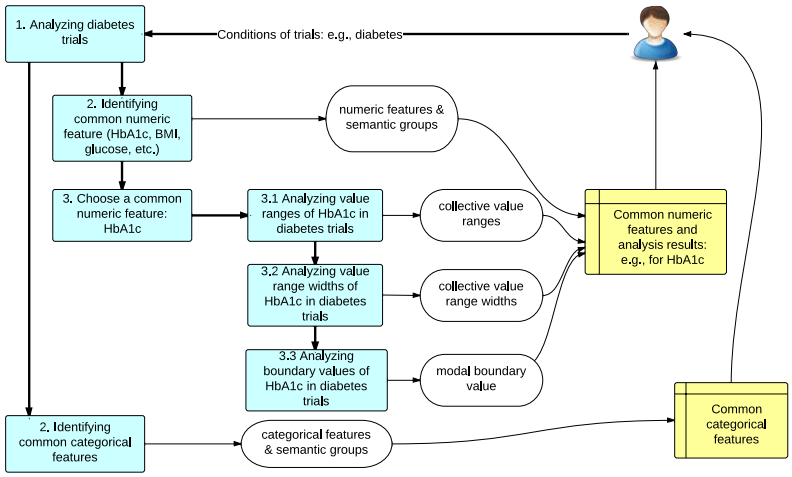


Figure 4. Workflow of CONECT with a concrete example of HbA1c patterns in diabetes trials.

In the example shown in **Figure 4**, a user may submit a query specifying the condition of trials to be "diabetes". In Step 1, CONECT will retrieve all the diabetes trials from COMPACT using the "trials_disease" table. In Step 2, CONECT will identify common numeric features (e.g., HbA1c, BMI, and glucose) and common categorical features in diabetes trials. In Step 3, the user may choose a numeric feature "HbA1c" for further analysis. Then all the pre-computed value ranges, value widths of the intervals, and boundary values of HbA1c in diabetes trials will be retrieved from COMPACT and aggregated upon the number of trials with the same value ranges, value widths, and boundary values. The output of the analysis including common numeric features, common categorical features, and the detail analysis of HbA1c in diabetes trials will be visualized and returned to the user.

Results

1. The Descriptive Statistics of COMPACT

To build the COMPACT database, we downloaded all 159,891 clinical trial records in XML format in ClinicalTrials.gov as of 01/27/2014. After excluding trials with no or non-informative eligibility criteria text, 159,187 trials were retained. The metadata fields of these trials were extracted and saved in the "Metadata" table. In order to quickly identify trials of a certain disease, we also created indexes for trials for 1,543 diseases with more than 50 trials in ClinicalTrials.gov. 229 out of these 1,543 diseases are listed as condition in more than 1,000. The trials of uncommon diseases can be retrieved by querying on the "condition" field using ClinicalTrials.gov's API. Valx extracted 1,045,893 numeric expressions for 176,986 unique features from 429,551 sentences in the inclusion criteria and 283,885 sentences in the exclusion criteria.

2. The Sample Query

In this section, we will use "Type 2 diabetes trials that recruit patients with HbA1C >= 7.0%" as a sample query to illustrate how common eligibility features are discovered from COMPACT. HbA1c is "a lab test that shows the average level of blood sugar (glucose) over the previous 3 months. It shows how well you are controlling your diabetes" [16]. In this paper, HbA1c refers to the test result. WebMD gives the normal ranges of HbA1c. For people without diabetes, the normal value range for HbA1c is between 4% and 5.6%. HbA1c levels between 5.7% - 6.4% indicate increased risk of diabetes, and 6.5% or higher indicate diabetes [17]. The American Diabetes Association recommends a HbA1c goal of less than 7.0% [18]. In COMPACT, 4,370 trials specify the value range of a patient's HbA1c in their eligibility criteria.

3. Common Eligibility Features and their UMLS Semantic Groups

In COMPACT, there are 4,493 Type 2 diabetes trials, out of which 700 trials (15.6%) are recruiting patients whose HbA1c must be >= 7.0%. According to the result of the analysis, five numeric features are used in the inclusion criteria of more than 50 such trials, whereas seven numeric features are used in the exclusion criteria of more than 50 such trials. Due to space limitations, **Table 1** shows the top five numeric features for each section, the semantic group, the number of trials, and the percentage of the qualifying trials (700 trials for the sample query) using the numeric feature. "HbA1c" is the only feature listed as one of the top five for both inclusion criteria and exclusion criteria of such trials, which conforms to the literature [18]. "HbA1c" and "Creatinine" are the most frequently used common numeric features in the inclusion and exclusion criteria of the qualifying trials, respectively.

Table 1. Top five numeric features frequently used in inclusion and exclusion criteria of Type 2 diabetes trials that recruit patients whose HbA1c >= 7.0%.

Numeric features used in the inclusion criteria				**Numeric features used in the exclusion criteria**			
Numeric features	**Semantic group**	**# Trials**	**Perc.**	**Numeric features**	**Semantic group**	**# Trials**	**Perc.**
HbA1c	Physiology	663	94.7%	Creatinine	Chemicals & Drugs	114	16.3%
BMI	Physiology	370	52.8%	Systolic blood pressure	Physiology	85	12.1%
Age	Physiology	327	46.7%	Diastolic blood pressure	Physiology	84	12%
Glucose	Chemicals & Drugs	114	15.9%	ALT	Chemicals & Drugs	73	10.4%
C-peptide	Chemicals & Drugs	56	8.0%	HbA1c	Physiology	71	10.1%

For common categorical features, nine are used in inclusion criteria of more than 50 trials, whereas 46 are used in exclusion criteria of more than 50 trials. **Table 2** shows the top five categorical features for each section, their sematic group, their use in inclusion or exclusion section, and their frequency. "Diabetes mellitus non-insulin-dependent" (Type 2 diabetes) and "diabetes mellitus insulin-dependent" (Type 1 diabetes) are the most frequently used categorical features in the inclusion and exclusion criteria, respectively. This is reasonable because usually Type 2 diabetes trials exclude Type 1 diabetes patients.

Table 2. Top five categorical features frequently used in inclusion and exclusion criteria of Type 2 diabetes trials that recruit patients whose HbA1c >= 7.0%. The names of categorical features are UMLS terms.

Categorical features used in the inclusion criteria				**Categorical features used in the exclusion criteria**			
Categorical features	**Semantic group**	**# Trials**	**Perc.**	**Categorical features**	**Semantic group**	**# Trials**	**Perc.**
diabetes mellitus non-insulin-dependent	Disorders	520	74.3%	diabetes mellitus insulin-dependent	Disorders	236	33.7%
sulfonylurea compounds	Chemicals & Drugs	118	16.9%	pharmacologic substance	Chemicals & Drugs	229	32.7%
antidiabetics	Chemicals & Drugs	94	13.4%	allergy severity - severe	Disorders	224	32.0%
pharmacologic substance	Chemicals & Drugs	91	13.0%	gravidity	Disorders	223	31.9%
contraceptive methods	Procedures	83	11.9%	malignant neoplasm	Disorders	190	27.1%

4. Collective Value Ranges of Numeric Eligibility Features

Among the top 5 numeric features discovered for the sample query, we chose the two most frequently used numeric features "HbA1c" and "BMI" to illustrate collective value ranges, collective value range widths, and modal boundary values. **Table 3** shows the five most frequently used value ranges of HbA1c and BMI in Type 2 diabetes trials. "[ ]" means being inclusive, while "( )" means being non-inclusive; -∞ refers to negative infinity, whereas +∞ refers to positive infinity. Out of 4,493 Type 2 diabetes trials, 2,058 trials (45.8%) use HbA1c, whereas 1,859 trials (41.4%) use BMI in their eligibility criteria. According to the analysis (after unifying inclusion and exclusion criteria), the most frequently used permissible value range for HbA1c is [7.0, 10.0], while the most frequently used value range for BMI is (-∞, 45.0].

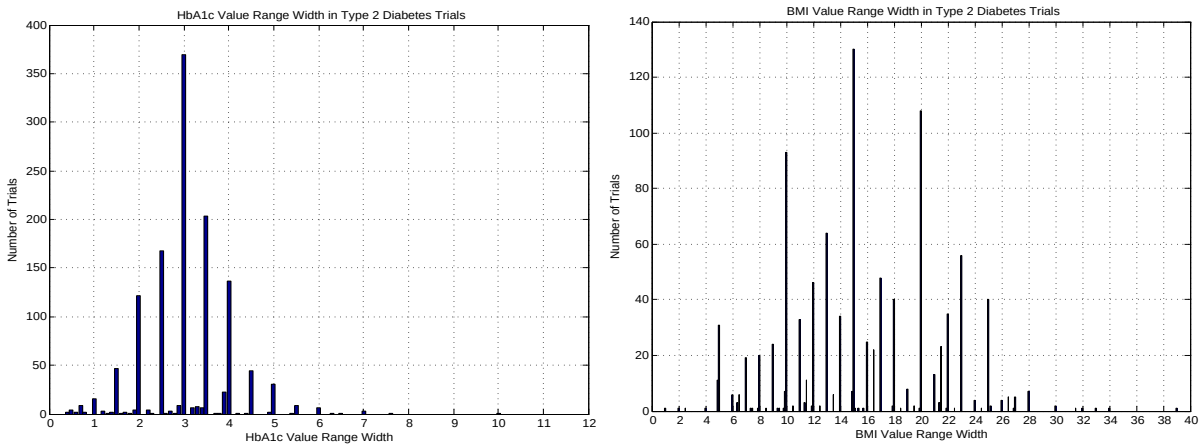
Table 3. Top five collective value ranges of HbA1c and BMI in Type 2 diabetes trials.

HbA1c value ranges	**Number of trials**	**BMI value ranges**	**Number of trials**
[7.0, 10.0]	228	(-∞, 45.0]	113
(7.0, +∞)	97	(-∞, 40.0]	104
(-∞, 7.0]	88	[25.0, 40.0]	72
[7.0, 11.0]	75	(-∞, 40.0)	61
[7.0, +∞)	57	(-∞, 35.0]	58

5. Collective Value Range Widths of Numeric Eligibility Features

In this study, when analyzing the collective value range widths of numeric features, we excluded all the numeric features without an upper bound or a lower bound and defer their analysis to future work.

Out of 2058 Type 2 diabetes trials with HbA1c in their eligibility criteria, 1156 trials (56.2%) specify a bounded value range for HbA1c, i.e., with both an upper bound and a lower bound. **Figure 5 (a)** shows the distribution of collective value range widths of HbA1c in Type 2 diabetes trials. We can see that the value range width "3" is used by 370 Type 2 diabetes trials for HbA1c. Out of 1859 Type 2 diabetes trials using BMI as a numeric feature, 981 trials (52.8%) specify a bounded value range for BMI. **Figure 5 (b)** shows that "15" is the most frequent value range width of BMI used by 130 Type 2 diabetes trials.

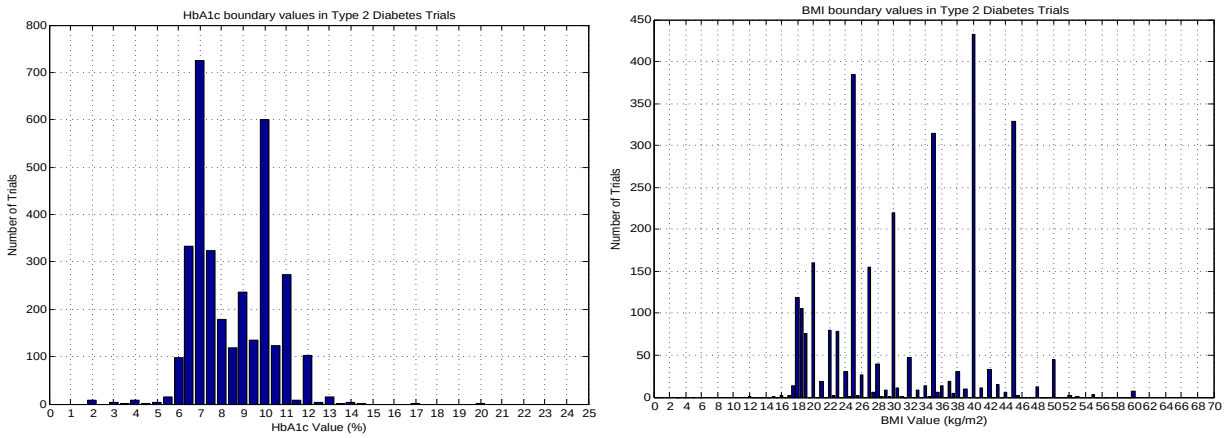


(a) (b)

Figure 5. Collective value range widths for (a) HbA1c and (b) for BMI in Type 2 diabetes trials.

6. Modal Boundary Values of Numeric Eligibility Features

For analyzing the modal boundary value of a numeric feature in trials of a certain disease, we aggregated the number of trials using a certain value for eligibility determination. **Figure 6 (a)** below shows the distribution of Type 2 diabetes trials using a specific HbA1c value in the eligibility criteria, i.e., below or above this value. The peak in **Figure 6 (a)** corresponds to the modal boundary value of HbA1c, which is 7.0% for Type 2 diabetes trials, whereas this modal boundary value of BMI is 40 kg/m^2, as shown in **Figure 6 (b)**.



(a) (b)

Figure 6. Collective boundary values for (a) HbA1c and (b) BMI, respectively, for Type 2 diabetes trials.

7. The Interface Design of CONECT (Commonalities in Target Populations in Eligibility Criteria)

We designed an interface for CONECT so that clinical investigators will be able to utilize the COMPACT database to inform the design of new trials. **Figure 7** shows the interface of the CONECT prototype. In this interface, an investigator can specify various characteristics of trials, e.g., disease, study type, study design (including intervention model, allocation, and time perspective), status, intervention type, phase, permissible values for specific numeric eligibility criteria features, and the number of common eligibility features to be retrieved. Based on the query of the investigator, CONECT will retrieve common eligibility features grouped by their semantic groups. When an investigator clicks on a certain numeric feature, its attributes, i.e., collective value ranges, collective value widths, and collective boundary values with a modal value if applicable, will be visualized in the lower space. CONECT informs the investigator what threshold other investigators have defined for the same variable in similar contexts. The analysis for contextual attributes for the categorical features is still under construction.

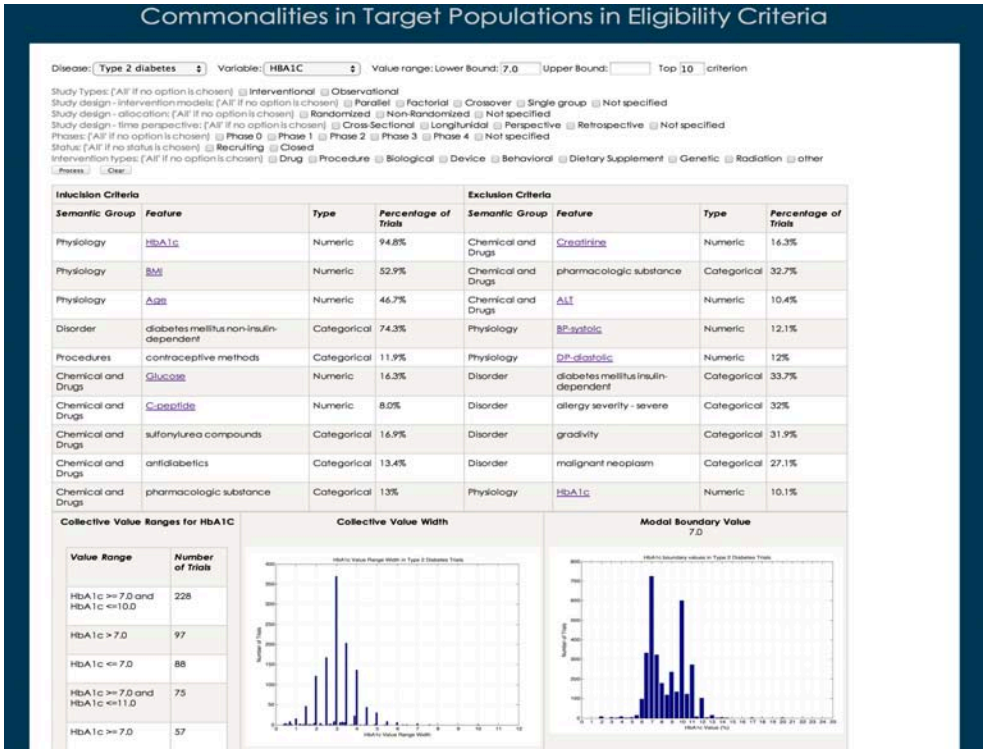


Figure 7. The interface of CONECT (Commonalities in Target Populations in Eligibility Criteria) prototype.

Discussion

This paper introduces a method for analyzing commonalities in eligibility criteria and a new data-driven approach to supporting actionable knowledge reuse for clinical trial eligibility criteria design. The COMPACT database built using this method should enable its target users to learn from the distributed clinical research community. Since its launch in 1997, ClinicalTrials.gov has accumulated about 160,000 studies, among which 48.6% are completed, and 27.3% are enrolling participants. With the large amount of information about clinical trial designs available, there is a compelling need to examine the commonalities in patient selection in this repository to inform the development of new clinical trials. Previously, there was no method going to this level of detail to parse the attributes of frequent eligibility features, e.g., their permissible value ranges, modal boundary values, etc. By coupling this information with descriptive characteristics of clinical trials, our method presents a new way for clinical trial sponsors and investigators to analyze fine-grained commonalities in clinical trial target populations and understand design patterns of eligibility criteria. Leveraging existing clinical trial records, our method intends to discover popular content used in the research community to facilitate future clinical trial design.

Utility of the COMPACT Database

The most important two design principles of the COMPACT database are generalizability and extensibility. In this paper, we demonstrate the utility of the COMPACT database with an example analytic module that supports disease-specific trial design. Analytic modules can be developed for other purposes. COMPACT has additional data such as drug name and location that have not been used by CONECT. Thus, one may develop a module to find common eligibility features in trials of a certain drug. COMPACT enables flexible selection of clinical trials of various characteristics, e.g., randomized or non-randomized, starting after a certain date, with a certain eligibility feature, recruiting only male patients, etc. For example, a clinical trial designer may be interested in discovering common eligibility features for breast cancer trials starting after January 2009, or for heart disease trials that recruit patients with systolic blood pressure greater than 140 mmHg, etc. The normalization of different units of the same numeric feature enables their aggregate analysis among different clinical trials.

Comparison with Other Databases Derived from ClinicalTrials.gov

The most notable effort related to our method is a database called Aggregate Analysis of Clinical Trials (AACT). It was developed by Tasneem *et al.* using descriptive characteristics, e.g., conditions and intervention, to aggregate

clinical trials from ClinicalTrials.gov [19]. Our method and AACT both use RDBMS to store clinical trial summaries from ClinicalTrials.gov. However, AACT does not analyze eligibility criteria in detail. In contrast, our method derives commonalities in participant selection from parsed eligibility criteria. Another database LinkedCT transformed the data of clinical trials on ClinicalTrials.gov into RDF to discover semantic links between clinical trial records [20]. However, it does not contain discrete eligibility features and hence does not support the analysis of eligibility criteria as COMPACT does. Thus, COMPACT can potentially make important contributions by facilitating knowledge reuse during eligibility criteria designs for new clinical trials.

Difference between Common Eligibility Features and Common Data Elements

Luo *et al.* used unsupervised machine learning to identify disease-specific Common Data Elements (CDEs) from clinical trial eligibility criteria [7]. Common eligibility features and CDEs can be both disease specific, but are different in the following three aspects: (1) CDEs are generated in a semi-automated fashion whereas common eligibility features generation is interactive and contextual, depending on the query; (2) Common eligibility features can be generated for trials of various characteristics on the fly, e.g., trials in a certain phase, eligibility criteria with a certain numeric feature in a certain range, etc., while CDEs are usually identified for a specific disease domain, and (3) Common eligibility features have contextual attributes. Numeric features have disease-specific value ranges, disease-specific value range widths and disease-specific modal boundary values. Investigators will get contextual knowledge on how these numeric features are used to define eligibility criteria for participant selection. This method overcomes the following limitations in traditional knowledge base development, including (1) centered around one or a small group of domain experts, (2) limited to one or few disease domains, (3) laborious and costly knowledge management processes, and (4) lack of scalability.

Limitations

This study has limitations. The accuracy of the commonalities in the target populations depends on the accuracy of the parsing of the eligibility criteria text. In a preliminary evaluation of the parser Valx, the precision, recall, and F-measure for extracting numeric expressions with the feature "HbA1c" were 99.6%, 98.1%, 98.8% for Type 1 diabetes trials, and 98.8%, 96.9%, 97.8% for Type 2 diabetes trials, respectively. The results of the measures for extracting numeric expressions with the feature "Glucose" were 97.3%, 94.8%, 96.1% for Type 1 diabetes trials, and 92.3%, 92.3%, 92.3% for Type 2 diabetes trials, respectively [10]. In the future, a comprehensive evaluation of Valx is necessary to assess its performance for other numeric features for their uses in clinical trials of various medical conditions. Our text mining method serves our goals presented here, but still has significant room for improvement. It currently does neither exhaustively recognize all the abbreviations nor normalize all the measurement units for every numeric feature in eligibility criteria text. Categorical features without contextual information (e.g., *within 12 hours of the onset of chest pain*) are sometimes not specific enough for direct reuse. Therefore, further collaborative research on natural language processing of free-text eligibility criteria is desired. In this study, we used one sample query on a common disease (i.e., Type 2 diabetes) for demonstration purposes. More studies are warranted to test how this method works for other eligibility features and other diseases.

Future Work

The COMPACT database needs to be updated on a regular basis. We will develop CONECT as a Web-based system and formally assess its value for clinical research stakeholders. To provide acceptable user experience when interacting with the system, it is imperative to improve the performance of our query processing method, which requires costly table join and value aggregation for numeric features. The current long waiting time, i.e., two minutes, to retrieve common eligibility features for some queries would impair the user experience of the system. We will create a repository to store common eligibility features for popular queries. When a popular question is asked again, the saved results can be returned to the user promptly. We will also analyze the queries submitted by the users, which may reveal new trends in trial design. With continuously improving natural language processing techniques, we will provide finer-grained common eligibility features. Temporal usage patterns will also be explored to show how commonalities in participants evolve over time and across disease domains. In the future, to enhance the analytic utility of COMPACT, we will add a table containing real world patient data and perform comparative analysis between clinical trial target populations and real world patient populations [21].

Conclusions

We introduced a novel data resource for analyzing common numeric and categorical features in eligibility criteria from public clinical trials records. We designed and built a database called COMPACT using all the clinical trial records on ClinicalTrials.gov. CONECT was introduced to illustrate example visualization of the COMPACT

content and the interactions between users and COMPACT. The query “Type 2 diabetes trials that recruit patients with HbA1c >= 7.0%” was used to illustrate the method. This research can potentially help clinical investigators understand frequently used eligibility criteria and how they have been shared across studies and can inform the design for participant selection for new clinical trials. Our future work includes further improvement of visual queries using COMPACT and user evaluation of this new decision aid for clinical research stakeholders.

Acknowledgments

We thank Dr. Riccardo Miotto for contributing parsed clinical trial summaries for building the COMPACT database. This study was sponsored by the National Library of Medicine grant **R01LM009886** (PI: Weng) and National Center for Advancing Translational Science grant **UL1 TR000040** (PI: Ginsberg).

References

1. Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?". Lancet. 2005;365(9453):82-93.
2. Fuks A, Weijer C, Freedman B, Shapiro S, Skrutkowska M, Riaz A. A study in contrasts: eligibility criteria in a twenty-year sample of NSABP and POG clinical trials. National Surgical Adjuvant Breast and Bowel Program. Pediatric Oncology Group. J Clin Epidemiol. 1998;51(2):69-79.
3. Van Spall HG, Toren A, Kiss A, Fowler RA. Eligibility criteria of randomized controlled trials published in high-impact general medical journals: a systematic sampling review. JAMA. 2007;297(11):1233-40.
4. Hao T, Rusanov A, Boland MR, Weng C. Clustering clinical trials with similar eligibility criteria features. J Biomed Inform. 2014;pii: S1532-0464(14)00011-2.
5. Kim EK, Seok JH, Oh JS, Lee HW, Kim KH. Use of hangeul twitter to track and predict human influenza infection. PLoS One. 2013;8(7):e69305.
6. ClinicalTrials.gov [February 2014]. Available from: http://www.clinicaltrials.gov/.
7. Luo Z, Miotto R, Weng C. A human-computer collaborative approach to identifying common data elements in clinical trial eligibility criteria. J Biomed Inform. 2013;46(1):33-9.
8. Boland MR, Miotto R, Weng C. A method for probing disease relatedness using common clinical eligibility criteria. Stud Health Technol Inform. 2013;192:481-5.
9. Miotto R, Weng C. Unsupervised mining of frequent tags for clinical eligibility text indexing. J Biomed Inform. 2013;46(6):1145-51.
10. Hao T, Weng C. Valx: A Knowledge-based System for Extracting and Structuring Numeric Comparison Statements in Clinical Research Eligibility Criteria Text. J Biomed Inform.Under review.
11. Hao T, Weng C. Valx - Numeric Expression Extraction and Normalization Tool 2013 [February 2014]. Available from: http://columbiaelixr.appspot.com/valx.
12. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004;32(Database issue):D267-70.
13. The UMLS Semantic Group [February 2014]. Available from: http://semanticnetwork.nlm.nih.gov/SemGroups/.
14. Ross J, Tu S, Carini S, Sim I. Analysis of eligibility criteria complexity in clinical trials. AMIA Summits Transl Sci Proc. 2010;2010:46-50.
15. Luo Z, Johnson SB, Weng C. Semi-Automatically Inducing Semantic Classes of Clinical Research Eligibility Criteria Using UMLS and Hierarchical Clustering. AMIA Annu Symp Proc. 2010;2010:487-91.
16. MedlinePlus [February 2014]. Available from: http://www.nlm.nih.gov/medlineplus/.
17. HbA1c value range [February 2014]. Available from: http://www.webmd.com/diabetes/guide/glycated-hemoglobin-test-hba1c.
18. American Diabetes Association Website [February 2014]. Available from: http://www.diabetes.org/.
19. Tasneem A, Aberle L, Ananth H, Chakraborty S, Chiswell K, McCourt BJ, Pietrobon R. The Database for Aggregate Analysis of ClinicalTrials.gov (AACT) and Subsequent Regrouping by Clinical Specialty. PLoS ONE. 2012;7(3):e33677.
20. Hassanzadeh O. LinkedCT [March 2014]. Available from: http://linkedct.org.
21. Weng C, Li Y, Ryan P, Zhang Y, Gao J, Liu F, Bigger JT, Hripcsak G. A Distribution-based Method for Assessing The Differences between Clinical Trial Target Populations and Patient Populations in Electronic Health Records. Applied Clinical Informatics. 2014;5(2):463-79.

Predicting Discharge Mortality after Acute Ischemic Stroke Using Balanced Data

King Chung Ho, MS[1,2], William Speier, MS[1,2], Suzie El-Saden, MD[2], David S. Liebeskind, MD[3], Jeffery L. Saver, MD[3], Alex A. T. Bui, PhD[1,2], Corey W. Arnold, PhD[1,2]

[1]Department of Bioengineering; [2]Medical Imaging Informatics, Department of Radiological Sciences; [3]UCLA Stroke Center, Department of Neurology University of California, Los Angeles, CA

Abstract

Several models have been developed to predict stroke outcomes (e.g., stroke mortality, patient dependence, etc.) in recent decades. However, there is little discussion regarding the problem of between-class imbalance in stroke datasets, which leads to prediction bias and decreased performance. In this paper, we demonstrate the use of the Synthetic Minority Over-sampling Technique to overcome such problems. We also compare state of the art machine learning methods and construct a six-variable support vector machine (SVM) model to predict stroke mortality at discharge. Finally, we discuss how the identification of a reduced feature set allowed us to identify additional cases in our research database for validation testing. Our classifier achieved a c-statistic of 0.865 on the cross-validated dataset, demonstrating good classification performance using a reduced set of variables.

1. Introduction

Stroke is a major cause of mortality and disability in the United States, with over 20% of affected individuals succumbing to the condition, and only one-fourth of surviving adults returning to normal health status [1-3]. There exist a variety of treatments for stroke, including intravenous and intra-arterial tissue plasminogen activator (tPA), as well as clot retrieval using mechanical devices (i.e. mechanical thrombectomy). However, the relationship among a patient's stroke presentation, treatment options, and functional outcomes is not well defined. To inform treatment, work in the literature has focused on building models to predict a patient's functional status given initial presentation, in hopes of detecting cases that need to be treated more or less aggressively.

To build a predictive model that may be used in practice and which generalizes to multiple institutions, it is important to determine an informative feature set that is collected in the course of normal clinical care. More importantly, the distribution of data and associated class labels is also important. Recently, the problem of imbalanced data has received increased attention as it has been shown to decrease model performance and lead to biased prediction [4-6]. While previous work has been focused on using logistic regression models to predict stroke outcomes [7-10], alternative machine learning techniques are less studied, but potentially useful.

In this work, we present a comparison between six models for predicting stroke mortality at discharge that includes methods for data balancing. We calculated the c-statistic for each classifier, following a systematic approach to identify the optimal set for each classifier by sequentially adding informative features until performance no longer increased.

2. Previous Work

There are many models for predicting acute stroke outcomes, which in general apply multivariate logistic regression techniques. Counsell *et.al.* [7] developed a six simple variable (SSV) multivariate logistic regression model to predict survival rate 30 days and 6 months after stroke using a 530 patient dataset. The six variables used were age, living alone, independence in activities of daily living before the stroke, the verbal component of Glasgow Coma Scale, arm power, and the ability to walk. The model was validated using two external independent cohorts of stroke patients with c-statistics from 0.84 to 0.88. Teale *et. al.* [8] reviewed different literature databases (e.g., MEDLINE) and concluded that the SSV model demonstrated statistical robustness, good discriminatory function in external validation, and comprised clinically feasible variables that were easily collected; it performed well among all prognostic models for predicting outcome after acute stroke.

Konig *et. al.* [9, 10] developed a multivariate logistic regression model to predict three month survival rate after acute ischemic stroke using age and the National Institutes of Health Stroke Scale (Pre-NIHSS) [11]. They showed

that Pre-NIHSS is a strong predictor variable that has significant impact on model accuracy. Further analyses [12, 13] have supported the effectiveness of Pre-NIHSS in 30-Day mortality prediction after acute ischemic stroke.

Saposnik *et. al.* [14-16] considered the utility of a predictive model in a clinical setting and developed an algorithm to predict mortality risk in ischemic stroke patients after hospitalization based on information routinely available in hospital, such as demographics, clinical presentation (diplopia, dysarthria, and aphasia, etc.), and patient comorbidities. The model, called iScore, was developed from multivariate logistic regression models by using a regression coefficient-based scoring method based on β coefficients. The iScore model is now available online [17].

In contrast to more general models, which attempt to predict an outcome based on a patient's presentation, additional algorithms are under development that attempt to predict response to a specific therapy. Examples include the DRAGON score for intravenous thrombolysis therapy [18], and HIAT2 score for outcomes after intra-arterial thrombolysis [19].

While significant work has been done in predicting outcomes after acute stroke, there is little work regarding the problem of between-class imbalance, which is common in binary prediction tasks when one class (the majority) is more common than the other class (the minority). We address this problem by using a sampling method, called the Synthetic Minority Over-sampling Technique (SMOTE) [20]. We compare the performance of various models (logistic regression, support vector machines, naïve bayes, decision tree, and random forest) for balanced and imbalanced datasets and determine the best model for predicting discharge mortality. We then identify the best six features that appear to be critical for good model performance.

3. Methods

This section describes the stroke dataset used, the imbalanced dataset problem, and the modeling methods compared for mortality prediction.

3.1 Dataset

UCLA maintains a REDCap [21] database that stores acute stroke patients who have been treated with one or more of the following treatments from 1992 to 2013: intra-arterial tissue plasminogen activator (IA tPA), intravenous tissue plasminogen activator (IV tPA), or mechanical thrombectomy. There are 778 patients in this study cohort, each with more than 500 features, including demographic information, laboratory results, and medications. Our goal is to predict patient mortality at hospital discharge, which is indicated by the discharge modified Rankin Scale (discharge mRS) [22, 23]. Discharge mRS is a commonly used scale for measuring the degree of disability or dependence in the daily activities of patients who have suffered stroke. The scale runs from 0-6, with 0 indicating perfect health and 6 indicating death. Our binary prediction model collapses this scale to two groups: alive (0-5) and dead (6).

Using this dataset, we defined a cohort subset for analysis in this study using inclusion criteria (Table 1). The features are summarized in Table 2. The continuous features are patients' demographics information and time-related information while the binary features are patients' presentation, and medication. This subset was used to build the model.

Table 1. Inclusion criteria for cohort subset.

Original dataset	778 patients; >500 features
Patients' inclusion criteria	• Only patients with discharge mRS recorded. • Only patients with ischemic stroke (excluding patients with subarachnoid hemorrhage). • Only patients who received treatment solely at UCLA. • Only patients who with hospital stays less than 20 days (patients who stay longer are more likely to have other conditions in addition to stroke). • Only patients without missing any features' values after using features inclusion criteria
Features inclusion criteria	• Only features that were available in over 90% of patient cases after the first four patients' inclusion criteria.
Cohort subset	190 patients (156 alive, 34 dead); 26 features

Table 2. Feature distribution of alive and dead patients in dataset before and after SMOTE.

		Alive	Dead (Before SMOTE)	Dead (After SMOTE)
Size		156	34	156
Continuous features		Average(SD)	Average(SD)	Average(SD)
	Age	67.91(17.05)	81.29(9.45)	81.41(8.06)
	Pre-NIHSS	12.16(7.20)	18.24(5.93)	18.16(4.84)
	Systolic blood pressure	151.03(28.20)	147.32(30.96)	145.54(27.54)
	Diastolic blood pressure	82.42(17.71)	79.12(19.12)	77.88(16.30)
	Blood glucose	131.72(48.14)	164.82(74.66)	160.81(61.98)
	Blood platelet count	224.99(75.63)	181.47(54.86)	179.43(50.50)
	Hematocrit	39.60(5.33)	38.66(5.37)	38.38(4.77)
	Time difference between first MRI image and time of symptoms (minute)	178.77(152.89)	166.15(109.27)	171.39(96.26)
	Time difference between first MRI image and admission (minute)	44.67(27.66)	49.35(34.73)	48.89(30.66)
Binary features		Percentage of true(%)	Percentage of true(%)	Percentage of true(%)
	Gender (Male)	44.87	52.94	37.82
	Hypertension	66.67	82.35	84.62
	Diabetes	16.67	20.59	6.41
	Hyperlipidemia	27.56	35.29	17.95
	Atrial fib	30.13	55.88	41.67
	Myocardial infarction	12.18	38.24	25.00
	Coronary artery bypass surgery	7.69	11.76	2.56
	Congestive heart failure	3.21	26.47	8.33
	Peripheral vascular disease	0.00	2.94	0.64
	Carotid endarterectomy angioplasty/stent	1.92	5.88	1.28
	Brain aneurysm	0.00	2.94	0.64
	Active internal bleeding	0.00	0.00	0.00
	Low platelet count	0.00	0.00	0.00
	Abnormal glucose	0.00	0.00	0.00
	Diabetes medication	14.74	32.35	10.90
	Hypertension medication	53.21	82.35	86.54
	Hyperlipidemia medication	16.67	23.53	7.69

3.2 Imbalanced Learning Problem

The cohort subset is imbalanced (156 alive vs. 34 dead). Previous research has shown that imbalanced learning can cause prediction bias, leading to inaccurate and unreliable results [4, 6]. In our case, models would predict all patients as alive to achieve high accuracy, but with low precision and recall. The imbalanced learning problem has received attention in both theoretical and practical application [5]. Most machine learning algorithms do not deal with imbalanced datasets during the training process [4]. In some cases, the minority are scattered in the feature space and the decision boundary is too specific. We addressed this issue by using Synthetic Minority Over-sampling

Technique (SMOTE) [20], a sampling technique combining under-sampling of the majority class with over-sampling of the minority class.

There are two parts to the SMOTE algorithm. In the first part, the minority class is over-sampled by taking each minority class sample and introducing new synthetic samples joining any or all of the *k* minority class nearest neighbors (by Euclidean distance). Neighbors from the *k* nearest neighbors are randomly chosen depending upon the amount of over-sampling required. Synthetic samples are generated in the following way:

1. Find the difference between the normalized feature vector (sample), F_{orig}, and a randomly selected normalized nearest neighbor, F_{near}.
2. Multiply the difference by a random number between 0 and 1.
3. Add the product to the feature vector to generate a new feature vector.

This approach essentially creates a random point along the line segment between two specific features and effectively forces the decision region of the minority class to become more general. The new feature vector, F_{new}, is defined as follows:

$$F_{new} = F_{orig} + rand(0,1) * (F_{near} - F_{orig}) \quad (1)$$

The above step is used for continuous features. For binary features, the new value is obtained by the majority vote (0 or 1) of all the neighbors. In our training process, the second part of the SMOTE algorithm, under-sampling, was not performed because the dataset was small and all data should be considered. We applied the SMOTE algorithm using MATLAB, denoting this new dataset as SMOTE-dataset. Five neighbors (*k*) were used and one of them was randomly selected to generate a synthetic sample. The synthetic step was repeated until the number of minority is equivalent to the number of majority.

3.3 Model comparison and feature selection

Early research has emphasized using logistic regression models [24] to predict stroke outcome [7-10]. To the best of our knowledge, none have compared the performance of different machine learning methods in stroke outcome prediction, particularly with a balanced dataset. Therefore, we compared the performance of five common machine learning methods: Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), Random Forests (RF), and Logistic Regression (LR). In addition, we compared the performance of a combined method: principal component analysis followed by support vector machine (PCA+SVM).

Briefly, NB is a probabilistic classifier algorithm based on Bayes' Rule that makes a conditional independence assumption between the predictor variables, given the outcome [25]. SVM is a supervised learning classification algorithm that constructs a hyperplane (or set of hyperplanes) in a higher dimensional space for classification [26]. DT is a tree-like prediction model in which each internal (non-leaf) node tests an input feature and each leaf is labeled with a class [27]. RF is an ensemble learning method in which a multitude of decision trees are constructed and the classification is based on the mode of the classes output by individual trees [28]. LR is a probabilistic classification model in which label probabilities are found by fitting a logistic function of feature values [29]. Principal component analysis is a statistical procedure that uses orthogonal transformation to reduce the dimensionality of the feature space containing only principal components (principal features) [30]. Yang *et. al.* [31] and Gumus *et. al.* [32] have shown the effectiveness of using PCA to extract principal components for SVM classification. We also used this combination of methods and determine whether PCA is beneficial before SVM on stroke patient mortality classification.

It is time-consuming to collect a comprehensive set of features for every patient, and not all the features are relevant. In addition, using all the features to construct a classifier may lead to decreased performance due to over-fitting, especially on small, imbalanced datasets, which are not uncommon in stroke. Thus, we sought a minimum feature set to mitigate these clinical and modeling challenges, and that could also be externally validated more easily.

Different machine learning methods may not perform equally on the same feature set. Therefore, optimal feature sets for each machine learning method were defined systematically, with the top performing method determined using the c-statistic. First, chi-square tests were used to weight the association between features (categorical variables/binned continuous variables) and discharge mRS in the original dataset. Then, the feature with the highest weight was used to construct a single-variable classifier and the c-statistic was calculated. The feature with the second highest weight was then added, and c-statistic was re-calculated. This process of adding features was repeated until an optimal feature set for a machine learning method was obtained. The optimal feature set (the first

optimum) was defined as the point at which adding any additional features did not increase the performance of the classifier. All methods were compared and the best one was chosen for stroke patient mortality at discharge. A summary of steps is shown in Figure 1. Models were fitted and compared using RapidMiner (RM) [33], an open-source software platform which provides an integrated environment for machine learning, data mining, and predictive analytics [34, 35].

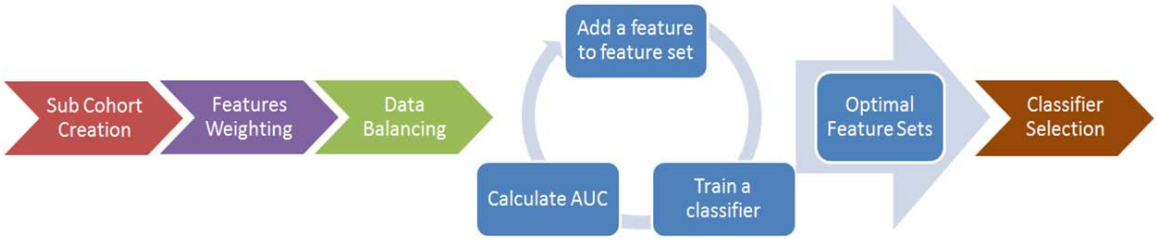


Figure 1. Summary of building classifier for predicting stroke patient mortality at discharge. A sub-cohort based on clinical and missing data factors was first created. Then, the relevance between features and the classes were weighted by chi-square statistics, and then the dataset is balanced by SMOTE. The fourth step was an iterative process in which the highest weighted feature was first used to build classifiers and AUC was calculated. Features were added to the training feature set sequentially in the order of weighting and classifiers were trained. After several iterations, optimal feature sets for all classifiers were obtained. Finally, performances were compared (SVM, PCA&SVM, DT, RF, NB, and LR) and the best classifier was selected.

4. Result and Discussion

4.1 SMOTE over-sampling

There were 190 patients in the training cohort with 156 class-1 patients (alive) and 34 class-2 patients (dead). The final SMOTE-dataset had 156 class-1 patients and 156 class-2 patients. Age and Pre-NIHSS have been shown to be two important features for stroke outcome prediction [9, 10, 12, 13]. Figure 2A shows the data distribution (x-axis: Pre-NIHSS; y-axis: age) between two the classes before SMOTE sampling and Figure 2B showed the distribution after SMOTE sampling. The original class-2 distribution was scattered, and it was hard to determine where the decision boundary should be. After oversampling by SMOTE, it was easy to observe that class-2 patients clustered in the region of high age and high Pre-NIHSS.

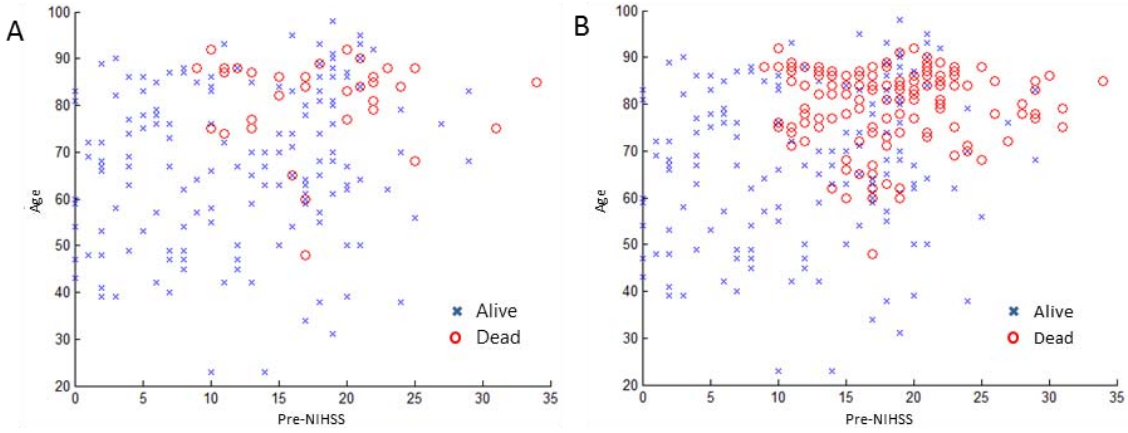


Figure 2. Two classes distribution before and after SMOTE sampling. (A) Before SMOTE, (B) After SMOTE. Two datasets for classifications: (A) original dataset with 156 class-1 patients and 34 class-2 patients), (B) SMOTE-dataset with 156 class-1 patients and 156 class-2 patients.

4.2 Feature Selection

Chi-square tests were used to weight the relevance between a feature and the classes in the original dataset. As our goal was to obtain a ranking of features (rather than finding statically significant features), we calculated the chi-squared statistics (weight) and normalized them for comparison. Continuous variables were discretized into ten bins. The top ten normalized weights are shown in Table 3.

Table 3. Top ten normalized weights by chi-squared statistic.

Feature	Weight
Pre-NIHSS	1.000
Age	0.631
Patient history of congestive heart failure	0.565
Platelet count	0.450
Patient history of myocardial Infarction	0.346
Serum glucose	0.325
Patient on hypertension medication	0.249
Time difference between the first MRI image and admission	0.243
Systolic blood pressure	0.223
Patient history of atrial fibrillation	0.209

4.3 Model Comparison

Ten-fold cross-validation was used to compare different models on the datasets, measuring the model performance via c-statistic and model bias via F1-score [36]. In each validation, nine groups of original data were used to train the classifier and one group of data was classified. For the SMOTE-dataset, nine folds of data were balanced using SMOTE and then used to train the classifier. The held-out unbalanced dataset was then classified.

Features were added sequentially to train a classifier in order of largest weight to smallest weight. The c-statistic of each classifier is shown in Figure 3, with its distribution shown in Figure 4. The enlarged markers represent the size of the optimal feature set for each classifier. With only one feature (Pre-NIHSS), all classifiers performed poorly. The performance gradually increased as more features were added. Each classifier reached to the first maximum c-statistic at a different number of features, with performance for each leveling off with additional features. This validated our hypothesis that more features might not necessarily improve the performance. The size of each optimal feature set is summarized in Table 4.

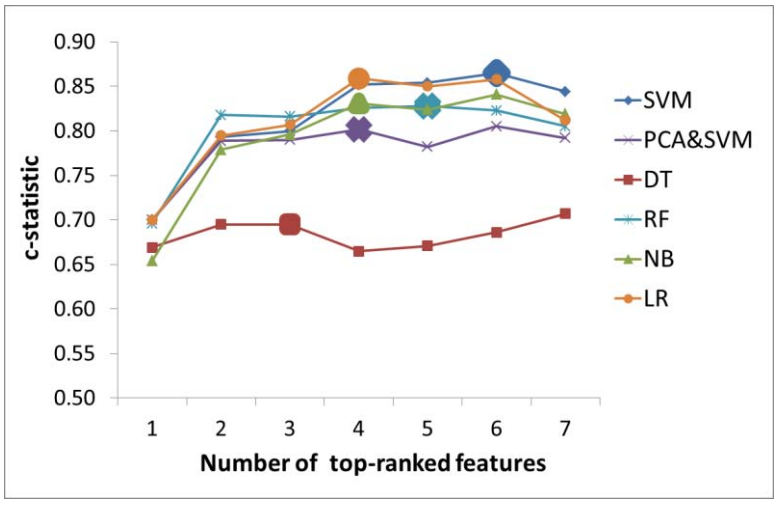


Figure 3. The c-statistics of classifiers with different number of top-ranked features based on Table 3. The size of the optimal feature set for each classifier was indicated by the enlarged marker. Each classifier has different size of optimal feature set. Among all classifiers, SVM has the highest c-statistic with 6 features being used.

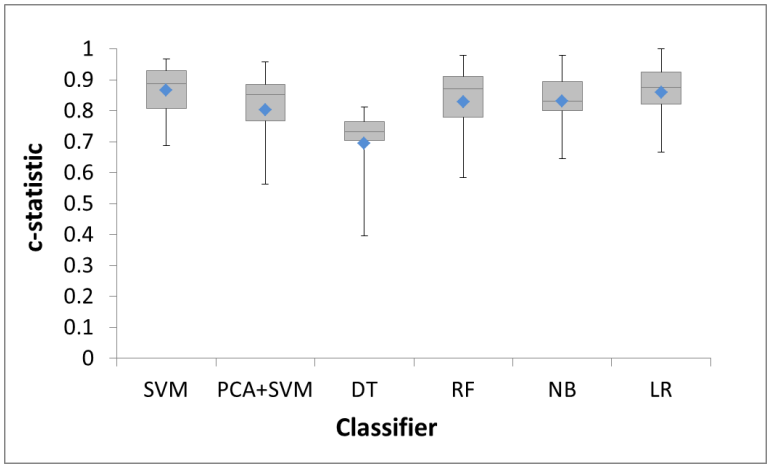


Figure 4. c-statistic distributions for classifiers with optimal feature set (means are represented by blue diamonds).

Table 4. The size of the optimal feature set for each classifier and the cross-validation result.

		SMOTE-dataset		Original dataset	
Classifier	**Optimal feature set size**	**c-statistic**	**F1-score**	**c-statistic**	**F1-score**
SVM	6	0.865	0.594	0.868	0.336
PCA&SVM	4	0.802	0.488	0.856	0.368
DT	2	0.695	0.441	0.709	0.061
RF	5	0.828	0.401	0.789	0.249
NB	4	0.831	0.560	0.839	0.354
LR	4	0.859	0.582	0.867	0.226

Among all the classifiers for the SMOTE generated dataset, SVM performed the best with a c-statistic of 0.865 and an F1-score of 0.594. PCA+SVM did not perform well, which could be due to a scenario where two classes overlapped more after projection [37]. RF and NB both generally performed more poorly than SVM in all feature sizes. DT had the worst performance with the first optimal c-statistic at 0.695. This is mostly because more features were required to build a good classifier based on the nature of DT. LR had the closest performance to SVM. However, previous research has shown that SVM performs better than LR in the case of multivariate and mixture of distributions with a better (or equivalent) misclassification rate [38, 39]. Therefore, SVM was determined to be the most suitable classifier.

The size of the optimal feature set for SVM was six: Pre-NIHSS, age, platelet count, serum glucose, congestive heart failure, and myocardial infarction. These features are routinely collected in acute stroke patients. Pre-NIHSS is a standard measure of impairment caused by a stroke, with most patients receiving Pre-NIHSS assessment immediately after hospital admission. Pre-NIHSS, age, platelet count, and serum glucose level have also been shown to be relevant to stroke outcomes [9, 40, 41]. Therefore, it is reasonable to expect other institutions may maintain these variables and could validate our model. Moreover, the c-statistic was nearly optimal (0.854) with only four features (Pre-NIHSS, age, congestive heart failure, and platelet count), suggesting the possibility of using even less information.

After identifying the top six features for the SVM model, we revisited the database and obtained an additional 39 patients that were not included in the initial filtering due to missing data, but had all six important features. These data were then used as an independent testing dataset to test the classifier. We compared the performance of six-variable SVM classifier trained with balanced and imbalanced testing dataset, and both c-statistic and F1-score were higher when the training dataset was balanced (Table 5). We also performed Mann-Whitney U tests to verify that both SVM1 and SVM2 are statistically better than random ($p=0.020$ and $p=0.011$, respectively), and used a weighted Wilcoxon signed-rank test to verify that the performance difference between SVM1 and SVM2 is

statistically significant (p=0.039). These results suggest that balancing data (i.e., SMOTE) yields an improved classifier for predicting mortality at discharge in acute stroke patients.

Table 5. Performance comparison of six-variable SVMs.

	SVM 1	**SVM 2**
Number of testing data	32 alive, 7 dead	
Number of features	6	
Applied SMOTE?	No	Yes
Number of training data	156 alive , 34 dead	156 alive,156 dead
c-statistic	0.750	**0.781**
F1-score	0.400	**0.500**

5. Conclusion and Future work

In this paper, we compared the performance of SVM, PCA-SVM, DT, RF, NB, and LR models for predicting stroke patient mortality at discharge and determined SVM was the best based on relative c-statistic and F1-score. We then developed an SVM predictive model with six common variables (Pre-NIHSS, age, platelet count, serum glucose level, congestive heart failure, and myocardial infarction) and predicted mortality on the testing data, achieving a 0.781 c-statistic. In addition, we demonstrated the importance of balancing the stroke dataset before training and illustrated the use of the SMOTE algorithm as a solution. We also discussed the benefits of identifying an optimal feature set for building the classifier.

There are a few limitations in our work. First, there are roughly 800 patients available in our dataset, but nearly half of them were missing discharge mRS or other feature values, reducing the size of available data. Our next step will be collecting this information via chart review. Also, we plan to seek independent external datasets on which to validate the proposed models. Second, the distributions of binary features became more imbalanced after SMOTE (Table 2). A possible solution to this problem would be to assign values to binary features by sampling the probability distribution of nearest neighbors, rather than a majority vote.

Lastly, mortality is a standard measure that most existing stroke models aim to predict. However, patients have different degrees of disability even though they are still alive. One clear difference is whether a patient can live independently or dependently [42]. We plan to investigate and improve existing methods in multi-class prediction [43-45] and extend our prediction ability to more classes (independent, dependent, and death), rather than just survival/non-survival.

6. Acknowledgements

This research was supported by National Institutes of Health (NIH) grant R01 NS076534.

References

1. Lloyd-Jones D, Adams R J, Brown TM, et al. Heart disease and stroke statistics—2010 update a report from the american heart association. Circulation. 2010:121: 46-215.
2. Prevalence of disabilities and associated health conditions among adults--united states, 1999. MMWR. Morbidity And Mortality Weekly Report. 50 (7), p. 120.
3. Dobkin BH. Rehabilitation after stroke. New England Journal Of Medicine. 2005:352:1677-1684.
4. Gu Q, Cai Z, Zhu L, Huang B. Data mining on imbalanced data sets. International Conference on Advanced Computer Theory and Engineering. 2008:1020-1024.
5. He H, Garcia EA. Learning from imbalanced data. Knowledge And Data Engineering, IEEE Transactions On. 2009:21 (9):1263-1284.
6. Wei Q, Dunbrack Jr R & L. The role of balanced training and testing data sets for binary classifiers in bioinformatics. Plos One. 2013:8 (7):67863.
7. Counsell C, Dennis M, Mcdowall M, Warlow C. Predicting outcome after acute and subacute stroke development and validation of new prognostic models. Stroke. 2002:33:1041-1047.

8. Teale EA, Forster A, Munyombwe T, Young JB. A systematic review of case-mix adjustment models for stroke. Clinical Rehabilitation. 2012:26 (9):771-786.
9. Weimar C, Konig IR, Kraywinkel K, Ziegler A, Diener H, et al. Age and national institutes of health stroke scale score within 6 hours after onset are accurate predictors of outcome after cerebral ischemia development and external validation of prognostic models. Stroke, 2004:35 (1):158-162.
10. Konig IR, Ziegler A, Bluhmki E, Hacke W, et al. Predicting long-term outcome after acute ischemic stroke a simple index works in patients from controlled clinical trials. Stroke, 2008:39 (6):1821-1826.
11. Adams H, Davis P, Leira E, et al. Baseline nih stroke scale score strongly predicts outcome after stroke a report of the trial of org 10172 in acute stroke treatment (toast). Neurology, 1999:53 (1):126-126.
12. Fonarow GC, Pan W, Saver JL, et al. Comparison of 30-day mortality models for profiling hospital performance in acute ischemic stroke with vs without adjustment for stroke severity. JAMA, 2012:308 (3):257-264.
13. Fonarow GC, Saver JL, Smith EE, et al. Relationship of national institutes of health stroke scale to 30-day mortality in medicare beneficiaries with acute ischemic stroke. Journal Of The American Heart Association. 2012:1 (1):42-50.
14. Saposnik G, Kapral MK., Liu Y, et al. Iscore clinical perspective a risk score to predict death early after hospitalization for an acute ischemic stroke. Circulation. 2011:123 (7):739-749.
15. Saposnik G, Raptis S, Kapral, MK, Liu Y, Tu JV, Mamdani, M. & Austin, P. C. The iscore predicts poor functional outcomes early after hospitalization for an acute ischemic stroke. Stroke, 2011: 42 (12):3421-3428.
16. Saposnik G, Fang J, Kapral MK, et al. The iscore predicts effectiveness of thrombolytic therapy for acute ischemic stroke. Stroke, 2012: 43 (5):1315-1322.
17. Sorcan.ca. (2014). Iscore - sorcan. [online] Retrieved from: http://www.sorcan.ca/iscore/
18. Strbian D, Meretoja A, Ahlhelm F, et al. Predicting outcome of iv thrombolysis--treated ischemic stroke patients the dragon score. Neurology. 2012:78 (6):427-432.
19. Sarraj A, Albright K, Barreto AD, et al. Optimizing prediction scores for poor outcome after intra-arterial therapy in anterior circulation acute ischemic stroke. Stroke, 2013:44 (12):3324-3330.
20. Chawla N, Bowyer K, Hall L, Kegelmeyer P. Smote: synthetic minority over-sampling technique. Journal Of Artificial Intelligence Research. 2002:16:321-357.
21. Harris PA, Taylor R, Thielke R, et al. Research electronic data capture (redcap)—a metadata-driven methodology and workflow process for providing translational research informatics support. Journal Of Biomedical Informatics. 2009:42 (2):377-381.
22. Bamford JS, Ercock P, Warlow C, Slattery J. Interobserver agreement for the assessment of handicap in stroke patients. Stroke. 1989:20 (6):828-828
23. A Shah N, Lin C, Close B, et al. Improving modified rankin scale assessment with a simplified questionnaire. Stroke. 2010:41 (5):1048-1050.
24. Altman DG. (1991). Practical statistics for medical research. London: Chapman And Hall.
25. Sesen MB, Kadir T, Alcantara R, et al. Survival prediction and treatment recommendation with bayesian techniques in lung cancer. AMIA Annual Symposium Proceedings. 2012:838-847.
26. Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995:20 (3):273-297.
27. Rokach L, Maimon OZ. (2008). Data mining with decision trees: theory and applications. Singapore: World Scientific.
28. Breiman L. Random Forests. Machine Learning. 2001:45 (1):5-32.
29. Bewick V, Cheek L, Ball J, et al. Statistics review 14: logistic regression. Crit Care. 2005:9 (1):112-118.
30. Abdi H, Williams LJ. Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics. 2010:2 (4):433-459.
31. Yang C, Duan X. Credit risk assessment in commercial banks based on svm using pca. Proceedings of the seventh International Conference on Machine Learning and Cybernetics. 2008:1207-1211.
32. Gumus E, Kilic N, Sertbas A, Ucan ON. Evaluation of face recognition techniques using pca, wavelets and svm. Expert Systems With Applications. 2010:37 (9):6404-6408.
33. Mierswa I, Wurst M, Klinkenberg R, Scholz M, Euler T. Yale: rapid prototyping for complex data mining tasks. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2006). 2006:935-940
34. Kosorus H, Honigl J, Kung J. Using r, weka and rapidminer in time series analysis of sensor data for structural health monitoring. 2011:306-310.
35. Doukas C, Goudas T, Fischer S, et al. An open data mining framework for the analysis of medical images: application on obstructive nephropathy microscopy images. 2010:4108-4111.

36. Powers DMW. (2011). Evaluation: from precision, recall and f-factor to roc, informedness, markedness & correlation. Journal of Machine Learning Technologies
37. Bishop CM (2006). Pattern recognition and machine learning. New York: Springer.
38. Verplancke T, Van Looy S, Benoit D, et al. Support vector machine versus logistic regression modeling for prediction of hospital mortality in critically ill patients with hematological malignancies. BMC Medical Informatics and Decision Making. 2008:8(1):56.
39. Salazar DA, Velez JI, Salazar JC, Comparison between SVM and Logistic Regression: Which One is Better to Discriminate?. Revista Colombiana de Estadistica Numero especial en Bioestadistica. 2012:35 (SPE2):223-237
40. O'Malley T, Langhorne P, Elton R, Stewart C. Platelet size in stroke patients. Stroke. 1995:26 (6):995-999
41. Alvarez-SabIn J, Molina CA, Ribo M, et al. Impact of admission hyperglycemia on stroke outcome after thrombolysis risk stratification in relation to time to reperfusion. Stroke. 2004:35 (11):2493-2498.
42. Bruno A, Shah N, Lin C, et al. Improving modified rankin scale assessment with a simplified questionnaire. Stroke. 2010:41 (5):1048-1050.
43. Hsu C, Lin C. A comparison of methods for multiclass support vector machines. Neural Networks, IEEE Transactions On. 2002:13 (2):415-425.
44. Liu Y, Wang R, Zeng Y. An improvement of one-against-one method for multi-class support vector machine. Proceedings of the Sixth International Conference on Machine Learning and Cybernetics. 2007:5:2915-2920.
45. Tsoumakas G, Katakis I, Vlahavas I. Mining multi-label data. Springer, 2010:667--685.

A Service Oriented Architecture Approach to Achieve Interoperability between Immunization Information Systems in Iran

Masoud Hosseini, MSc[1, 2], Maryam Ahmadi, PhD[3], Brian E. Dixon, MPA, PhD[1, 2, 4]
[1] Indiana University, School of Informatics and Computing, Department of BioHealth Informatics
[2] Regenstrief Institute, Center for Biomedical Informatics
[3] Tehran University of Medical Science, School of Health Information Management
Tehran, Iran
[4] Center for Health Information and Communication, Department of Veterans Affairs, Veterans Health Administration, Health Services Research and Development Service
Indianapolis, Indiana, USA

Abstract

Clinical decision support (CDS) systems can support vaccine forecasting and immunization reminders; however, immunization decision-making requires data from fragmented, independent systems. Interoperability and accurate data exchange between immunization information systems (IIS) is an essential factor to utilize Immunization CDS systems. Service oriented architecture (SOA) and Health Level 7 (HL7) are dominant standards for web-based exchange of clinical information. We implemented a system based on SOA and HL7 v3 to support immunization CDS in Iran. We evaluated system performance by exchanging 1500 immunization records for roughly 400 infants between two IISs. System turnaround time is less than a minute for synchronous operation calls and the retrieved immunization history of infants were always identical in different systems. CDS generated reports were accordant to immunization guidelines and the calculations for next visit times were accurate. Interoperability is rare or non-existent between IIS. Since inter-state data exchange is rare in United States, this approach could be a good prototype to achieve interoperability of immunization information.

Introduction

Health information technology systems support clinical decision-making, providing access to information or suggesting diagnostic strategies (1, 2). However, clinical information is usually scattered among several independent systems that may be syntactically or semantically incompatible. This incompatibility prevents providers' access to patients' comprehensive medical history and makes the exchange of information very challenging (3).

Standards in the healthcare domain play crucial role in facilitating health information exchange and interoperability. However, information systems today are developed based on different technologies and various proprietary protocols and communication standards are not good solutions for inter-organization interoperability (4). Interoperability between systems is one of the key aspects of enterprise solutions, but because of semantic heterogeneity between applications it is very difficult to achieve (5).

To address interoperability, HL7 works as one of the leading standards for exchange of clinical and administrative data among healthcare information systems. HL7 is a syntactic lingua franca used by healthcare computers to talk to other computers and provide information when and where needed (6). Version 2.x of this standard is largely adopted in immunization information systems (IISs) which are confidential, population-based, computerized information systems that attempt to collect vaccination data about all children within a geographic area (7). IISs can be more efficient if they are improved by immunization decision support system (DSS). Immunization DSS can generate reminder and recall vaccination notices regarding each child's immunization history. Also they can provide official vaccination reports and vaccination coverage assessments for caregivers (8).

To date, IISs are usually isolated systems and querying other systems in order to assemble immunization history is barely possible (9). Service-oriented architecture (SOA) provides a way for these isolated systems to remain viable and responsive to increasing demands for information and analysis (10). By overcoming interoperability limitations, SOA allows existing systems to be integrated by exploiting the pervasive infrastructure of the internet and offers a new chance to continue to use and reuse the business functions provided by legacy systems (11).

Currently in Iran, immunization records are captured on paper records and the information is not stored electronically. Accordingly, providers in Iran's immunization centers do not have access to complete records of an individual's hitherto received vaccines, making it difficult to forecast needed vaccines or schedule the next

immunization event. In this paper, we describe the development of two IISs and provide a solution to achieve interoperability between them. Our goal is to facilitate complete immunization history management and Immunization DSS to help providers in Immunization decision making.

System Development

Immunization Information System

To communicate with other systems, an IIS should be able to capture necessary immunization data into a local database and exchange this information with other systems. Our three main objectives for developing IIS are:

- Support basic business requirements of immunization information system according to immunization guidelines for Iran's vaccination centers (12) such as small clinics.
- Ability to interoperate with other systems through standard healthcare web services.
- Ability to send/receive HL7 v3.0 messages based on standard terminologies.

To meet these objectives, business and system requirements were analyzed and necessary use cases were specified. Then an Immunization Minimum Data Set (MDS) was designed and information model was developed in compliant with the HL7 Reference Information Model (RIM). Afterwards, the local IIS's database was designed and implemented using Microsoft SQL Server.

The HL7 RIM specifies the grammar of v3.0 messages of HL7 standard, specifically, the basic building blocks of the language (nouns, verbs etc.), their permitted relationships and Data Types. These messages are all based on XML. All of the XML tags and attributes used in v3.0 messages are derived from the HL7 RIM and the HL7 v3.0 Data Types (6).

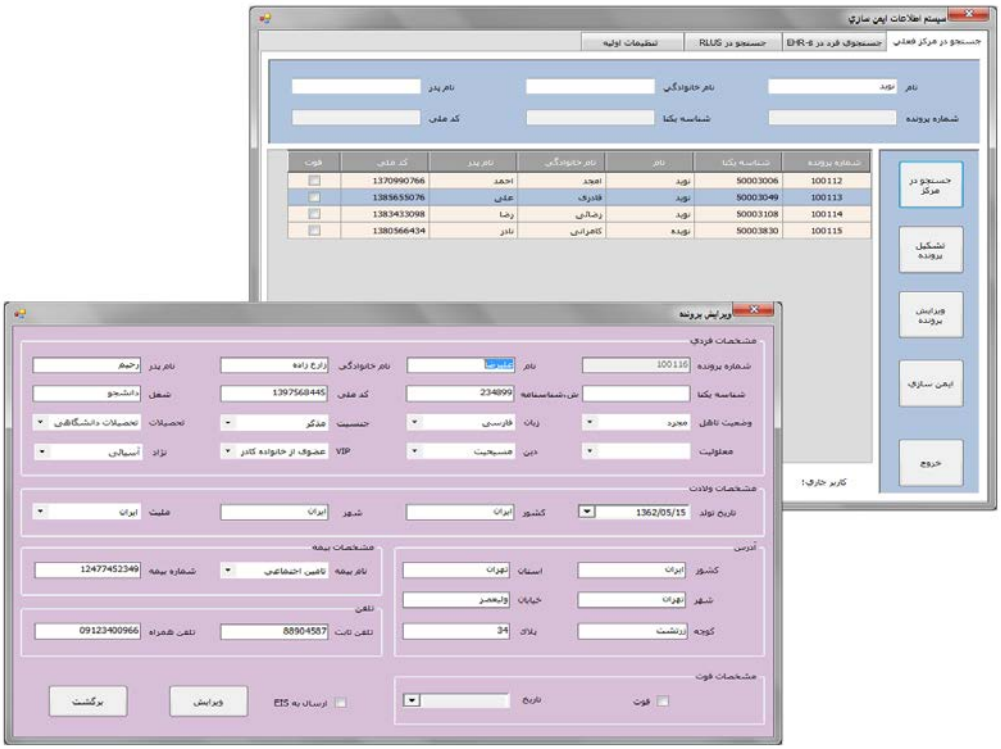


Figure 1. Patient registration forms of implemented IIS

Required vocabularies for concept exchange was extracted from various terminologies like LOINC, ISO 3166, ISO 639-1 and HL7 vocabularies. The system was developed based on three-tier architecture and for transmission of data between these layers we used Data Transfer Object component. Data transfer object (DTO) is an object that carries data between processes in order to reduce the number of method calls. When you are working with a remote interface, each call to it is expensive. The solution is to create a Data Transfer Object that can hold all the data for

the call (13). The system was built in C# language over Microsoft Visual Studio. Figure 1 demonstrates two patient registration forms of implemented IIS which fulfills the required information for HL7 RIM domain model.

EIS and RLUS web services

When a mother takes her child to an immunization center, healthcare providers want to know when and what vaccines has the child received up to now. In other words, they are interested in knowing the child's immunization history. If the entire history is contained in the local center's IIS, there is no need to communicate with other systems to get further information. However, in many cases a child receives vaccines in other locations and data must be retrieved from multiple IISs, making interoperability an important feature for IISs. Deciding if it is required to deliver a vaccine to child, providers need to know what vaccines were previously administered at the other centers. Service Oriented Architecture can help streamline data exchange and promises interoperability between heterogeneous information systems. By using specific web services available in a network, caregivers can retrieve the most current immunization records. SOA environment allows different IISs in heterogeneous platforms and architectures, to negotiate and interoperate with web services without any need to change their architecture.

When there is no information about the current patient in the local system, the first thing the provider needs to know about the child is an identifier which specifies the child as a unique individual among multiple organizations. It is common that the information system of each organization or even department often assigns its own ID that uniquely identify the individuals for its own purposes, with the result that these ID values are meaningless outside that system or organization. These autonomously managed IDs suit the purposes of recording and retrieval of information for the single department or organization, but interoperability requires an ID or matching process to identify an individual uniquely among multiple IISs and then retrieve required information.

In this regard, HSSP recognizes the need for service specifications to support healthcare IT. HSSP is a collaborative effort between HL7 and the Object Management Group (OMG) standards group to address interoperability challenges within the healthcare sector. The activity is an effort to create common "service interface specifications" that ultimately can be tractable within a Health IT context. The stated objective of the HSSP project is to create useful, usable healthcare standards that define the functions, semantics, and technology bindings supportive of system-level interoperability (14).

The Entity Identification Service (EIS) is one of the services recommended by HSSP which is charged with defining the functional specifications of a set of service interfaces to uniquely identify various kinds of entities (e.g. people: patients, providers etc., devices) within disparate systems within a single enterprise and/or across a set of collaborating enterprises(15). EIS defines 'generic' interfaces that would allow name-value pairs to be associated with an entity. EIS is based upon the creation and maintenance of an index consisting of a linked set of Source ID/Entity ID pairs representing the same Real World Entity (RWE). A Source ID and Entity ID are supplied in pairs in order that they may uniquely identify an Entity with the Domain of the EIS. An Entity ID alone uniquely identifies an Entity within the Domain of the Source (16).

By exchanging HL7 messages between IIS and EIS web service we can obtain the child's Unique ID and his/her demographic information. In this effort EIS was implemented according to the service specifications which are published by OMG and HL7 (15, 16). Communicating with this web service, our IIS was able to search using the child's traits in web service and get respective unique ID and demographic information to register a new record in its own local data base.

After resolving the child's identification, the next step is retrieving individual's immunization history based on this unique ID. The Retrieve, Location, and Updating Service (RLUS) is another service specified by HSSP which provides a set of interfaces through which information systems can access and manage information. RLUS allows health data to be located, accessed and updated regardless of underlying data structures, security concerns, or delivery mechanisms (17).

This service is designed for general purposes in the healthcare domain, so we implemented a customized RLUS web service according to (17, 18) with constrained functions to retrieve individual's immunization information. For integrating immunization information of different IIS systems, each IIS sends the information of every Immunization event occurred in the local center to the web service through HL7 v3.0 XML based messages. Figure 2 shows a high level picture of SOA-based interoperability between IISs and web services. In the picture, HL7 messages are wrapped inside SOAP messages and they are exchanged though internet. Also it demonstrates the two immunization DSSs which are embedded in IISs.

Our EIS and RLUS Web services, were built based on Microsoft WCF technology in C# language and the protocol used to exchange XML-based HL7 messages is SOAP and the location of the services and their operations are described through the web services description language (WSDL). The service layer in both web services (EIS and RLUS) is very thin and does not include any business processes or service implementations. Service, data and operation contracts were exposed just in service layer. Regarding the security of interoperability we take advantage of WCF security. Microsoft WCF provides several security features such as transfer security, authorization, and auditing by default, which are responsible for providing message confidentiality, data integrity, and authentication of communicating parties.

Each web service took advantage of multi-tier architecture and has its own independent data base in Microsoft SQL Server. Operations in both of the web services are practically equivalent to HL7 v3.0 interactions specified in HL7 universal domains and Input parameters and return values of the operations are HL7 messages. When a remote IIS invokes one of these operations, depending on operation and its input message, the processes are performed in the web service and the results are returned in the form of an HL7 message. If any error occurs during the message processing, the error code and its description will be returned to the calling IIS. Messages are validated once received by the web service. Message validation is performed locally in EIS and RLUS servers for the pilot system. Messages are checked for both syntax (e.g., Is the message conformant with HL7 message structure?) and semantics (e.g., is valid information from standard terminologies used?). For improper messages, an error message with appropriate error description is returned to the sender.

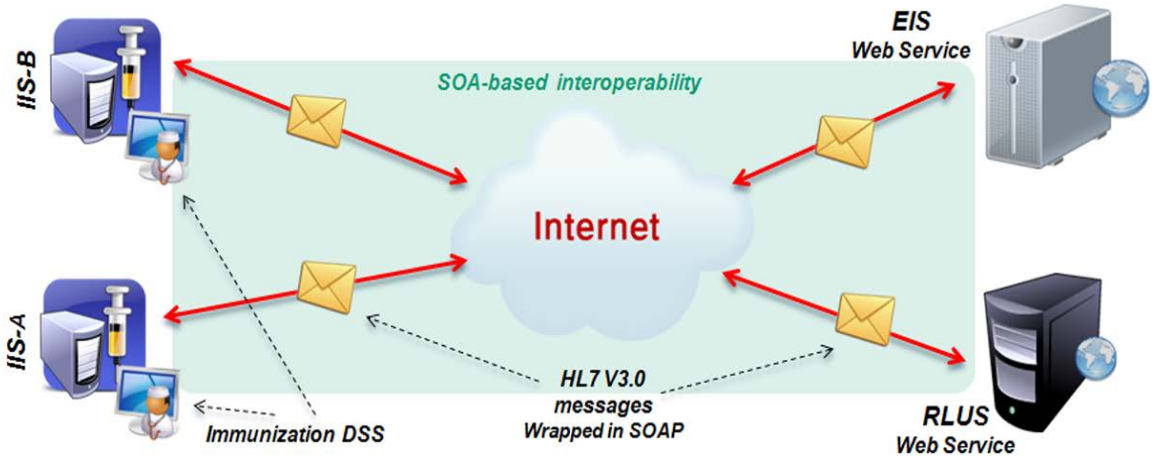


Figure 2. High level picture of SOA-based interoperability between IISs and web services

HL7 v3.0 Messages

The HL7 standard represents the foundation of many healthcare information management systems (19). Version 2.x of this standard is extensively adopted in health care domain and usually IISs use version 2.x rather than v3.0. Immunization information systems which are exchanging HL7 v3.0 messages are rare or non-existent. The reason is that the structure of messages completely changed in version 3.0 and it is cumbersome to upgrade systems from older version to newer one. However, version 2.x enables some optionality in message generation that makes it very hard to apply conformance tests and forces implementers to investigate whether the messages coming from different parties have the same structure of optional fields for a single message (6). This optionality provides multiple ways for generating the same message, therefore when you get one implementation of version 2.x it is likely that you get another one and that is different because of many optional fields and segments in message structure (20).

By presenting RIM data model and using “rigorous analytic and message building techniques”, HL7 addressed the issues of version 2.x and provided messages without optionality issues in version 3.0 (6). By adoption of RIM in this study as information model of IISs and exchanging HL7 v3.0 messages, we tried to improve immunization information interoperability.

```
<id root="15892" AssigningAuthorityName="IdCardNo"/>
<id root="1380790681" AssigningAuthorityName="NationalCode"/>
<name>
    <given>رضا</given>
    <family>مهدوی</family>
</name>
<fatherName>
    <given>احمد</given>
    <family>مهدوی</family>
</fatherName>
<telecom use="HP" value="88909216"/>
<telecom use="MC" value="9124000456"/>
<administrativeGenderCode code="M" displayName="Male" />
<birthTime value="19720330"/>
<addr use="HP">
  <HouseNumber>56</HouseNumber>
  <Alley>کوچه وهاب زاده</Alley>
  <StreetName>خیابان ولیعصر</StreetName>
  <CityName>تهران</CityName>
  <ProvinceName>تهران</ProvinceName>
  <country>US</country>
  <postalCode>13597-55675</postalCode>
</addr>
<maritalStatusCode code="12218"/>
<educationLevelCode code="19179"/>
<disabilityCode code="10182"/>
<religiousAffiliationCode code="19208"/>
<raceCode code="15814"/>
<religiousAffiliationCode code="19208"/>
<languageCommunication>
    <languageCode code="99101"/>
</languageCommunication>
 <birthPlace>
    <addr>
      <country>ایران</country>
      <city>تهران</city>
    </addr>
 </birthPlace>
```

Figure 3. Localized HL7 message in XML format with Father Name and customized home address

According to EIS and RLUS specifications, messages are defined to cover universal needs; however the structure of universal messages does not meet Iran's needs for message exchange. For example, "Father Name" is mandatory in patient's demographic information in Iran, which is not considered in universal messages. Also the structure of patients' address is totally different in Iran. Each universal message includes plenty of classes and attributes that are not in the scope of our work. We solved these problems in two steps. First the messages were customized according to our business requirements and then an intermediary component was developed as an API which isolates the complexity of messages inside of itself and presents very simple operations for message generation. Working with this component is much easier because it plays as a mediator between operations and HL7 complicated classes and data types.

HL7 messages have many data types which are added into the web service's data contracts and by exposing the services through WSDL the local IISs can access to all of HL7 data types in order to massage making. The localized HL7 message in XML format with Father Name and customized home address is shown in Figure 3.

Immunization Decision Support System (DSS)

Timely vaccinations decrease a child's risk of contracting vaccine-preventable disease and prevent disease outbreaks. Reminders and comments generated by immunization DSS can advise providers on which immunizations are needed for a particular patient and provide valuable information for caregivers to deliver better vaccination services to individuals and help people to live healthy. Moreover, Clinical DSS can improve the adherence of providers to clinical guidelines (21).

In this study, the immunization DSS is designed as a guideline based rule engine and embedded into local IISs by which the system can forecast immunization events and give appropriate alerts and recommendations to providers. All of the information needed for rule engine component is provided through communicating with EIS and RLUS.

It is important for providers to know when and what vaccines were administered to patients up to now and whether they were delivered on time or late. Also, they need to know where a patient received the vaccines and who the caregiver was or what vaccines are required to be administered in future and what would be the accurate schedule for next immunization events according to guidelines. Identifying individuals uniquely by exchanging messages with EIS and retrieving integrated immunization history from RLUS web service, the rule engine can apply standard immunization rules on retrieved comprehensive immunization history and generate useful information for providers to address their requirements. Also system can alert the provider that what vaccine is not allowed to be administered or what is needed.

Figure 4 is translation of a sample report from Persian language to English generated by the immunization DSS. This report is based on fake immunization scenario with many delays in attending at immunization center which shows the DSS detects late delivery of vaccines to child. Also, it schedules the future events considering the information from previous events. Moreover, the implemented DSS warns the provider that the Diphtheria, Tetanus and Pertussis (DPT) Vaccination is already recorded for five times and is not allowed to be administered any more when a provider tries to add a new immunization record into system. For each immunization event, the report informs the location of vaccine delivery, provider and some necessary comments. The rules of DSS component have been extracted from guideline of immunization approved by national immunization committee of Iran (12).

Immunization Forecasting Report

Patient Demographics

Patient ID: 30005046 **Name:** Homayoun **Last Name:** Asadi **DOB:** 07/23/2011

Vaccine type: Diphtheria, Tetanus, and Pertussis (DPT)

Vaccine Schedule	Encounter Date	Delay (day)	Next Dose Date	Center	Provider	comments
1st time 2 month old	10/03/2011	10	11/23/2011	IIS_A	Reza Ahmadi	This dose should be delivered to baby between 2 months old to 2 months and 29 days old
2nd time 4 month old	01/31/2012	69	02/29/2012	IIS_A	Reza Ahmadi	Next dose in 1 month
3rd time 6 month old	03/04/2012	4	01/23/2013	IIS_A	Reza Ahmadi	Next dose at the age of 18 months
4th time 18 month old	02/25/2013	33	07/23/2017	IIS_B	Kazem Nabavi	Next dose at the age of 6 years
5th time 6 year old	08/06/2017	14	N/A	IIS_B	Kazem Nabavi	Administration is not allowed after this dose
6th time	07/29/2019	Warning	Warning	IIS_B	Kazem Nabavi	Administration is not allowed!

Figure 4. Sample report in Farsi language generated by the immunization DSS

Results

In order to explore the feasibility of our approach, we conducted a system test and deployed two independent IISs connected to internet without any direct connection to each other and published the EIS and RLUS web services. Then approximately 1500 immunization records for roughly 400 infants from Alzahra Education and Treatment Center in Iran were collected. These records are collected on paper and didn't include any identifying information, so the fake demographic data and ID for children were generated.

The data were randomly entered into two IISs. Considering that each child has more than one immunization record, by random data insertion we guaranteed that immunization information of each individual is scattered in different IISs. During the registration of patients' demographics and immunization records in a random way, the information automatically were sent to EIS and RLUS web services based on HL7 v3.0 messages and integrated in databases located in web services. Then we randomly selected 50 patients and searched their immunization records in both IISs. Based on received immunization history we got and immunization forecast from DSSs embedded in systems.

The HL7 v3.0 messages are specialized from abstract RIM data model to carry domain specific concepts. These messages are converted to large XML documents to transmit between IISs and web services. Although XML messages are large and complex, however, the systems' turnaround time is less than a minute for synchronous operation calls and there was no delay even when the HL7 XML messages are very large.

Due to communication in SOA environment, information is gathered from both IISs and consolidated in web services so the immunization history of patients are always identical in different systems. Systems can update their local data base by synchronizing with web services.

The reliable information supports DSS to generate reliable reports and alerts which is very important in health care sector. For the 50 individuals that we retrieved immunization history, generated warnings or forecasts for the next vaccines were completely accordant to Immunization guidelines and the calculations for next visit dates were 100% accurate. Also the result of DSS was the same in both IISs and there was no difference in generated reports.

Discussion

It has been more than a decade since Iran's government invested in Electronic Health Record (EHR) systems to overcome the challenge of data exchange between Health Information Systems (HIS). However, there has been little progress in computerization of Immunization registries and vaccination information is still captured in traditional methods in Iran. In the United States, minimum functional standards for the operation of immunization information systems were developed by the U.S. Centers for Disease Control and Prevention (CDC), however, currently IIS systems are usually state-centric and there is significant need to data exchange among states and improve data quality and interoperability of systems. Given that residents travel between facilities and across states, healthcare providers require access to comprehensive immunization history of individuals across fragmented EHR, HIS, and IIS. We present a novel prototype for integration and exchange of immunization information based on localized HL7 v3.0 messages for Iran and which takes advantage of SOA specifications. This same approach could be applied in the U.S. for integrating individuals' immunization records by exchanging data in SOA environment and accessing valid and up-to-date information in all of the states. For example, the Nationwide Health Information Network (now referred to as the eHealth Exchange) uses a SOA-based framework for the exchange of HL7 v3.0 continuity of care documents (CCDs) among health information exchanges and federal agencies (22, 23).

To date, v2.x of HL7 standard has been more popular than v3.0 in U.S., mainly due to inertia among commercial EHR companies to move away from legacy platforms. However, the implementation of CDA-requirements in compliance with the 'meaningful use' initiative has resulted in several commercial EHR vendors adopting v3.0 web services, even if they are simply wrappers (24) built on top of existing messaging endpoints. Therefore the U.S. market may see increased availability of v3.0 services in the coming years, and the pilot work in Iran might be useful for EHR vendors that wish to implement interoperable interfaces between various IISs among the various state health departments. Furthermore, given Europe's broader adoption of HL7 v3.0 messaging, the prototype described here might also be useful for sharing immunization records across nation borders.

Our system is built upon standard profiles and technical specifications that will help to be translated by US audience. IHE is one of leading organizations that tries to improve the utilization of computer systems in healthcare to support optimal patient care. EIS web service developed in this study, has many overlaps with IHE Patient Identifier Cross-Referencing (PIXV3) and IHE Patient Demographics Query HL7 V3 (PDQV3) which are presented in IHE IT Infrastructure Technical Framework (25). Both our study and IHE profiles use HL7 RIM model for design of messages and are corresponding with HL7 interactions for message calls. Also, in this study we adopted service profiles and specifications from HSSP and OMG, which are prominent organization in standard developing. Although there are some difference between our work and IHE profile such as service or message customization, however, we believe vendors can use presented model in this work along with IHE profiles which will not necessarily bad thing and they can reproduce this method.

Because of the inherent complexity in HL7 v3.0 messages, it can be confusing to extract and map all of the required IIS information into specific message fields when generating a message. We therefore encapsulated all message generation complexities inside an API component and, by exposing some simple operations to external software developers, better facilitated the process of message generation. This API can be re-used in other systems where the data model is based on the HL7 RIM.

When developing CDSSs, knowledge engineers try to use expression languages like GELLO to represent knowledge in a shareable format among other CDSSs. However, expression languages require an Object Oriented information

model to work efficiently. In this study, we designed information model based on HL7 RIM which is Object Oriented. This model enables our system to use expression languages for defining immunization knowledge and make this knowledge available for any other IIS. Furthermore, CDSSs are also moving towards SOA-based frameworks to enable distributed knowledge delivery across a range of clinical guidelines to improve health care quality, safety, and efficiency (26).

Conclusion

Providers require up-to-date access to information to accurately forecast vaccines as well as make other clinical decisions. EHR, HIS, and IISs have generally been developed and deployed as silos enabling fragmented, cumbersome access to patient information. Greater deployment of health information exchange technologies like SOA and HL7 make it possible to engineer distributed IIS and CDSS. We describe an approach for Iran that has potential for the U.S. and other nations. Future research and development is promising to incrementally advance health information technologies towards meeting provider and patient needs for real-time access to information wherever it may be located.

References

1. Vittorini P, Michetti M, di Orio F. A SOA statistical engine for biomedical data. Computer Methods and Programs in Biomedicine. 2008;92(1):144-53.
2. Zolnoori M, Jones JF, Moin M, Heidarnejad H, Fazlollahi MR, Hosseini M. Evaluation of user interface of computer application developed for screening pediatric asthma. Universal Access in Human-Computer Interaction Applications and Services for Quality of Life: Springer; 2013. p. 563-70.
3. Martínez-Costa C, Menárguez-Tortosa M, Fernández-Breis JT. An approach for the semantic interoperability of ISO EN 13606 and OpenEHR archetypes. Journal of Biomedical Informatics. 2010;43(5):736-46.
4. Blazona B, Koncar M. HL7 and DICOM based integration of radiology departments with healthcare enterprise information systems. International Journal of Medical Informatics. 2007;76(Supplement 3):S425-S32.
5. Ralyté J, Jeusfeld MA, Backlund P, Kühn H, Arni-Bloch N. A knowledge-based approach to manage information systems interoperability. Information Systems. 2008;33(7-8):754-84.
6. Tim Benson. Principles of Health Interoperability HL7 and SNOMED. First ed: Springer; 2010.
7. Grannis S, Dixon BE, Brand B. Leveraging Immunization Data in the e-health Era: Exploring the Value, Tradeoffs, and Future Directions of Immunization Data Exchange. Atlanta: Public Health Informatics Institute, 2010. Archived at: http://www.webcitation.org/6O2vVosgF
8. Centers for Disease Control (CDC). Immunization Information Systems 2010. Available from: http://www.cdc.gov/vaccines/vac-gen/policies/ipom/downloads/chp-03-immz-info-sys.pdf.
9. Public Health Data Consortium. BUILDING A ROADMAP FOR HEALTH INFORMATION SYSTEMS INTEROPERABILITY FOR PUBLIC HEALTH. 2007.
10. Arzt NH, Metroka A. Service-Oriented Architecture: Immunization Information System Case Studies. 44th National Immunization Conference; April 212010.
11. Canfora G, Fasolino AR, Frattolillo G, Tramontana P. A wrapping approach for migrating legacy system interactive functionalities to Service Oriented Architectures. Journal of Systems and Software. 2008;81(4):463-80.
12. Schedule and Guideline of Immunization Approved by National Immunization Committee. 7ed: Iran Ministry of Health and Medical Education; 2010.
13. Fowler M, Rice D, Foemmel M, Hieatt E, Mee R, Stafford R. Data Transfer Object. Patterns of Enterprise Application Architecture: Addison Wesley; 2003. p. 347-56.
14. Healthcare Services Specification Project (HSSP). Available from: http://hssp.wikispaces.com.
15. Eckman B, Honey A, Batra V, Bennett C, Dutta A, Forslund D, et al. Service Functional Model Specification - Entity Identification Service (EIS). HL7 Corporation; 2006.
16. Object Management Group Inc. (OMG). Entity Identification Service (EIS) Specification. 2009.
17. Koisch J, Rubin K, Honey AP, Robinson S, Kawamoto K, Koisch J. Service Functional Model Specification - Retrieve, Locate, and Update Service. HL7 Corporation; 2006.
18. Object Management Group Inc. (OMG). Retrieve, Locate, and Update (RLUS) Service Specification. 2009.
19. Health Level Seven web site. "About HL7". Available from: http://www.hl7.org/about/index.cfm.

20. Beeler GW. HL7 version 3-an object-oriented methodology for collaborative standards development. Int J Med Inform. 1998 Feb;48(1-3):151-61.
21. Zhu VJ, Grannis SJ, Rosenman MB, Downs SM. Implementing broad scale childhood immunization decision support as a web service. AMIA Annu Symp Proc. 2009;2009:745-9.
22. Simonaitis L, Dixon BE, Belsito A, Miller T, Overhage JM. Building a production-ready infrastructure to enhance medication management: early lessons from the nationwide health information network. AMIA Annu Symp Proc. 2009;2009:609-13.
23. Bouhaddou O, Bennett J, Teal J, Pugh M, Sands M, Fontaine F, et al. Toward a virtual lifetime electronic record: the department of veterans affairs experience with the nationwide health information network. AMIA Annu Symp Proc. 2012;2012:51-60.
24. Saez C, Breso A, Vicente J, Robles M, Garcia-Gomez JM. An HL7-CDA wrapper for facilitating semantic interoperability to rule-based Clinical Decision Support Systems. Comput Methods Programs Biomed. 2013 Mar;109(3):239-49.
25. IT Infrastructure Technical Framework. Available: http://www.ihe.net/technical_frameworks/#IT
26. Paterno MD, Goldberg HS, Simonaitis L, Dixon BE, Wright A, Rocha BH, et al. Using a Service Oriented Architecture Approach to Clinical Decision Support: Performance Results from Two CDS Consortium Demonstrations. AMIA Annu Symp Proc. 2012;2012:690-8.

Syndromic surveillance in an ICD-10 world

Achala Jayatilleke MBBS, MSc, PhD, Jeffrey Kriseman MS, PhD, Lisa H Bastin MA, Umed Ajani MBBS, MPH, Peter Hicks MA, MPH

Centers for Disease Control and Prevention, Atlanta, GA, USA

Abstract

The Centers for Disease Control and Prevention's BioSense program is an integrated national public health surveillance system that uses electronic medical record (EMR) data to provide situational awareness for all-hazard health-related events. Because the system leverages International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) coded data from EMRs for syndromic surveillance, the upcoming Health and Human Services-mandated transition from ICD-9-CM to ICD-10-CM will have a significant impact. To translate across the two encoding systems, we developed a Mapping Reference Table (MRT) for the ICD-9/10 transition. We extracted ICD-9-CM codes binned to predefined syndromes and mapped each to its corresponding ICD-10-CM code(s). Then, we translated the output ICD-10-CM codes back to ICD-9-CM through a reverse translation validation process. Throughout the translation process, we examined outputs manually and incorporated annotated results into the MRT. The resulting MRT can be used to refine and update each existing syndromic surveillance definition in BioSense to be compatible with ICD-10-CM and consistently classify or bin any given emergency department visit into the correct syndrome regardless of coding system.

Disclaimer

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention (CDC).

Introduction

The Public Health Security and Bioterrorism Preparedness and Response Act of 2002 mandated the establishment of public health surveillance systems for early detection and rapid assessment of potential bioterrorism-related illness. As a response to that mandate, the Centers for Disease Control and Prevention (CDC) launched the BioSense Program in 2003 to establish an integrated national public health surveillance system.(1) BioSense began receiving data feeds in 2003 from outpatient clinics maintained by the Department of Veteran Affairs (VA) and Department of Defense (DoD).(2, 3) In 2005 BioSense began to receive data from non-federal civilian hospitals as well. Initially, BioSense defined 11 broad major syndromes (botulism-like, fever, gastrointestinal, hemorrhagic illness, localized cutaneous lesion, lymphadenitis, neurological, respiratory, rash, severe illness or death, and specific infection) and more-specific sub-syndromes based on chief complaint and final diagnosis data.(4) Although BioSense was initially launched as a surveillance system for early event detection and rapid assessment of potential bioterrorism-related illness over time, it has transformed to a public health surveillance system that provides situational awareness for all-hazard health-related events.(5)

CDC started redesigning BioSense in 2010, and the system was launched as BioSense 2.0 in 2012. The latest iteration of BioSense is a streamlined collaborative data-exchange system that enables its users to share health-related data, quickly track adverse or anomalous health issues, and share this information rapidly among other public health jurisdictions participating in the system.(5–7) Currently, as of July 2014, BioSense receives data feeds from 3,377 (1,920 non-federal and 1,457 federal) facilities and tracks approximately 130 pre-defined syndromes.(8)

Leveraging data from electronic medical records (EMRs) is a modern cornerstone for public health surveillance activities.(9) BioSense uses EMR data (such as chief complaint and final diagnosis codes) to provide national, regional, and local situational awareness for all-hazard health-related events and to inform a wide range of public health activities. Within the BioSense program, International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM)-coded data are extracted from EMR data provided by participating healthcare facilities.(4,8) These ICD-9-CM codes are binned and classified into syndromes and used to create temporal and geographic indicators and trend-lines for injury, disease, conditions, healthcare utilization, and adverse or anomalous event monitoring.(4,5)

In 2009, the Department of Health and Human Services mandated that a transition from ICD-9 to ICD-10 will occur on October 1, 2014 (now extended to October 1, 2015), for all entities covered under the Health Insurance Portability Accountability Act (HIPAA) and for healthcare systems that submit reimbursements to the Centers for Medicare and Medicaid (CMS) in the United States.(10) As many other secondary users of final diagnosis-coded data, public health surveillance systems will be significantly affected by this transition. To receive, analyze, interpret, and report upon ICD-9 and ICD-10 encoded data, public health surveillance systems that used ICD-9-CM codes must modify the existing database structure, modify data extraction rules, and create well-defined Mapping Reference Tables (MRT) to bridge the gap across the two encoding systems.

The most challenging aspect of the ICD-9/10 transition for public health surveillance systems will be to develop flexible and standardized solutions to accommodate analysis across two different code sets in 2015 and beyond. In addition, a meaningful baseline must be developed that can be used for benchmarking after the transition. Public health surveillance systems should expect to receive ICD-9-coded data for the first nine months of 2015 and ICD-10-coded data for the last three months of 2015; in addition, public health surveillance systems should plan for the possibility of receiving and analyzing a

mixture of both ICD-9 and ICD-10-coded data for some period of time following October 1, 2015, which will make the situation more complicated. In this context it is essential to have a method to identify the type of original codes received by public health surveillance systems and to have the compatibility to seamlessly map across the two code sets.

We carried out this project to describe the process of developing an MRT for ICD-9/10 transition to be used to translate across the code sets and to propose a methodology for a meaningful baseline to be used in the BioSense system after October 1, 2015.

Methods

We extracted ICD-9-CM codes binned to all predefined BioSense syndromes from the list of syndromes developed by the BioSense community, which includes federal, state, and local public health partners.(8) We updated each of the extracted codes to 2014 ICD-9-CM codes by using the Code Translation Tool (CTT) developed by 3M for the translation process. (11) By using CTT, we translated the resulting ICD-9-CM 2014 codes to ICD-10-CM 2014 codes. Then, we translated the output ICD-10-CM codes back to ICD-9-CM by using a reverse translation validation process to ensure that the appropriate codes were correctly identified at the onset of the translation process. Throughout the translation process, outputs were individually examined manually and annotated results were incorporated into the MRT by an epidemiologist and a clinician. We used ICD-9 and ICD-10 complete official code sets for the manual process.(12, 13)

By using the resulting MRT, we computed ICD-9-CM codes for use cases of ICD-10-CM codes and vice versa. Computed ICD-9-CM codes were then binned to existing BioSense syndromes. Similarly, we attempted to bin computed ICD-10-CM codes to the existing BioSense syndromes. During this process, we also developed an algorithm that can identify whether the original code belongs to ICD-9-CM or ICD-10-CM. In addition, we used the resulting MRT to refine and update each of the existing syndromic surveillance definitions used in the BioSense program to be ICD-10-CM compatible.

Results

The output MRT comprises four columns: 1) the original ICD-9-CM, 2) the updated 2014 ICD-9-CM codes, 3) corresponding ICD-10-CM codes, and 4), lastly, a value to indicate the level of confidence for each individual mapping in regards to a specific syndrome. Table 1 depicts a section of MRT for syndrome asthma as an example. We indicated the level of confidence by using four values (Table 2).

In some instances where a single ICD-9-CM code is insufficient to represent the relevant ICD-10-CM codes, we introduced an additional column in the MRT named *Supplementary ICD-9-CM Code*. Similarly, we introduced supplementary ICD-10-CM codes where a single ICD-10-CM code is insufficient to represent some concepts represented by ICD-9-CM.

Table 3 shows the level of confidence of ICD-10-CM codes for five selected syndromes. Figure 1 illustrates the number of ICD-9-CM and ICD-10-CM codes binned to five selected syndromes.

Leveraging the MRT, we computed ICD-9-CM and ICD-10-CM codes for hypothetical use cases and binned them to the existing BioSense syndromes. Table 4 shows the computed codes and binning of codes to the syndromes asthma and chronic obstructive pulmonary disease (COPD).

Figure 2 shows the algorithm we used to identify the type of original code. This algorithm describes how an ICD-9-CM diagnosis code can be differentiated from that of ICD-10-CM.

Discussion

The ICD-9-CM to ICD-10-CM transition will have a significant impact on the BioSense program. Due to the complexity and higher level of specificity in ICD-10-CM codes, ICD-9-CM codes pertinent to syndromes within BioSense cannot be automatically translated into ICD-10-CM codes. Existing translation tools are insufficient and frequently provide results that are either inaccurate or incompatible with the syndromic surveillance concept under review. Therefore, we developed an MRT to be used by the BioSense program with manual expert review and input based on General Equivalence Mappings (GEMs) files prepared by CMS.(14)

By leveraging the developed MRT, ICD-10-CM codes can be down-coded and translated into ICD-9-CM codes. The resulting ICD-9-CM codes can be binned to existing BioSense syndromes. Similarly, ICD-9-CM codes can be up-coded and translated to ICD-10-CM by using the MRT. The output (the resulting ICD-10-CM codes) can then be incorporated into new syndromic surveillance visit (case) definitions that can be incorporated into the overall BioSense system or appended to a given record to allow a given event to be analyzed in either code set. The developed MRT can be used to refine and update each of the existing syndromic surveillance definitions used in the BioSense program to be ICD-10-CM compatible. However, due to the higher specificity of ICD-10-CM codes, existing BioSense 2.0 syndromes should be reviewed based on ICD-10-CM codes to achieve syndromic surveillance objectives more effectively.

Syndromic surveillance data are of limited use without a reliable and interpretable referential baseline. The BioSense program uses the past two years of data to define a referential baseline. However, after October 1, 2015, it will be impossible to compute a referential baseline unless there is a platform to accommodate both ICD-9-CM and ICD-10-CM codes. Referential base lines can be computed for both ICD-9 and ICD-10 code sets by using our developed MRT. These calculated referential baselines can be used until all healthcare facilities providing data to the BioSense program fully transition to the ICD-10 coding system or until a complete two-year referential baseline is computed (October 2017).

After October 1, 2015, public health surveillance systems may receive a mixture of both ICD-9 and ICD-10 codes. Because ICD-10-CM diagnosis codes begin with an alpha character (e.g., J45.30, J45.20), a majority of ICD-10-CM codes can be differentiated from ICD-9-CM codes that begin with a numeric character (e.g., 493.0, 493.1). However, there are similarities in ICD-10-CM codes and ICD-9-CM codes starting with an alpha character (E and V).(15) To address this issue, we developed an algorithm for this project that can be used to validate and differentiate ICD-10 codes from ICD-9 codes after the transition. However, in some occasions where ICD-10 codes are identical to ICD-9 codes, additional information will be required for the differentiation.

The lack of ICD-10-CM-based data to test the developed MRT is a limitation of this project. Currently, we are developing a synthetic ICD-10-CM data set by using probability theories based on historical ICD-9-CM data. This data set can be used to test the developed MRT several months before the real transition occurs.

In conclusion, the developed MRT can be used as a common language and reference to receive, integrate, and classify healthcare visits to syndromic surveillance definitions and develop computed referential baseline data regardless of ICD code set. However, to receive the benefit of the more specific ICD-10 codes, existing BioSense syndrome definitions should be revised based on ICD-10-CM codes to achieve syndromic surveillance objectives more effectively. The MRT and the code set identification algorithm can be leveraged beyond the BioSense program and will be made available to state and local partners for potential inclusion in their specific syndromic surveillance systems.

Table 1. Mapping Reference Table (MRT) for selected codes of asthma syndrome

Original ICD-9 CM	2014 ICD-9 CM	Corresponding ICD-10 CM	Level of Confidence
493	493.00	J45.20	A
493	493.00	J45.30	A
493	493.00	J45.40	A
493	493.00	J45.50	A
493	493.01	J45.22	A
493	493.01	J45.32	A
493	493.01	J45.42	A
493	493.01	J45.52	A
493	493.90	J45.909	A
493	493.90	J45.998	A
493	493.91	J45.902	A
493	493.92	J45.901	A
493	493.81	J45.990	A
493	493.82	J45.991	A
--	491.20*	J44.9	D
--	491.21*	J44.1	D
--	491.22*	J44.0	D
--	496*	J44.9	D

* Resulted from reverse translation process

Table 2. Level of confidence

Level of Confidence	Description
A	Consists of codes that reflect general symptoms of the syndrome group and also include codes for the bioterrorism diseases of highest concern or those diseases highly approximating them.
B	Consists of codes that might normally be placed in the syndrome group, but daily volume could overwhelm or otherwise detract from the signal generated from the Category 1 code set alone.
C	Consists of codes that might normally be placed in the syndrome group, but daily volume could overwhelm or otherwise detract from the signal generated from the Category 1 code set alone.
D	Not a match

Table 3. Level of confidence of translated ICD-10 codes for selected syndromes

Syndrome	Level of Confidence	Number of ICD-10 codes	%
Asthma	A	33	89.2
	B	0	0.0
	C	0	0.0
	D	4	10.8
Diphtheria	A	10	62.5
	B	6	37.5
	C	0	0.0
	D	0	0.0
Lymphadenitis	A	11	36.7
	B	0	0.0
	C	3	10.0
	D	16	53.3
Anthrax	A	7	100.0
	B	0	0.0
	C	0	0.0
	D	0	0.0
Convulsions	A	4	100.0
	B	0	0.0
	C	0	0.0
	D	0	0.0

Figure 1. Number of ICD-9-CM and ICD-10-CM codes per selected syndromes

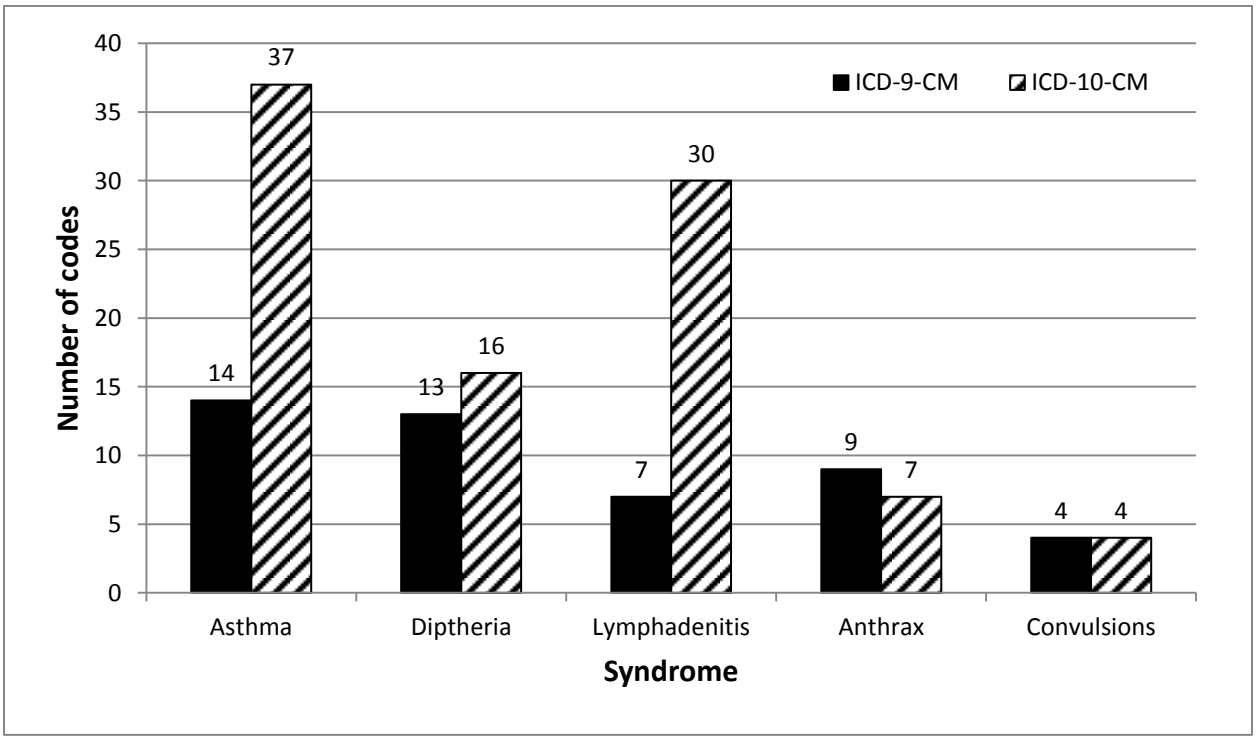


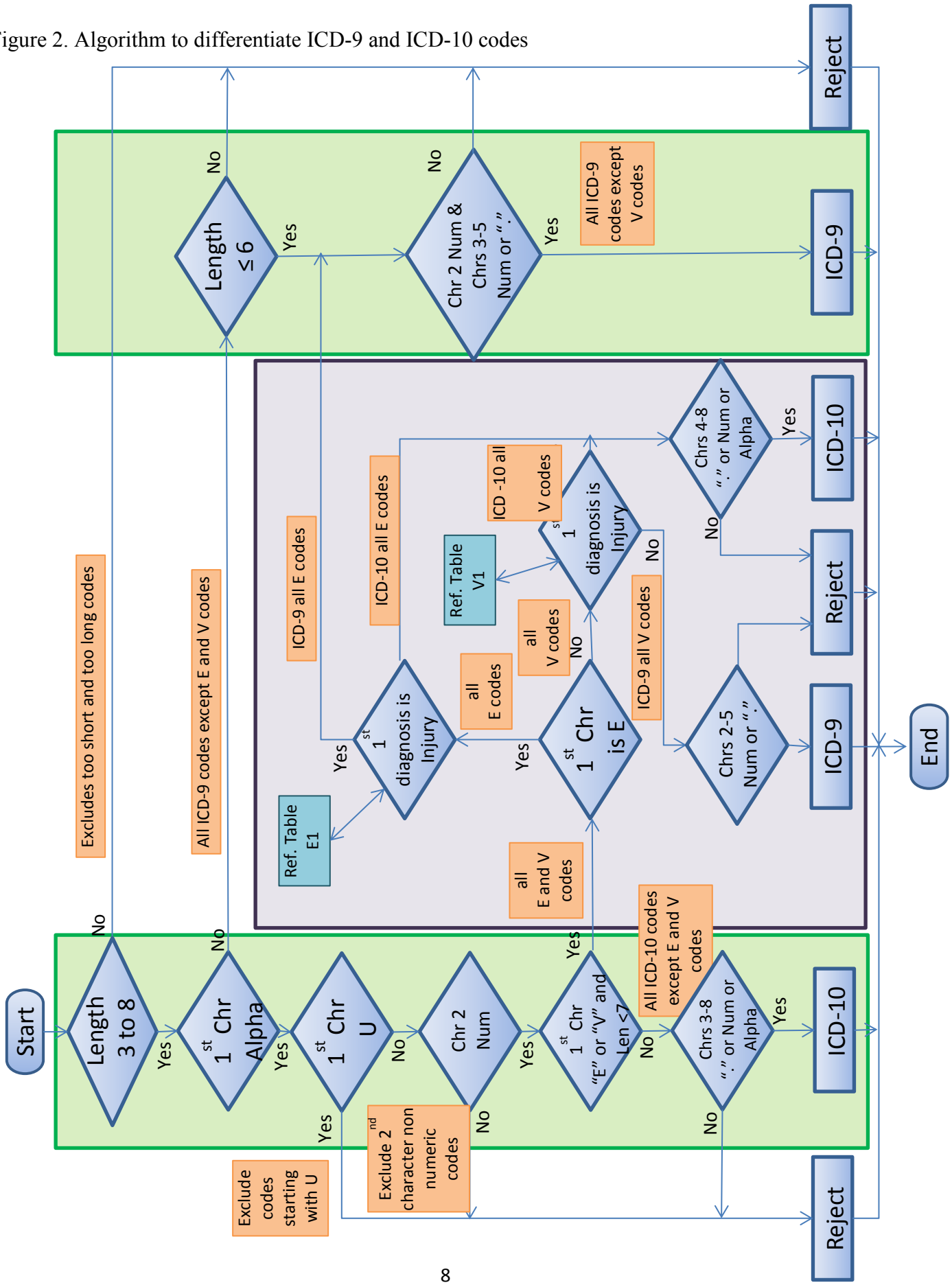
Table 4 Binning of computed codes to asthma and chronic obstructive pulmonary disease (COPD) syndromes

Visit ID	Original code [A]	Code type	Computed ICD-10 code [B]	Computed ICD-9 code [C]	ICD-9 [A+C]	ICD-10 [A+B]	Syndrome ICD-9	Syndrome ICD_10
XXX1	493.00	ICD-9	J45.20	--	493.00	J45.20	Asthma*	Asthma
XXX2	493.20	ICD-9	J44.9	--	493.20	J44.9	Asthma	COPD
XXX3	J45.909	ICD-10	--	493.90	493.90	J45.909	Asthma	Asthma
XXX4	J44.9	ICD-10	--	493.20	493.20	J44.9	Asthma	COPD
XXX5	J44.9	ICD-10	--	491.21	491.21	J44.9	COPD†	COPD
XXX6	493.90	ICD-9	J45.998	--	493.90	J45.998	Asthma	Asthma
XXX7	J44.9	ICD-10	--	496	496	J44.9	COPD	COPD
XXX8	J44.0	ICD-10	--	493.21	493.21	J44.0	Asthma	COPD
XXX9	J44.0	ICD-10	--	491.20	491.20	J44.0	COPD	COPD

* ICD-9 CM codes included in Asthma syndrome definition - 493

† ICD-9 CM codes included in COPD syndrome definition - 491, 492

Figure 2. Algorithm to differentiate ICD-9 and ICD-10 codes



References

1. Centers for Disease Control and Prevention. BioSense background. Available at: http://www.cdc.gov/biosense/background.html. Accessed on 12 July 2013.
2. Centers for Disease Control and Prevention. BioSense: Implementation of a National Early Event Detection and Situation Awareness system. *MMWR* 2005, 54(1):11-19.
3. Tokars J, English R, McMurray P, Rhodes B. Summary of data reported to CDC's national automated biosurveillance system, 2008. *BMC Med Informatics and Decision Making* 2010, **10**:30
4. Centers for Disease Control and Prevention. Syndrome definitions for diseases associated with critical bioterrorism-associated agents, October 23, 2003. Available at: http://www.bt.cdc.gov/surveillance/syndromedef/. Accessed on 5 July 2013.
5. Centers for Disease Control and Prevention. BioSense program. Available at: http://www.cdc.gov/biosense/index.html. Accessed on 10 August 2013.
6. Centers for Disease Control and Prevention. BioSense 2.0. Available at: http://www.cdc.gov/biosense/biosense20.html. Accessed on 25 September 2013.
7. BioSense Redesign Collaboration Site-Google Sites. Available at: https://sites.google.com/site/biosenseredesign/about. Accessed on
8. BioSense Syndrome List. Available at: https://biosen.se/syndrome-info.php#syndromes. Accessed on 15 February 2014.
9. Eggleston EM, Weitzman ER. Innovative uses of electronic health records and social media for public health surveillance. *Curr Diab Rep*, 2014;14:468
10. Center for Medicaid and Medicare. The ICD-10 Transition: An Introduction. Available at: https://www.cms.gov/Medicare/Coding/ICD10/Downloads/ICD10_Introduction_060413[1].pdf. Accessed on 5 July 2014.
11. 3M. 3M ICD-10 Code Translation Tool (CTT). Available at: https://www.3micd-10maps.com/CTT/. Accessed on 22 August 2013.
12. American Medical Association. 2014 ICD-9-CM for hospitals, volumes 1,2 and 3. ISBN 978-0323186742.
13. Optum. ICD-10-CM The Complete Official Draft Code Set. Sept 2013. ISBN - 978-1622540679. 214 edition.
14. Center for Medicaid and Medicare. 2014 ICD-10-CM and GEMs. Available at: https://www.cms.gov/Medicare/Coding/ICD10/2014-ICD-10-CM-and-GEMs.html . Accessed on 20 November 2013.
15. Centers for Disease Control and Prevention. International Classification of Diseases, (ICD-10-CM/PCS) Transition. Available at: http://www.cdc.gov/nchs/icd/icd10cm_pcs_background.htm. Accessed on 6 July 2014.

Divisive Hierarchical Clustering towards Identifying Clinically Significant Pre-Diabetes Subpopulations

Era Kim, MS[1], Wonsuk Oh, MS[1], David S. Pieczkiewicz, PhD[1], M. Regina Castro, MD[2], Pedro J. Caraballo, MD[2], Gyorgy J. Simon, PhD[1],
[1]Institute for Health Informatics, University of Minnesota, Minneapolis, MN.
[2]Mayo Clinic, Rochester, MN

Abstract
Type 2 Diabetes Mellitus is a progressive disease with increased risk of developing serious complications. Identifying subpopulations and their relevant risk factors can contribute to the prevention and effective management of diabetes. We use a novel divisive hierarchical clustering technique to identify clinically interesting subpopulations in a large cohort of Olmsted County, MN residents. Our results show that our clustering algorithm successfully identified clinically interesting clusters consisting of patients with higher or lower risk of diabetes than the general population. The proposed algorithm offers fine control over the granularity of the clustering, has the ability to seamlessly discover and incorporate interactions among the risk factors, and can handle non-proportional hazards, as well. It has the potential to significantly impact clinical practice by recognizing patients with specific risk factors who may benefit from an alternative management approach potentially leading to the prevention of diabetes and its complications.

Introduction

Type 2 Diabetes Mellitus is one of the fastest growing chronic diseases in the United States, with a profound influence on public health quality and cost[1,2]. It is a progressive disease, associated with an increased risk of developing serious cardiac, vascular, renal and ophthalmological complications, and it is one of leading causes of death[2]. With no cure per se, prevention and management are of paramount importance. As effective preventive measures such as lifestyle change and drug therapy exist[3,4], early identification and management of patients at high risk is an important healthcare need.

Numerous diabetes risk indices aimed at early identification of patients at high risk have been developed[5]. Arguably, the most popular such index is the Framingham score[6], which has gained wide acceptance in clinical practice. The Framingham model assigns a risk to a patient based on the risk factors the patient presents with and the resulting score can be used to stratify patients into low, moderate, or high-risk groups. Almost all indices, the Framingham score included, estimate the risk of diabetes in an additive fashion, assuming that the risk factors act independently.

Interactions among risk factors are known to exist[7–11]. Recent work[7–10] aimed to address interactions, most prominently through the application of association rule mining (ARM)[12–15]. ARM was specifically designed to discover sets of associated risk factors, along with the affected subpopulations. While association does not always translate into (non-additive) interaction, it often does. Given its ability to seamlessly incorporate interactions, ARM has successfully identified patient subpopulations that face significantly increased or decreased risk of diabetes[7,11]. Another beneficial characteristic of the ARM model lies in the straightforward interpretability of the individual rules. Thus the ARM model does not just provide a risk estimate, but it also offers a "justification" in the form of the associated risk factors in the rule.

ARM has its own shortcomings. While interpretability is one of the hallmarks of the ARM modeling approach, ARM algorithms tend to extract combinatorially large sets of redundant rules, which quickly erodes interpretability. Under these conditions, it is necessary to offer fine control of the amount of details the ARM model extracts; however, this is precisely where ARM falls flat. When ARM discovers a manageable number of rules, they tend to be too general to be useful; when the model is sufficiently detailed to give the user new insights, the sheer number of rules impedes interpretation. There is a reason for this phenomenon. The ARM rule set is highly redundant: the same subpopulation is described by an exponential number of rules, each rule associating the subpopulation in question with a different set of risk factors. This unfortunate property obfuscates the disease mechanism.

In this work, we propose the use of a novel divisive hierarchical clustering[16] technique, which retains most of the advantages of ARM, while it alleviates the interpretability issues. From a hierarchical clustering, depending on the desired amount of detail, many clusterings can be extracted. Each clustering consists of a varying number of clusters, is complete (they include all patients) and non-overlapping (each patient belongs to exactly one cluster).

Our proposed approach retains all advantageous properties of ARM and alleviates its primary shortcoming: interpretability is enhanced through the elimination of redundancy and through lending the user fine control over the amount of details the clustering should incorporate.

Methods

Data

In this study we utilized a large cohort of Olmsted County, Minnesota, residents identified by using Rochester Epidemiology Project resources. The Rochester Epidemiology Project (REP)[17] is a unique research infrastructure that follows residents of Olmsted Co., MN over time. The baseline of our study was set at Jan. 1, 2005. We included all adult Mayo Clinic patients with research consent, who are part of the REP, resulting in a study cohort of 69,747 patients. From this cohort, we excluded all patients with a diagnosis of diabetes before the baseline (478 patients), missing fasting plasma glucose measurements (14,559 patients), patients whose lipid health could not be determined (1,023 patients) and patients with unknown hypertension status (498 patients). Our final study cohort consists of 52,139 patients (overlaps between the groups exist) who were followed until the summer of 2013.

We collected demographic information (age, gender, body mass index BMI), laboratory information (primarily fasting plasma glucose and lipid panel), vital signs (blood pressure and pulse), relevant diagnosis diagnoses (obesity, hyperlipidemia, hypertension, renal failure and various cardiac and vascular conditions), aspirin use, and medications used to treat hypertension and hypercholesterolemia. Additional known risk factors for diabetes (such as tobacco usage) were also included.

Features

To enhance the interpretability of our results, the variables were transformed into binary variables to indicate the presence and severity of risk factors. These variables are typically constructed as a meaningful combination of diagnoses, abnormal vital signs, abnormal laboratory results, and use of medications by drug class. Laboratory results were considered abnormal when they exceeded the cutoffs published in the American Diabetes Association (ADA)[18] guidelines. Table 1 shows the definitions of the variables used henceforth.

Table 1. Predictors and their definitions

Predictors	Definitions
Demographics	
age.18+	Age > 18 and < 45
age.45+	Age ≥ 45 and < 65
age.65+	Age ≥ 65
genderM	Male
Comorbidities	
obese	Obesity (BMI ≥ 30 or diagnosis)
tobacco	Current smoker
renal	Renal disease
chf	Congestive Heart Failure
ihd	Ischemic Heart Disease
Major risk factors and their severities	
ifg.no	Normo-glycemic patients: fasting plasma glucose (FPG) ≤ 100
ifg.pre1	Impaired Fasting Glucose level 1: FPG > 100 and ≤ 110
ifg.pre2	Impaired Fasting Glucose level 2: FPG > 110 and ≤ 125
htn.no	No indication of Hypertension: no diagnosis of HTN, no hypertensive drugs are described and blood pressure results (if present) are normal.

htn.any	Indication of Hypertension exists in the form of either a HTN diagnosis or abnormal blood pressure measurement
htn.tx	Hypertension required therapeutic intervention; however, at most 3 HTN drugs were prescribed.
htn.pers	Persistent Hypertension. Patients present with abnormal blood pressure measurements despite having been prescribed 3 or more drugs; or they are prescribed 4 or more drugs (regardless of blood pressure results).
hyperlip.no	No indication of Hyperlipidemia: no diagnosis of hyperlipidemia, no cholesterol drugs and no abnormal lipid panel results are present.
hyperlip.any	Indication of Hyperlipidemia exists in the form of diagnosis or abnormal laboratory results.
hyperlip.tx	Hyperlipidemia with therapeutic intervention: a diagnosis code or abnormal laboratory result indicates hyperlipidemia and a single cholesterol drug is prescribed.
hyperlip.multi	Hyperlipidemia requiring multi-drug intervention: multiple cholesterol drugs are prescribed.

Patient Clustering

The purpose of clustering is to partition patients into groups (clusters), such that patients within the same cluster are more similar to each other than to patients in a different cluster. Formally, in our application, a **cluster** is a set of patients, who share risk factors relevant to diabetes progression and have similar diabetes risk. A **clustering** is a non-overlapping complete set of clusters. A clustering is *complete* in the sense that all patients in the population are assigned to a cluster, and it is *non-overlapping*, as each patient is assigned to a single cluster in a clustering. Our goal is to create a patient clustering, where the clusters correspond to clinically meaningful patient subpopulations.

To identify such subpopulations, we applied bisecting divisive hierarchical clustering. The algorithm iteratively constructs a hierarchy of clusters in a top-down (divisive) fashion, in each iteration bisecting a cluster into two new (child) clusters. A cluster is bisected using a *splitting variable*. One of the two child clusters contains all patients from the parent cluster for whom the splitting variable is true, and the other child contains all patients for whom the splitting variable evaluates to false. For example, if the parent cluster (cluster to split) consists of patients with hypertension (htn is true) and the splitting variable is ifg.pre2 (fasting plasma glucose FPG > 110), one of the child clusters is comprised of hypertensive patients having high FPG (ifg.pre2=true) and the other cluster is comprised of hypertensive patients with lower FPG (ifg.pre2 is false).

The algorithm proceeds by recursively bisecting each cluster into two child clusters starting with a cluster that represents the entire population. The algorithm terminates when no cluster can be bisected without having insufficient number of patients in the resultant child clusters; or when the patients in the cluster are sufficiently similar to each other.

The splitting variable is selected on the basis of how much variability in the diabetes outcome it can explain; bisections that explain a large amount of variability are preferred. Let t_j denote the follow-up time (in days) and δ_j the diabetes status at the end of follow-up for patient j. This patient is censored when the diabetes outcome is negative ($\delta_j = \text{false}$) at the end of follow-up. The martingale residual $M_j(t)$ for a patient j at time t_j is computed as the difference between the observed number $\delta_j(t)$ of event (1 if a patient j had developed diabetes before (or exactly at) time t_j, 0 if censored) and the estimated number $H_j(t)$ of events (cumulative hazard)

$$M_j(t) = \delta_j(t) - H_j(t).$$

To calculate the cumulative hazard, we use the Nelson-Aalen estimator,

$$H_j(t) = \sum_{t_i \le t} h_j(t_i) = \sum_{t_i \le t} \sum_k \frac{dN_k(t_i)}{Y_k(t_i)},$$

where $h_j(t_i)$ denotes the (non-cumulative) hazard of patient j at time t_i, k iterates over all patients, $dN_k(t_i)$ denotes the number of diabetes incidents that patient k suffers exactly at time t_i (0 or 1) and $Y_k(t_i)$ indicates whether patient k is at risk at time t_i. The formula for the non-cumulative hazard can be thought of as the number of events

occurring exactly at time t_i divided by the number of patient at risk at that time. When multiple patients suffer events at exactly the same time, these events are arbitrarily serialized.

Suppose we have a cluster C_1, which we need to bisect into clusters C_{11} and C_{12} using a particular splitting variable. Further, let SSR(C) denote the sum of squared martingale residuals for any cluster C. Bisecting C_1 will decrease the total SSR by

$$G = SSR(C_1) - [SSR(C_{11}) + SSR(C_{12})].$$

Each splitting variable produces a different G value. Among the possible splitting variables, we select the one that reduces the SSR the most, or equivalently, maximizes G. This is the splitting variable that explains the diabetes outcome in C_1 the best, thus it can be thought of as 'most relevant' to diabetes in the subpopulation corresponding to cluster C_1.

Once the cluster hierarchy has been constructed, the final clustering can be extracted. A **leaf cluster** is a cluster that is not bisected. Our hierarchical clustering algorithm ensures that each patient falls into exactly one leaf cluster, thus the collection of leaf clusters form a non-overlapping complete clustering of the patient population.

We wish to make two notes. First, our clustering algorithm is similar to the survival tree construction algorithm[19]; in fact, one can think about it as an adaptation of the recursive partitioning algorithm[20] for censored outcomes to a clustering application. Indeed, we follow Therneau et al.[21] in broad strokes and adapt their ANOVA criterion for censored outcome: we use sum squared martingale residuals instead of sum squared error. Second, the analogy between recursive partitioning and our algorithm goes deeper. The martingale residual can be rescaled into a deviance residual. Just as the sum squared error relates to the "deviance" of two nested Gaussian models, the sum squared deviance residuals relate to the deviance of two nested survival models, enabling the use of likelihood ratio tests for significance testing. Since our purpose is to construct the full hierarchy of clusters, we do not perform significance testing and use the martingale residuals instead of the deviance residuals.

Clustering and statistical analysis were conducted with the use of R version 3.0.1.

Results

In what follows, we demonstrate that our clustering algorithm successfully identified potentially interesting clusters that consist of patients with substantially higher or lower risk of diabetes than the general patient population. The clustering (the collection of leaf clusters) assigns each patient to exactly one leaf cluster and each leaf cluster possesses a cumulative hazard curve (specific to the subpopulation that the leaf consists of). Thus the clustering can be used to estimate a patient's risk of diabetes and can hence serve as a diabetes index. We will also demonstrate that our clustering used in this fashion outperforms the popular Framingham score. Thirdly, we will also show by constructing the entire hierarchy of clusters, that we can extract clusterings that encompass varying amounts of detail. Finally, we will demonstrate that our clustering can model non-proportional hazards as well as interactions among the risk factors.

Identifying high and low risk subpopulations

We performed hierarchical clustering of our patient population under the user-defined constraint that clusters with less than 50 patients are not bisected further. We identified 275 leaf clusters. From these leaves, we selected two: one indicative of very high risk and one indicative of very low risk. We then compared these leaves with the general population.

In Figure 1, we display the Kaplan-Meier survival curve and the cumulative hazard curve for the entire population (blue dotted line) and for the above two clusters: red solid line is used for the high-risk cluster and green dotdash line for the low-risk one. The patients (n = 61) in the high-risk cluster have fasting plasma glucose greater than 110 mg/dL (ifg.pre2), hyperlipidemia that requires therapeutic interventions (hyperlip.tx), and are current smokers (tobacco). The low-risk cluster consists of the patients (n = 498) characterized by fasting plasma glucose level equal to or lower than 100 mg/dL, no indication of hypertension, no indication of hyperlipidemia, no obesity, no renal disease, no congestive heart failure, no ischemic heart disease, no aspirin use, male non-smokers being between 45 and 65 years of age.

From Figure 1, the difference in diabetes progression among these three subpopulations is obvious and the risk factors for these patients are consistent with our clinical expectation.

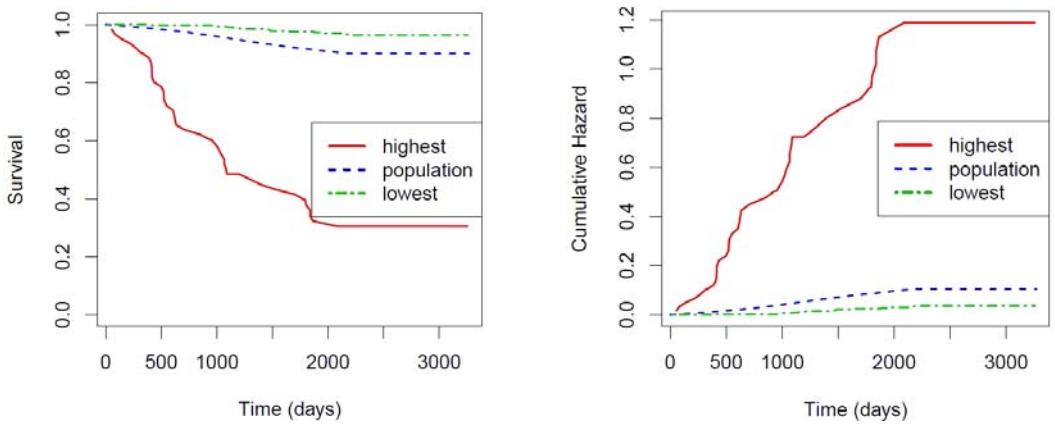


Figure 1. Kaplan-Meier survival curve (Left) and Cumulative incidence of diabetes (Right) for two pre-diabetic subpopulations (red sold and green dotdash) and the entire population (blue dotted)

Clustering as a diabetes index

As we described earlier, the leaf clusters form a non-overlapping clustering, where each patient belongs to exactly one leaf cluster. Since each leaf cluster contains a survival function of the corresponding subpopulation, it is possible to use the clustering as a diabetes index, providing a risk estimate for each patient. In this section, we compare the clustering as a diabetes index against the popular Framingham score, an actual diabetes index in clinical use. Specifically, we use concordance as our evaluation measure. Concordance is the probability that for any two patients where one progressed to diabetes earlier than the other, the one that progressed earlier has a higher predicted risk. The clustering achieved a concordance of 0.78, while the Framingham score achieved a lower concordance of 0.70, signifying the clustering has imporved discriminatory power.

Controlling the amount of detail

An important advantage of the proposed clustering technique over alternative methods, such as association rules, is that it offers fine control over the amount of detail it presents to the user. This control can be achieved by **cutting** the hierarchy at a particular level. To illustrate this point, we depict the entire cluster hierarchy in Figure 2. The leaf clusters are listed along the horizontal axis and the vertical axis indicates the SSR of the cluster. The hierarchy is represented by a dendrogram, which can be interpreted as follows. The root of the dendrogram is at the top (SSR=4645) and it represents a cluster that includes all patients. The root is split into two clusters (on ifg.pre2; not shown) one with SSR 730 (ifg.pre2=true) and one with SSR 3914 (ifg.pre2=false). The subpopulation having ifg.pre2=true is split on htn.pers into two clusters, one with SSR 41 (persistent hypertension present) and one with SSR 688 (htn.pers=false). In short, the dendrogram allows us to trace the bisections our algorithm performed and it also depicts the SSR of the resultant clusters.

We can cut the dendrogram at any SSR of our choice. Cutting the hierarchy produces a new set of leaf clusters, which in turn forms a non-overlapping complete clustering of the patient population. For example, if we cut the dendrogram at SSR of 4000 (which is very close to the top), we obtain only two leaf clusters: ifg.pre2=true with SSR of 730 and ifg.pre2=false with SSR of 3914. If we cut the dendrogram at a lower SSR, say at 500, we will obtain a larger set of leaf clusters (26 in this example) each having lower SSR. This particular cut is shown in Figure 2 as a magenta line. Larger number of leaf clusters offers a larger amount of detail. Selecting an SSR for cutting the dendrogram is what allows us to control the amount of detail (number of leaf clusters) in a predictable fashion. The SSR of the resulting leafs also shows the within-cluster similarity of the patients.

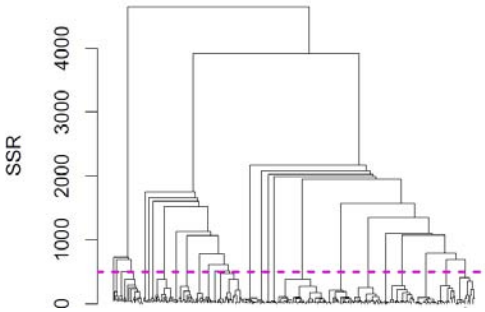


Figure 2. Dendrogram of the entire hierarchy of clusterings

Naturally, a tradeoff exists between the amount of detail and the predictive capability of a clustering. In Figure 3, we visualize this tradeoff. The horizontal axis represents the SSR at which the dendrogram was cut and the vertical axis represents the resultant clustering with the number of clusters depicted in left pane and the predictive capability (as measured by concordance) in the right pane. The figure shows that as we increase the SSR (move right on the horizontal axis), we decrease the amount of detail (number of clusters) and along with the decreased amount of detail, the predictive capability of the clustering decreases, as well.

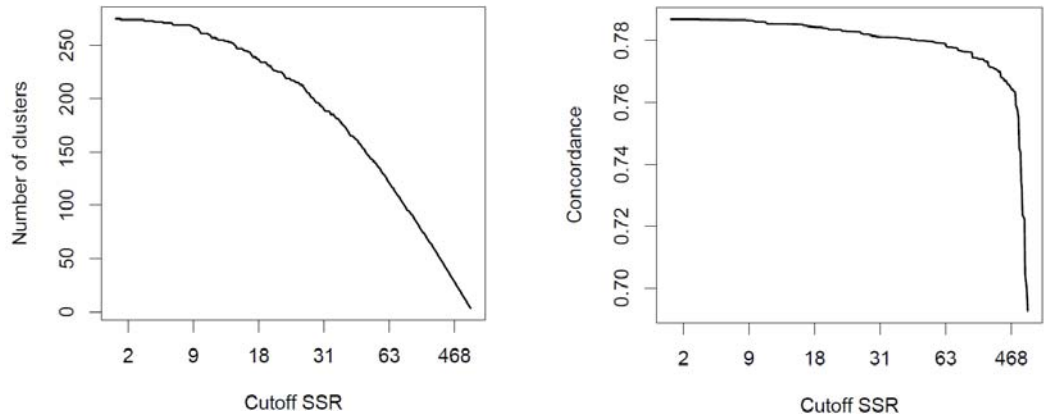


Figure 3. Tradeoff between the amount of detail (number of clusters) and the predictive capability (concordance)

Non-proportional hazard

We plot the cumulative hazard functions for the 26 clusters we extracted earlier (by cutting the hierarchy at SSR 500) in Figure 4. To avoid overcrowding the image and preserve good visibility of the lines, we separated the 26 clusters into four panes essentially at random. The IDs in the legend refer to their original IDs (IDs before cutting), thus they can exceed 26. Cumulative hazards across the clusters are not proportional because the LOGLOGS plots of the 26 clusters (shown in Figure 5) do not appear as parallel lines, indicating interactions between time and subpopulations. This non-proportionality was correctly captured by our approach. To show these clusters are clinically relevant, we selected the 12 highest risk clusters out of the 26 and described in Table 2.

Interactions among risk factors

Cluster 4 consists of patients who have ifg2.pre2 and htn.pers, and the estimate of the cumulative hazard at the end of the study is 1.02. To estimate the hazard under the assumption that the two risk factors are additive (act independently), we fit a Cox regression model to the entire population using only the above two variables as predictors. We used this Cox model to make a prediction for the subpopulation represented by cluster 4 and their cumulative hazard is 1.44. The difference between this prediction (of 1.44) and the prediction of 1.02 by the clustering strongly suggests that an interaction between these two risk factors (FPG and HTN) exists. While we do not know the exact risk of diabetes (true value for the cumulative hazard) in this subpopulation, it is between the observed prevalence of diabetes in this subpopulation, which is 0.57, and 1.0 (each patient can only experience at most one event of diabetes). 1.02 is closer to this range than 1.44, thus the additive Cox regression model overestimated the risk

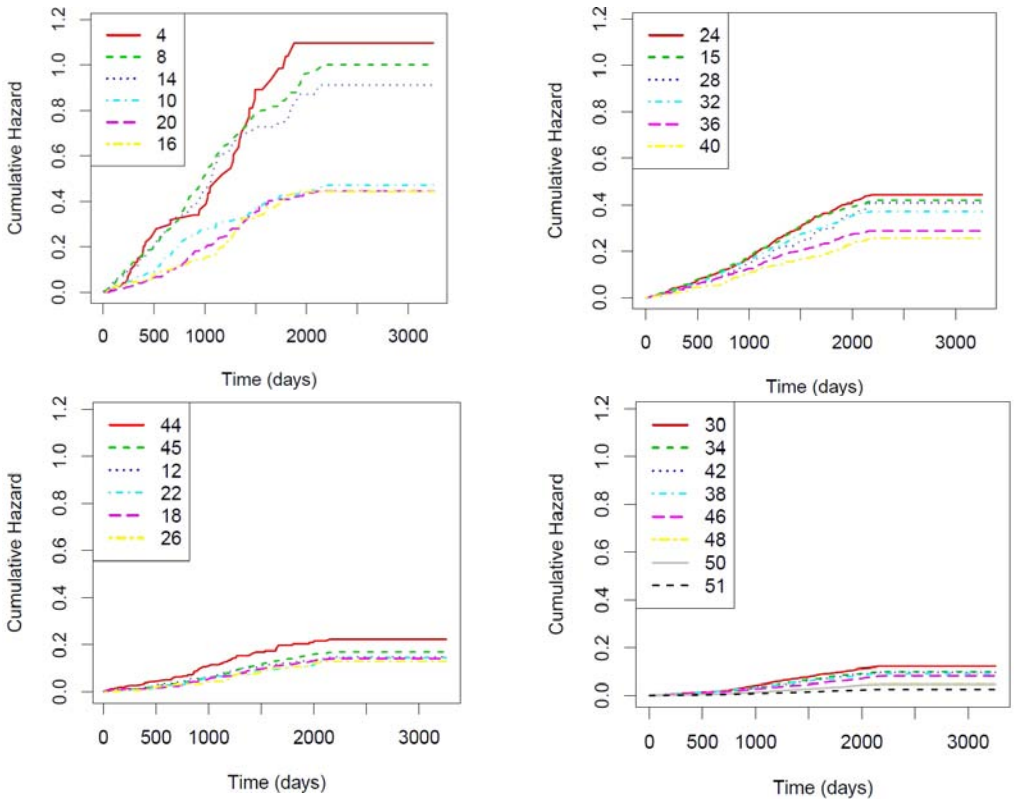


Figure 4. Identified pre-diabetic subpopulations based on cumulative hazard after infinite follow up time

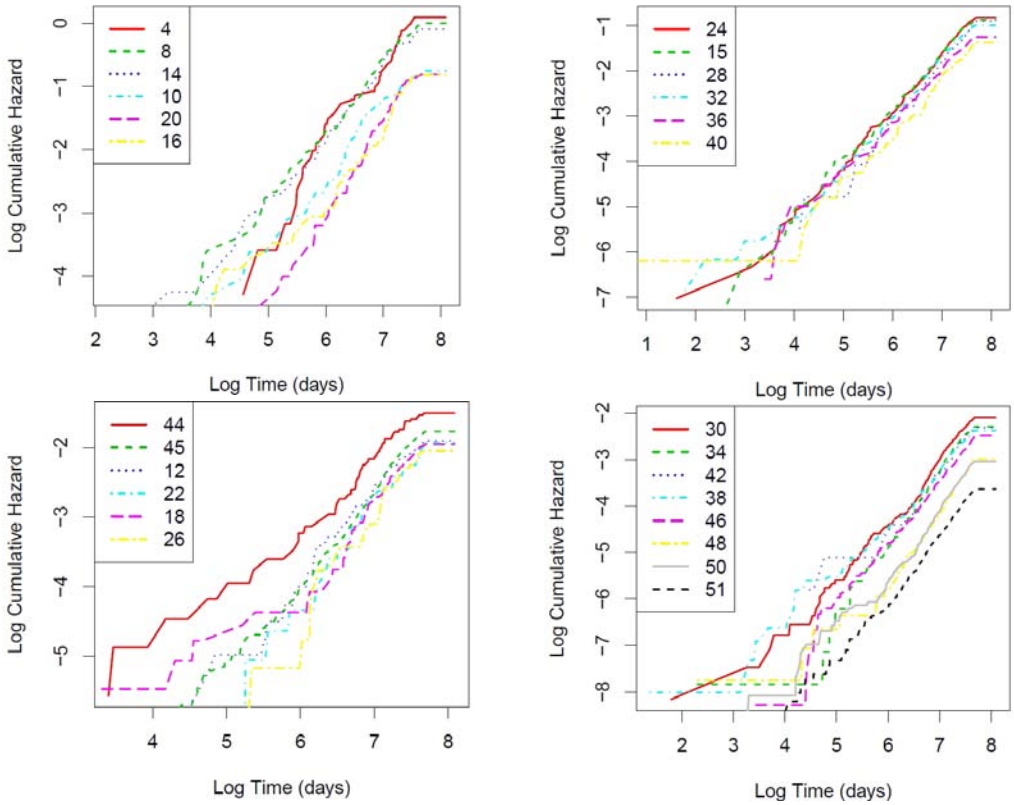


Figure 5. LOGLOGS plot of cummulative hazard

Table 2. Subpopulation summarization in terms of cumulative hazard at the end of the study.

Cluster ID	Patient Count	SSR	Cumulative hazard	Risk factors
4	74	40	1.02	ifg.pre2=true, htn.pers=true
8	297	180	1.00	ifg.pre2=true, htn.pers=false, obese=true
14	212	121	0.91	ifg.pre2=true, htn.pers=false, obese=false, hyperlip.tx=true
10	227	81	0.47	ifg.pre2=false, ifg.pre1=true, hyperlip.multi=true
20	280	88	0.45	ifg.pre2=false, ifg.pre1=true, hyperlip.multi=false, renal=false, htn.pers=true
12	736	88	0.15	ifg.pre2=false, ifg.pre1=false, htn.pers=true
24	1130	380	0.44	ifg.pre2=false, ifg.pre1=true, hyperlip.multi=false, renal=false, htn.pers=false, hyperlip.tx=true
15	1276	384	0.42	ifg.pre2=true, htn.pers=false, obese=false, hyperlip.tx=false
28	241	68	0.41	ifg.pre2=false, ifg.pre1=true, hyperlip.multi=false, renal=false, htn.pers=false, hyperlip.tx=false, ihd=true
32	949	277	0.37	ifg.pre2=false, ifg.pre1=true, hyperlip.multi=false, renal=false, htn.pers=false, hyperlip.tx=false, ihd=false, obese=true
36	735	166	0.28	ifg.pre2=false, ifg.pre1=true, hyperlip.multi=false, renal=false, htn.pers=false, hyperlip.tx=false, ihd=false, obese=false, htn.tx=true
40	493	102	0.25	ifg.pre2=false, ifg.pre1=true, hyperlip.multi=false, renal=false, htn.pers=false, hyperlip.tx=false, ihd=false, obese=false, htn.tx=false, aspirin=true

Discussion

In this paper, we presented a novel bisecting divisive hierarchical clustering algorithm to identify clinically relevant patient subpopulations using type 2 diabetes as the endpoint. In a good clustering, patients within the same cluster are more similar to each other than to patients in a different cluster. Patients in our clusters are similar to each other because they share the same risk factors that are most relevant to diabetes and also have similar risk of developing diabetes. We have shown that our clustering can be used as a diabetes index: when the clustering is sufficiently detailed, it outperformed the Framingham score in terms of concordance (ability to distinguish high-risk patients from low-risk patients). While ARM models have also shown excellent predictive performance, their high level of redundancy leads to unnecessary computational cost. In the following discussion, we examine the beneficial properties of the clustering particularly compared to ARM models and the potential of overfitting.

Comparison to Association Rule Mining

Recent developments of ARM[22], including survival association rule mining[9] have demonstrated its applicability in the EHR mining domain and its appropriateness to serve as a diabetes index. The key advantage of the ARM methodology lies in its interpretability: individual rules are straightforward to interpret and the interpretation provides a context around the risk estimate (e.g. the high risk is due to persistent hypertension and severe hyperlipidemia). As previously discussed, the disadvantage of ARM is that it generates an exponentially large, redundant rule set. With many rules applying to the same patient, making prediction for an individual becomes non-trivial[7,9], making a direct comparison between ARM and clustering leave room for arguments. Just to show that the predictive performance of ARM and clustering are similar, we performed a simple, albeit admitted imperfect, comparison. We largely followed the methodology outlined in the studies using ARM to assess the risk of type 2 diabetes[7,14]: we built a Cox model with age and gender as the predictors, and extracted distributional association rules[7,15] indicating an association between the martingale residual and the major risk factors (IHD, hypertension and hyperlipidemia as defined in Table 1 that covered at least 50 patients (same coverage as used for clustering). For each patient, we made a prediction using the most specific rule. The concordance of the resultant model was .7601 with a standard error of 004. This is comparable to the performance of the clustering model. Additionally, both the ARM-based models and our clustering have the ability to automatically discover interactions among risk factors and seamlessly incorporate them into the model or clustering.

Our proposed method goes beyond the state of the art by allowing the user to control the amount of details the clustering should incorporate. This is particularly beneficial, because the amount of detail can be adjusted to the needs of the consumer of the model. For example, when the user of the clustering is an automated clinical decision support system, a highly detailed clustering may be desirable. Computational systems can handle complex models, even as complex as the ARM models, and thus a clustering that incorporates a large amount of details (without overfitting) can be most appropriate.

Another potential use of the clustering produced by our method concerns clinical investigation, where clinicians, rather than computers, view the clustering results. Presenting excessively detailed complex models to investigators can be more distracting than useful, thus a moderately complex clustering may be most desirable. Our method constructs the entire cluster hierarchy upfront allowing investigators to drill down for further details. This can be achieved through further clustering a specific subpopulation (leaf), as needed.

Overfitting

Models as flexible as the clustering-based model or the association rule set are susceptible to overfitting the data. In this application, we were not particularly concerned with the predictive performance of the model as it is secondary to its interpretability. To avoid overfitting, we required the presence of at least 50 patients in each node (or association rule), which is sufficient to reliably estimate their risk. Also, Figure 3 shows no sign of overfitting: increased number of nodes have consistently led to improved performance on a validation set. Nonetheless, when the clustering is used as a predictive modeling tool, the number 50 needs to be tuned more carefully and attention must be paid to the potential overfitting.

In summary, we have demonstrated that our clustering method retains the benefits of existing diabetes risk models and adds its own advantage through allowing for fine control of detail that is presented to the user. This promises great potential of contributing to clinical practice.

References

1. Centers for Disease Control and Prevention. National diabetes fact sheet: national estimates and general information on diabetes and prediabetes in the United States, 2011. 2011.

2. Centers for Disease Control and Prevention. Diabetes Report Card 2012: National and State Profile of Diabetes and Its Complications. 2012.

3. Lindström J, Peltonen M, Eriksson JG, et al. Improved lifestyle and decreased diabetes risk over 13 years: long-term follow-up of the randomised Finnish Diabetes Prevention Study (DPS). *Diabetologia*. 2013;56(2):284-93. doi:10.1007/s00125-012-2752-5.

4. Knowler WC, Barrett-Connor E, Fowler SE, et al. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *N Engl J Med*. 2002;346(6):393-403. doi:10.1056/NEJMoa012512.

5. Collins GS, Mallett S, Omar O, Yu L-M. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC Med*. 2011;9(1):103. doi:10.1186/1741-7015-9-103.

6. Wilson PWF, Meigs JB, Sullivan L, Fox CS, Nathan DM, D'Agostino RB. Prediction of incident diabetes mellitus in middle-aged adults: the Framingham Offspring Study. *Arch Intern Med*. 2007;167(10):1068-74. doi:10.1001/archinte.167.10.1068.

7. Simon GJ, Caraballo PJ, Therneau TM, Cha SS, Castro MR, Li PW. Extending Association Rule Summarization Techniques to Assess Risk of Diabetes Mellitus. *Knowl Data Eng IEEE Trans*. 2013;PP(99):1-13. doi:10.1109/TKDE.2013.76.

8. Kim HS, Shin AM, Kim MK, Kim YN. Comorbidity study on type 2 diabetes mellitus using data mining. *Korean J Intern Med*. 2012;27(2):197-202. doi:10.3904/kjim.2012.27.2.197.

9. Simon GJ, Schrom J, Castro MR, Li PW, Caraballo PJ. Survival association rule mining towards type 2 diabetes risk assessment. *AMIA Annu Symp Proc*. 2013;2013:1293-302.

10. Schrom JR, Caraballo PJ, Castro MR, Simon GJ. Quantifying the Effect of Statin Use in Pre-Diabetic Phenotypes Discovered Through Association Rule Mining.

11. Simon GJ, Kumar V, Li PW. A simple statistical model and association rule filtering for classification. *Proc 17th ACM SIGKDD Int Conf Knowl Discov data Min - KDD '11*. 2011:823. doi:10.1145/2020408.2020550.

12. Agrawal R, Srikant R, others. Fast algorithms for mining association rules. In: *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*.Vol 1215.; 1994:487-499.

13. Liu G, Feng M, Wang Y, et al. Towards exploratory hypothesis testing and analysis. *2011 IEEE 27th Int Conf Data Eng*. 2011:745-756. doi:10.1109/ICDE.2011.5767907.

14. Caraballo PJ, Castro MR, Cha SS, Li PW, Simon GJ. Use of Association Rule Mining to Assess Diabetes Risk in Patients with Impaired Fasting Glucose. *AMIA Annu Symp Proc*. 2011.

15. Simon GJ, Li PW, Jack CR, Vemuri P. Understanding atrophy trajectories in alzheimer's disease using association rules on MRI images. In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '11*. New York, New York, USA: ACM Press; 2011:369. doi:10.1145/2020408.2020469.

16. Tan P-N, Steinbach M, Kumar V. Cluster Analysis: Basic Concepts and Algorithms. In: *Introduction to Data Mining*. Addison-Wesley; 2005.

17. Rocca WA, Yawn BP, St Sauver JL, Grossardt BR, Melton LJ. History of the Rochester Epidemiology Project: half a century of medical records linkage in a US population. *Mayo Clin Proc*. 2012;87(12):1202-13. doi:10.1016/j.mayocp.2012.08.012.

18. American Diabetes Association. Executive summary: standards of medical care in diabetes—2014. *Diabetes Care*. 2014;37 Suppl 1(January):S5-13. doi:10.2337/dc14-S005.

19. Davis RB, Anderson JR. Exponential survival trees. *Stat Med*. 1989;8(8):947-61.

20. Olshen LBJHFRA, Stone CJ. Classification and regression trees. *Wadsworth Int Gr*. 1984.

21. Therneau TM, Atkinson EJ. An introduction to recursive partitioning using the RPART routines. 1997.

22. Srikant R, Agrawal R. Mining generalized association rules. *Futur Gener Comput Syst*. 1997;13(2-3):161-180. doi:10.1016/S0167-739X(97)00019-8.

Extracting Patient Demographics and Personal Medical Information from Online Health Forums

Yang Liu, B.Eng.[1], Songhua Xu*, Ph.D.[1], Hong-Jun Yoon, Ph.D.[2], Georgia Tourassi, Ph.D.[2]
[1]: Department of Information Systems, College of Computing Sciences
New Jersey Institute of Technology
University Heights, Newark, NJ, 07102, USA
[2]: Biomedical Science and Engineering Center, Health Data Sciences Institute
Oak Ridge National Laboratory
One Bethel Valley Road, Oak Ridge, Tennessee, USA, 37830

Abstract

Natural language processing has been successfully leveraged to extract patient information from unstructured clinical text. However the majority of the existing work targets at obtaining a specific category of clinical information through individual efforts. In the midst of the Health 2.0 wave, online health forums increasingly host abundant and diverse health-related information regarding the demographics and medical information of patients who are either actively participating in or passively reported at these forums. The potential categories of such information span a wide spectrum, whose extraction requires a systematic and comprehensive approach beyond the traditional isolated efforts that specialize in harvesting information of single categories. In this paper, we develop a new integrated biomedical NLP pipeline that automatically extracts a comprehensive set of patient demographics and medical information from online health forums. The pipeline can be adopted to construct structured personal health profiles from unstructured user-contributed content on eHealth social media sites. This paper describes key aspects of the pipeline as well as reports experimental results that show the system's satisfactory performance in accomplishing a series of NLP tasks of extracting patient information from online health forums.

Introduction

Natural language processing (NLP) has been widely leveraged in the biomedical domain in recent years. It has been shown effective in extracting biomedical information from free-structured text data [1, 2, 3]. Many applications have been developed to enhance the performance of clinical information retrieval. The extracted information is sometimes restructured or encoded in a special machine-friendly format for automatic computer processing. Most biomedical NLP applications focus on specific areas and thus extract information from text data in single clinical application areas such as radiology reports [4], discharge summaries [5], medication instructions [6, 7] and nursing narratives [8]. Typical input to these text extraction tasks is formal medical reports that contain precise clinical information about patients. However these reports do not comprehensively reveal the overall health conditions of a patient because of their narrow focus, definitive purpose, and limited scope of information collection aimed at by a single medical report in a specific area. To better understand patient conditions with more depth and breadth, we need to extract health-related information of patients more richly and diversely. A good class of information sources useful for this purpose is the increasingly emerging and popular online health forums, which contain posts in large numbers written by a variety of patients, caregivers, and supporters. Many of these posts voluntarily offer, through narrative text, a board range of information regarding the personal background and medical conditions of patients, including patients' demographic and health-related information, such as patients' living habits, working circumstances, and family history. Such an enriched body of information is typically not included in formal clinical reports. Having access to the personal medical information through the non-traditional communication channel can be a valuable additional resource for clinical and epidemiological research. By extracting the comprehensive scope of patients' health-related information from the relevant posts on online health forums, this study exploits a new opportunity to access patient information that differs from traditional information extraction approaches relying on formal and specialized clinical reports.

*Correspondence can be addressed to S. Xu through songhua dot xu at njit dot edu.

It is noted that many existing studies have attempted to extract information from online forums. For example, Ritter and Mausam propose a method to extract open domain event information from Twitter [9] through topic modeling. Jung propose a name entity recognition method for microtexts in social networks such as Twitter [10]. In [11] Sondhi and Gupta use support vector machine and conditional random field to classify sentences in health forums into two medical categories, including physical examination/symptoms and medications. However they neither extract detailed attributes from those sentences nor attempt to structure or restructure the extracted information. In this paper we propose an algorithmic pipeline that automatically extracts a comprehensive set of patient demographic and health-related information about their posts on online health forums. The goal is to construct structured patient health profiles from unstructured user-contributed content on eHealth social media sites. The derived structured patient health profiles are encoded in the XML format, containing structured and categorized patient information that is easy for computer to read, process, retrieve, and analyze. Such output format can also be easily transformed into different representation models and formats for automated consumption by various third-party systems, e.g. Entity-Relationship (ER) and UML models.

Methods

System Framework

The proposed system comprises a sequence of modules, each responsible for a type of NLP function, which collectively composes our overall application pipeline. The initial input to the pipeline is a corpus of webpages downloaded from a collection of online health forums. The system output is a set of structured patient health profiles. Each module transforms its input data according to its functional role and then subsequently passes the intermediate output to its following module until the final form of the target structured patient profile is generated. The first module is a preprocessor, which extracts the text content of all posts written by or about a user and then merges the result into one user profile. This module also detects and segments the post text into individual sentences for further processing. The second module is a classifier. It assigns each sentence a set of class labels that specify the particular category of patient information delivered by the sentence. The next module is a parser, which utilizes multiple NLP tools to extract health-related attributes from each category of sentences. The last module is an encoder. It generates the final form of the profile of patient information in the XML format. Figure 1 illustrates the overall framework of the proposed system pipeline.

Preprocessing

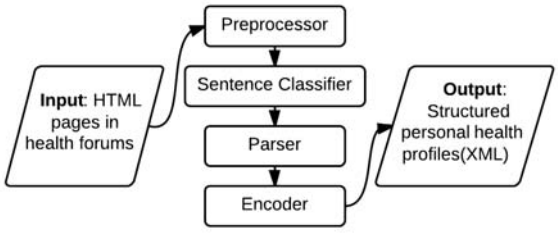


Figure 1: The overview framework of our system pipeline.

The raw data downloaded from the Internet is a corpus of HTML webpages collected from a range of online health forums. These raw HTML files cannot be directly fed into the text mining modules of our pipeline since they carry lots of text irrelevant to our text mining purpose such as HTML tags and Javascripts codes. To address this problem, we develop an HTML parser to filter out such irrelevant content from the raw webpage downloads. The remaining text is retained for the subsequent analysis. The parser has different parsing rules according to different HTML patterns exhibited by various health forums. To identify and gather all posts written by or about a specific patient in the cyberspace, we perform link analysis on each online forum. All analysis results derived for the same user are then merged into a single file for the person. When identifying a user, we also use the HTML parser to detect the user's name according to different tag patterns exhibited in HTML pages downloaded from different forums. For example, posts on the American Cancer Society forum express a user's name through the following string pattern: " .* ", where ".*" means a character string of arbitrary letters and of any length. If we encounter a string in the pattern of " mxperry ", we can parse this string and extract the user name as "mxperry". Each forum adopts its own source code pattern. We respectively explore each site's particular encoding pattern for string parsing when extracting user names. For the tracking purpose, each user is assigned with a unique patient ID number. As mentioned earlier, all detected content corresponding to a common user is then consolidated into a single text file for the user, which we call the user's *online content file*. For each resultant online content file corresponding to a specific user, we then apply a sentence splitter to segment the consolidated post content into its constituent

Class	Description	Subclass	Example
DMO	Demographic information	AGE (age)	11 years old
		GEN (gender)	male, female
		ETH (ethnic)	Asian, Hispanic
BVR	Patient's behavior	WOK (work)	I was a teacher at 2009
		LIF (living habit or behavior)	I have been smoking for 8 years
		STU (study)	I studied for my master degree
		CLI (clinical response)	My back pain relieved after the therapy
FMA	Family information		My father died of cancer
SPM	Symptom		chest pain, cough
ETR	Exam/Test and Results	TES (test)	chest x-ray
		RES (result)	Blood pressure is normal
DIG	Diagnosis		I was diagnosed with stage IV lung cancer
PRO	Medical Procedure /Treatment		surgery, chemotherapy
MED	Medication		I took Lisinopril one time a day
PHY	Physical State		tired, sleepy
PSY	Psychological State		sad, stressed
NOC	Not belonging to any class		

Table 1: The list of sentence classes and subclasses on eHealth forums supported in this work.

sentences. Overall, the preprocessing module transforms the original unorganized collection of webpages downloaded from the Internet to a corpus of personal information files. Each online content file of a user contains the user's patient ID, name or nickname, followed by a list of sentences written by or about the user on online health forums.

Extracting and Classifying Sentences Containing Patient Demographic and Health-Related Information

Our classification procedure regards each sentence in a user's online content file as the basic processing element. By classifying every such sentence, the method offers a finer granularity for information access and processing. For example, different researchers may be interested in different aspects of a patient's information for knowledge discovery and pattern mining. Our method recognizes 11 basic classes of sentences for those appearing on online health forums. Some classes also have sub-classes. The 11 classes are selected based on the medical concepts from the Medical Entity Dictionary (MED) [12], Unified Medical Language System (UMLS) [13], and some extension recommended by the domain experts we consulted with for the project. These resources collectively cover nearly all categories of health-related information supplied by online health forums that are encountered in this study. Table 1 lists all supported classes and sub-classes in this study.

Every sentence in the training set is manually labeled with one or multiple applicable class labels and their corresponding subclass labels, if the latter type of finer-level sentence classification is available. Each sentence may have multiple labels since these sentence classes are not necessarily mutually exclusive, given that a sentence may convey one or more information classes simultaneously. We explore a number of text classification methods such as Naive Bayesian, Bayesian Net, Adaboost, K-Nearest Neighbor, Support Vector Machine (SVM), Sequential Minimal Optimization (SMO), Logistic Regression, Neutral Network, and Decision Tree. Multiple types of text features such as TF-IDF, unigram, bigram, trigram, and words sequence are extracted for input to each candidate classification method. Our implementation chooses the Naive Bayesian classifier as its multi-class sentence classifier in the end because of its robustness and efficiency for short text classification according to empirical evidence. Note that the sequence of sentences may bear an important impact on sentence labeling outcome. For example, a sentence that states a patient's medical test results (RES) is very likely to follow a sentence that states the patient's medical test (TES). According to this observation, we utilize the Hidden Markov Model (HMM) to boost our basic Naive Bayesian classifier. Our implementation utilizes the HMM boosting procedure proposed in [14]. The states in the HMM model correspond to the 11 classes of sentences recognized in our method. Experimental results show that our Naive Bayesian classifier boosted by HMM outperforms other candidate classifiers explored. Using the multi-label classifier mentioned in the above, the method can automatically assign one or multiple class labels for each sentence in the patient's online content file. Those sentences labeled with "NOC," which stands for "not belonging to any class," are filtered out.

Parsing Sentences using NLP Tools to Extract Detailed Patient Information

After we extract sentences that contain a patient's demographic and health-related information using the Naive Bayes classifier, the method can further extract detailed attributes of patient information from each sentence. For sentences of different classes, different natural language processing toolkits are adopted for the parsing task. The patient information extraction result is formatted following the inline XML schema.

Extracting Patient Demographic Information

We use and extend DIVER, a tool developed by Hsieh and Doan [15] that identifies and standardizes demographic variables, to extract patient demographic information from sentences belonging to the class of Demographic information (DMO). In particular, we extract three types of attributes: age, ethnic, and gender. An example is as follows:

Input: <DMO:AGE>I am 35 years old now.</DMO:AGE>
Output: <DMO:AGE>I am <Age>35 years old</Age>now .</DMO:AGE>

Extracting Heath-Related Patient Information

We use MedLEE [1], a medical natural language processing system, to parse sentences that belong to the classes of SPM, ETR, DIG, and PRO. MedLEE has been proved to yield satisfactory performance in processing free-structured clinical text, such as X-ray reports, discharge summaries, sign-out notes, and nursing narratives [2, 8]. MedLEE covers a broad range of concepts and semantic rules in the medical domain such as medication, procedure, body measurement, and laboratory test. The output of MedLEE is represented using a frame-based format. Each frame contains the frame type, frame value, and several modifier slots, which are also represented following a frame-based format. We show an example below that demonstrates how MedLEE structures a sentence that belongs to the symptom class.

Input: severe pain in back
Output: [problem, pain, [degree, severe, [bodyloc, back]]

We can easily transform the frame-based output of MedLEE into its corresponding XML format by extracting each value of the frame as an attribute with the type of the frame used as the tag name. An example is shown below on the transformation process.

Input: <SYP>I suffered from a severe pain in my back.</SYP>
Output: <SYP>I suffered from a <Degree>severe</Degree> <Problem>pain</Problem> in my <Bodyloc>back</Bodyloc>.</SYP>

As medical information is often related to time, organization, and location, we also need to extract those attributes. For such purpose, we adopt the Stanford Core NLP toolkit [16]. It contains a named entity recognition tool that can tag seven types of entities, including Time, Location, Organization, Person, Money, Percent, and Date. Leveraging this tool, we can extract and tag the time, date, organization, and location attributes from a sentence. Below is a parsing example on a sentence that discusses a medical procedure:

Input: <PRO>I received chemotherapy at CTCA in Oklahoma in 2009.</PRO>
Output: <PRO>I received <Procedure>chemotherapy</Procedure> at <Organization>CTCA</Organization> in <Location>Oklahoma </Location> in <Date>2009</Date>.</PRO>

Extracting Medication Information

For sentences that discuss medication information, we use the MedEX, a medication NLP tool introduced in [6], to extract concrete attributes involved, such as drugName, strength, route, frequency, form, dose amount, intake time, duration, and dispense amount. MedEx can extract more detailed attributes than MedLEE in the above area. Below is an example:

Input: <MED>I was discharged on Lopressor 50 mg, take 1 Tablet by mouth two times a day.</MED>
Output: <MED>I was discharged on <DrugName>Lopressor</DrugName> <Strength>50 mg</Strength>,

take <Dose>1 Tablet</Dose> <Route>by mouth</Route> <Frequency>two times a day<Frequency>.
</MED>

Extracting Patient Behavior

For sentences that describe patients' behaviors, physical and psychological states, as well as family information, the involved information often does not fall into any single domain. To extract those attributes, we need a general-purpose parsing tool. In this subsection, we are concerned with the method for extracting patient behaviors. Our system implementation utilizes the part-of-speech parser provided by the Stanford Core NLP toolkit, which can generate a parse tree for each input sentence. This parser has a good reputation in processing text in the biomedical domain [17]. An example of the generated parse tree is as follows:

Input: <BHV:LIF>I smoked very often in the last 10 years</BHV:LIF>.
Parse Tree: (ROOT (S (NP (PRP I)) (VP (VBD smoked) (ADVP (RB very) (RB often)) (PP (IN in) (NP (DT the) (JJ last) (CD 10) (NNS years))))))

We first identify the subject of the sentence by finding the noun phrase (NP) that appears immediately before the verb phrase (VP), where both phrases are in a simple declarative clause (S), which must be the child node of the sentence root. In this example, we identify "I" as the subject following the above heuristic. For sentences that describe patient behaviors, our approach assumes that the verb phrase in a predicate usually contains the direct description of patient behavior. Such verb phrase usually locates at the right sibling of the sentence subject. After identifying the verb phrase, we then filter out the descriptive words in it, including the adverb phrase (ADVP) and preposition phrase (PP), so that only verbs would remain. We then tag each remaining verb as a patient behavior attribute. We also use the Stanford named entity parser mentioned in the previous section to label the time, date, organization, and location appearing in an input sentence. An example output is shown below for the same input displayed in the above:

Output: <BHV:LIF>I <Behavior>smoked</Behavior> very often in the <Date>last 10 years</Date>
</BHV:LIF>.

Extracting Patient Physical and Psychological States

To extract patient physical and psychological states, we use the same parse tree method mentioned in the above. The difference is that this time the main focus is on extracting adjective phrases in an input sentence instead of verb phrases because adjective phrases are generally more descriptive of patients' physical and psychological states. In particular, by identifying the adjective words (JJ) in adjective phrases (ADJP), we can extract attributes regarding patient states. An example is as follows:

Input: <PYS>I feel scared and stressed in the first two weeks.</PYS>
Parse Tree: (ROOT (S (NP (PRP I)) (VP (VBP feel) (ADJP (JJ scared) (CC and) (JJ stressed)) (PP (IN in) (NP (DT the) (JJ first) (CD two) (NNS weeks))))))
Output: <PSY>I feel <Psy-state>scared</Psy-state> and <Psy-state>stressed</Psy-state> in the first two weeks.</PSY>

Extracting Patient Family Information

For those sentences that contain family information (FMA) such as family medical history, we extract any mentioning of family member(s) in the sentences. We first use the parse tree to find the subject of an input sentence. We then search the appearance of any subject word according to a vocabulary list regarding different roles of family members according to English Grammar Online [18]. The list contains 45 words. If the subject word appears in the list, we then label it using the tag "Member." An example is as follows:

Input: <FMA>My father died of lung cancer at the age of 83. </FMA>
Output: <FMA>My <Member>father </Member> died of lung cancer at the age of 83. </FMA>

Results

Data

We downloaded a total of 11274 webpages from a number of US health forums and hospital association websites, such as American Cancer Society's Cancer Survivors Network, eHealth Forum, PatientsLikeMe, etc. After the pre-processing stage, we constructed 3196 patients' personal online content files using the pipeline discussed earlier. After segmenting these content files into individual sentences, we obtained 9730 unique sentences as our full experimental data set. We then invited 4 researchers in the domain of medical informatics to manually assign sentence class labels for all these sentences with the help of MED and UMLS. Table 2 shows the 10 online health forums where we collect our data and their size of patient population respectively.

Forum Name	URL	Population
American Cancer Society	http://www.cancer.org	401
eHealth Forum	http://ehealthforum.com	319
PatientsLikeMe	http://www.patientslikeme.com	222
HealingWell.com	http://www.healingwell.com	335
Seattle Cancer Care Alliance	http://www.seattlecca.org	380
Gibis Cancer Center	http://www.gibbscancercenter.com	303
Team Draft	http://www.teamdraft.org	346
American Lung Association	http://www.lung.org	298
HuffPost Impact	http://www.huffingtonpost.com/impact	323
APIAHF	http://www.apiahf.org	296

Table 2: Ten online forums and their respective URLs and user populations.

Sentence Classification

To explore the effectiveness of the new pipeline, we tested the overall performance of different classifiers mentioned in Section 3. We use Weka [19], a widely used machine learning software, to run our experiments. From the whole dataset we randomly chose 7000 sentences as the training data and the rest 2730 sentences as the testing data. We adopted the cross-validation schema to benchmark the learning performance. After we train our sentence classifier through the training dataset, we measured its performance on the testing dataset. We performed ten-fold cross validation to measure the precision, recall, and F-rate of the method.

Table 3 shows that in terms of the sentence classification performance, the Naive Bayesian classifier boosted by HMM outperforms all other classifiers. Although the rest of the classifiers may be more effective for tackling certain text classification tasks, for the particular sentence classification task concerned in this study, we adopted Naive Bayes as the optimal learning device for our implementation. Table 4 shows that the sentence classification module implemented in our pipeline yields generally satisfactory performance in terms of its precision, recall, and F-rate, especially in the medically related areas. In such areas, text tends to be written using more medical terms and formal language than those in the non-medical areas. This particular language characteristic may affect the choice of the optimal text classifier. Figure 2 shows the percentages of sentences in each class. According to the statistics, each sentence class has enough samples for training the corresponding classifier.

Classifier	Precision (%)	Recall (%)	F-rate (%)	Classifier	Precision (%)	Recall (%)	F-rate (%)
NB-HMM	93.2	90.9	91.7	SVM	82.2	70.3	75.8
NB	91.3	89.6	90.4	NN	76.3	74.5	75.4
BN	90.8	87.6	89.1	SMO	77.9	70.8	73.8
AD	86.6	83.2	80.1	DT	87.5	81.1	84.2
KNN	82.7	68.9	75.2				

Table 3: Performance comparision regarding the 9 sentence classifiers: Naive Bayes boosted by HMM (NB-HMM), Naive Bayes (NB), Bayes Network (BN), Adaboost (AD), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Neutral Network (NN), Sequential Minimal Optimization (SMO), Logistic, and Decision Tree (DT).

Sentence Class	Precision (%)	Recall (%)	F-rate (%)	Sentence Class	Precision (%)	Recall (%)	F-rate (%)
DMO	90.6	89.6	90.1	SPM	85.7	84.3	85.0
DMO:AGE	89.6	87.2	88.3	ETR	91.1	91.5	91.3
DMO:GEN	93.3	90.6	91.9	ETR:TES	91.1	89.8	90.4
DMO:ETH	87.0	85.6	86.3	ETR:RES	91.3	91.1	91.1
BHV	81.7	78.0	79.8	DIG	93.2	90.2	91.6
BHV:WOK	78.8	75.7	77.2	PRO	82.9	79.3	81.1
BHV:LIF	78.1	74.3	76.2	MED	95.5	91.1	93.2
BHV:STU	82.5	80.0	81.2	PHY	80.3	78.8	79.5
BHV:CLI	85.3	81.2	83.1	PSY	89.1	85.4	87.2
FMA	88.6	85.8	87.2	NOC	80.2	86.5	83.2

Table 4: Performance of different sentence classifiers.

Class	Precision(%)	Recall(%)	F-rate(%)	Class	Precision(%)	Recall(%)	F-rate(%)
DMO	91.3	86.3	88.7	SPM	95.7	87.3	91.3
DMO:AGE	87.2	83.3	85.2	ETR	91.0	88.5	89.7
DMO:GEN	97.6	96.1	96.8	ETR:TES	89.1	86.8	87.9
DMO:ETH	90.6	87.6	89.0	ETR:RES	91.9	88.2	90.0
BHV	81.7	78.0	79.8	DIG	96.2	91.3	93.6
BHV:WOK	88.8	79.7	84.0	PRO	90.9	85.4	88.0
BHV:LIF	78.0	76.3	77.1	MED	94.5	92.1	93.1
BHV:STU	81.5	78.1	79.8	PHY	83.1	79.9	81.4
BHV:CLI	89.3	86.2	87.7	PSY	88.7	84.2	86.3
FMA	96.6	95.8	96.2				

Table 5: Performance of information extraction in different sentence classes.

Extracting Detailed Patient Information

For the 9730 sentences analyzed in our experiments, those labeled with "NOC" are first excluded. We then utilize a set of NLP tools integrated in our prototype system implementation to extract attributes containing detailed patient information for different classes of sentences respectively. For the evaluation and benchmarking purpose, the information extraction accuracy is manually verified by four domain experts according to MED, UMLS, and some extensions made by them based on the sentence classes not related to MED and UMLS directly. Since the demographic information usually requires more precise extraction, we further report the information extraction performance for each demographic attribute. To further demonstrate the effectiveness of the information extraction module, we compare its performance with that of another peer concept extraction tool—BioTagger-GM [20]. The peer tool uses machine learning techniques to train semantic taggers for extracting medical concepts from clinical documents. We apply both our method and the peer method to extract ten classes of information concerned in this study, i.e. all sentence classes except for the "NOC" class. The results yielded by both methods are then compared. Figure 3 shows that our proposed information extraction pipeline, which collaboratively leverages multiple NLP tools, consistently outperforms the peer method.

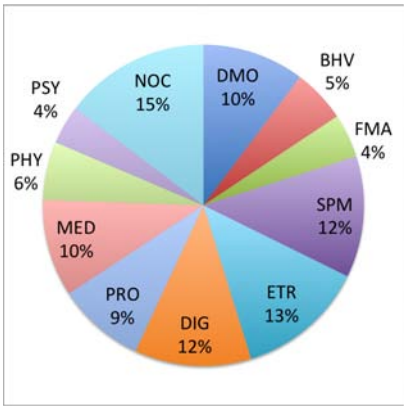
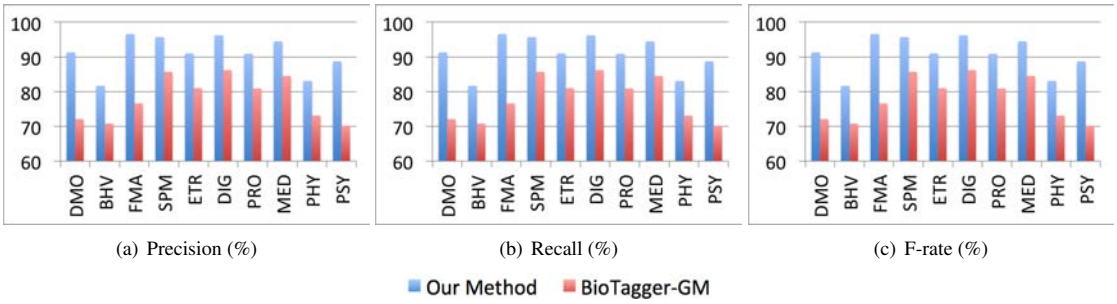


Figure 2: Percentage of sentences in each class.

Table 5 shows that the information extraction module of our implemented pipeline performs satisfactorily for all sentence classes. Some categories received relatively low F-rates such as life behavior. A potential reason was due to the more complex language and ambiguous vocabularies often used for conveying information of these categories.



(a) Precision (%) (b) Recall (%) (c) F-rate (%)

Figure 3: Comparison of information extraction performance between our proposed method and the peer method.

Coverage(%) Class / Forum	DMO	BHV	FMA	SPM	ETR	DIG	PRO	MED	PHY	PSY
American Cancer Society	77.6	70.3	65.5	86.6	83.2	87.9	74.3	79.6	68.1	70.0
eHealth Forum	83.6	70.8	61.4	80.4	71.9	86.3	86.4	84.4	73.9	63.2
PatientsLikeMe	80.3	87.4	62.9	86.0	79.0	78.6	86.5	71.6	72.6	73.4
HealingWell.com	77.8	66.6	86.0	71.2	77.9	80.5	78.3	81.3	72.5	65.8
Seattle Cancer Care Alliance	78.6	70.3	59.6	73.3	72.1	77.4	73.9	79.7	66.7	69.0
Gibis Cancer Center	78.4	61.0	84.7	75.3	78.4	80.9	88.8	78.3	69.6	66.0
Team Draft	84.0	63.3	60.7	83.9	83.4	73.5	72.5	89.9	63.4	70.6
American Lung Association	81.2	87.6	83.3	73.8	77.3	79.2	89.6	73.1	67.1	72.8
HuffPost Impact	78.5	69.4	62.4	81.7	74.5	77.6	81.6	75.0	75.8	82.3
APIAHF	70.0	78.2	56.7	88.1	80.3	76.7	77.5	70.3	69.9	65.8

Table 6: Coverage of different classes of information on top ten most popular online health forums in this study.

Coverage of Patient Information on Different Health Forums

We examine the distributions of different categories of patient information on different online health forums and websites. The amount and range of demographic and health-related patient information available on a forum indicate the forum's richness of information on certain patient conditions. We measure the information coverage of a forum using the ratio of the users on the forum whose extracted health profiles contain the desired class of information. Table 6 shows the respective sizes of patient populations on top ten most popularly encountered health forums in our study. We can see that although the information coverage varies among different domains and health forums, our system can extract a relatively comprehensive set of patient information since the information coverage ratios for all classes on each of the top ten forums consistently exceed 60%.

Distribution of Patient Demographic Information on Different Forums

By extracting patient's demographic information we can further analyze the demographical composition of different patient groups, which can help medical researchers more insightfully analyze the common characteristics and behavior features of the group of the patients. For this need, the implemented prototype system also aims to extract comprehensive patient demographic information from online health forums. Figures 4 and 5 respectively present the age, gender, and ethnic distribution of users on the top ten most popular health forums concerned in this study. Figure 4 shows the age curves of patients extracted from the ten forums respectively. Their peaks

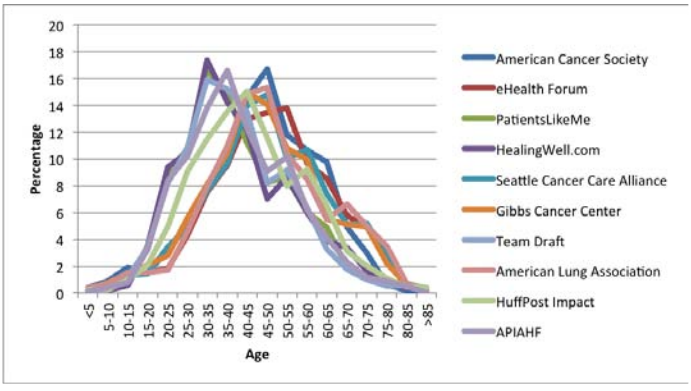
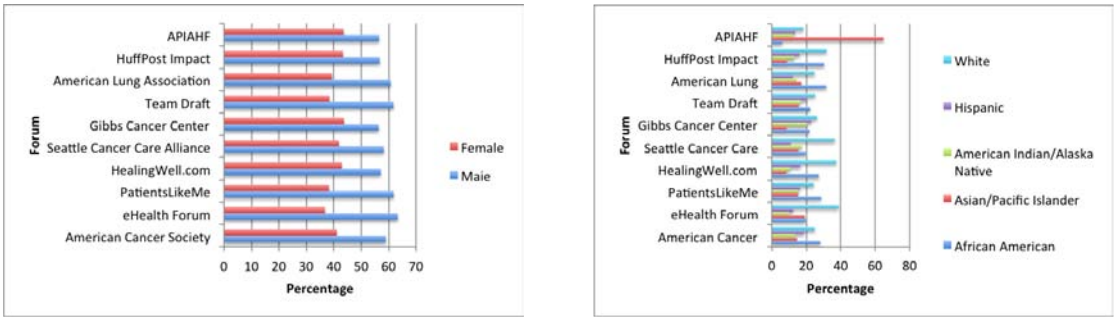


Figure 4: Patient age distributions on each of the ten most popular forums analyzed in this study.

fall into the range of 30 to 55 years old. We assume that the sparse presence of old patients might be partially caused by the inability of those people in accessing and participating on online health forums. We also find that the average age of patients on cancer forums is greater than that on general health forums, which indicates cancer may be more likely to affect older people than young people. In general, the patient demographic statistics extracted can provide valuable clues for analyzing online patient content in an age-adjusted manner.



(a) Gender distribution

(b) Ethnic distribution

Figure 5: Patient gender and ethnic distributions on each of the ten most popular forums analyzed in this study.

Discussion

The proposed biomedical NLP pipeline and its prototype system implementation extract patients demographic and health-related information in a wide range of categories from online health forums. For each category of patient information present on health forums, a corresponding set of NLP procedures are applied to extract values of detailed category attributes. The set of supported information categories is also adaptive, which can be adjusted and expanded over the time to extract more comprehensive patient information. One potential limitation of this work is that a patient may register on mutiple health forums and possibly using different user names or nicknames, none of which may be their real names. How to detect and merge different profiles of the same patient into one consolidatd record is a meaningful algorithmic problem deserving immediate future studies.

Conclusion

We proposed and developed a software pipeline for extracting and structuring patient's personal and health related information from online health forums. Our experimental results demonstrated the usability and effectiveness of the prototype implementation. The automatically acquired health profiles of online patients can provide valuable informational aid for cyber-mining based eHealth research.

Acknowledgement

This study was performed under the Protocol, F 186-14, approved by the Institutional Review Board (HHS FWA #00003246) as well as the Protocal, ORNL(12)-130, approved by the Oak Ridge Site-Wide IRB (HHS FWA #00005031). The study was funded in part by the National Cancer Institute (Grant #: 1R01CA170508).

References

[1] Carol Friedman. Towards a comprehensive medical language processing system: methods and issues. In *Proceedings of the AMIA annual fall symposium*, page 595. American Medical Informatics Association, 1997.

[2] Carol Friedman. A broad-coverage natural language processing system. In *Proceedings of the AMIA Symposium*, page 270. American Medical Informatics Association, 2000.

[3] N Sager, M Lyman, C Bucknall, N Nhan, and LJ Tick. Natural language processing and the representation of clinical data. *Journal of the American Medical Informatics Association*, 1(2):142–160, 1994.

[4] C Friedman, Philip O Alderson, John HM Austin, James J Cimino, and Stephen B Johnson. A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association*, 1(2):161–174, 1994.

[5] Sigfried GM, N Elhadad, XX Zhu, JJ Cimino, and G Hripcsak. Extracting structured medication event information from discharge summaries. 2008.

[6] H Xu, SP Stenner, S Doan, KB Johnson, LR Waitman, and JC Denny. Medex: a medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association*, 17(1):19–24, 2010.

[7] L Deléger, C Grouin, and P Zweigenbaum. Extracting medical information from narrative patient records: the case of medication-related information. *Journal of the American Medical Informatics Association*, 17(5):555–558, 2010.

[8] S Hyun, SB Johnson, and S Bakken. Exploring the ability of natural language processing to extract data from nursing narratives. *Computers Informatics Nursing*, 27(4):215–223, 2009.

[9] A Ritter, O Etzioni, S Clark, et al. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1104–1112. ACM, 2012.

[10] Jason J Jung. Online named entity recognition method for microtexts in social networking services: A case study of twitter. *Expert Systems with Applications*, 39(9):8066–8070, 2012.

[11] P Sondhi, M Gupta, CX Zhai, and J Hockenmaier. Shallow information extraction from medical forum data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1158–1166. Association for Computational Linguistics, 2010.

[12] JJ Cimino, PD Clayton, G Hripcsak, and SB Johnson. Knowledge-based approaches to the maintenance of a large controlled medical terminology. *Journal of the American Medical Informatics Association*, 1(1):35–50, 1994.

[13] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270, 2004.

[14] R Xu, K Supekar, Y Huang, A Das, and A Garber. Combining text classification and hidden markov modeling techniques for structuring randomized clinical trial abstracts. In *AMIA Annual Symposium Proceedings*, volume 2006, page 824. American Medical Informatics Association, 2006.

[15] A Hsieh, S Doan, M Conway, KW Lin, and H Kim. Demographics identification: Variable extraction resource (diver). In *Healthcare Informatics, Imaging and Systems Biology (HISB), 2012 IEEE Second International Conference on*, pages 40–49. IEEE, 2012.

[16] NLP Stanford. Toolkits.

[17] Y Huang, HJ Lowe, D Klein, and RJ Cucina. Improved identification of noun phrases in clinical radiology reports using a high-performance statistical natural language parser augmented with the umls specialist lexicon. *Journal of the American Medical Informatics Association*, 12(3):275–285, 2005.

[18] English Grammar Online. http://www.ego4u.com/. [Online; accessed 1-March-2014].

[19] IH Witten, E Frank, LE Trigg, MA Hall, G Holmes, and SJ Cunningham. Weka: Practical machine learning tools and techniques with java implementations. 1999.

[20] M Torii, K Wagholikar, and HF Liu. Using machine learning for concept extraction on clinical documents from multiple data sources. *Journal of the American Medical Informatics Association*, pages amiajnl–2011, 2011.

Data Model for Personalized Patient Health Guidelines: An Exploratory Study

Mary McNamara[1], Karthik Sarma[3], Denise R. Aberle MD[1,2], Alex A. T. Bui PhD[1,2], Corey Arnold PhD[1,2],

[1]Department of Bioengineering, University of California, Los Angeles
[2]Medical Imaging Informatics, Department of Radiological Sciences, University of California, Los Angeles
[3]David Geffen School of Medicine, University of California, Los Angeles

Abstract

Practitioner guidelines simultaneously provide broad overviews and in-depth details of disease. Written for experts, they are difficult for patients to understand, yet patients often use these guidelines as a source of information to help them to learn about their health. Using practitioner guidelines along with patient information needs and preferences, we created a method to design an information model for providing patients access to their personal health information, linked to individualized, relevant supporting information from guidelines within a patient portal. This model consists of twelve classes of concepts. We manually reviewed and annotated medical records to demonstrate the validity of our model. Each class of the model was found within at least one patient's record, and seven classes of concepts appeared in over half of the patients' records annotated. These annotations show that the model produced by the method can be used to determine what guideline information is relevant to an individual patient, based on concepts in their health information.

Introduction

Patient portals are web-based applications designed to allow patients direct access to content from their medical record. Portals can also provide patients with general content relevant to their health, similar to what they would find at MedlinePlus[1], the Mayo Clinic website[2], or other consumer health websites. While not yet commonplace, large healthcare institutions are beginning to design and implement patient portals. Compelling reasons to do so include: the potential for an informed patient population, government endorsements via policy and funding, and ubiquitous digital infrastructure that allows patients easy access to information. Portals have the potential to empower patients; accessing personal health information encourages patient involvement, enabling them to make decisions based on the information they have received[3]. Applications that promote patient empowerment have also been shown to improve clinical outcomes and health statuses[4].

Searching for health information is now the third most popular task completed online[5]. However, the content available to consumers spans a wide array of quality in their accuracy and completeness of information[6]. In addition, not all accurate sources regarding health and disease will be relevant to an individual. Here, we define "relevance" as how well the information meets the information needs of the user, based on the idea that information that satisfies the information need will fill the knowledge gap of the user[7]. Searching for health information, patients are left to mull content, attempting to determine what is applicable to their personal health and often having difficulty doing so[1,8]. Medical content intended for patient consumption lacks personal context; while letting individuals view information on subjects that concern them, such content does not provide a contextual overview or specific details regarding their diagnosis and the process of their care. Professional medical guidelines also lack a tailored view of the healthcare process relative to a given patient, often enumerating a spectrum of medical concepts (symptoms, test, diagnoses, etc.) that requires an expert to make logical inferences. As such, a lay patient looking at a guideline may be overwhelmed with information, and may not properly comprehend or appreciate the nuances within the guideline.

Motivation

A model for personalized patient guidelines can provide consumers with supporting information regarding their health, presented in a fashion that provides context and chronology, and allowing them to see how individual concepts and processes relate together and to their health. The design of this information model should tailor the content of guidelines to the content of individual medical records in order to support the understanding of relevant information and to lessen cognitive overload. In this work, we address the issue of developing a methodology to model relevant supporting information to medical record content. The construction of the model was defined by the overlap between patient information needs and preferences, relevant professional clinical guidelines, and medical record content. More specifically, content from each of these three domains was examined to determine shared concepts to be included in the model. Examples of each information source area may be seen in Figure 1. Using this model creation method, we were able produce a model that matches relevant supporting content from guidelines to those concepts patients are interested in within their medical records.

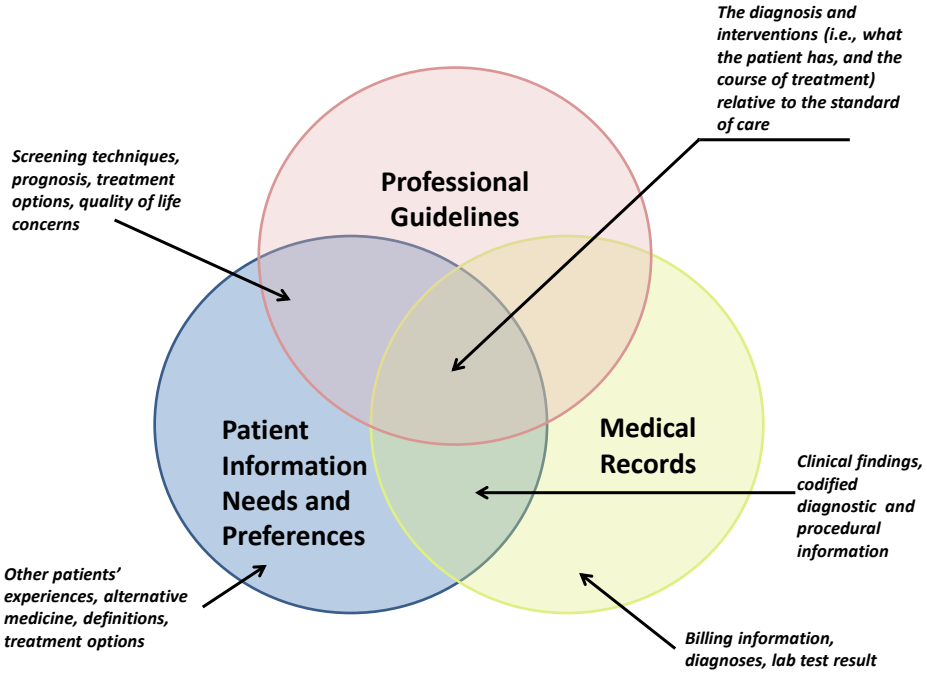


Figure 1. Venn diagram illustrating cross section of information areas of interest, which falls in the overlap of the three domains.

Patients undergoing diagnostic tests have numerous sources of traditional supporting information including pamphlets, informational clinics, and support groups. The provision of electronic health information is also widely available, as noted above, in the form of consumer health information sources[1,2,9-11], and personal health records (PHRs). Even before a diagnosis is made, data is collected on symptoms, procedures and test results. The testing process is often a stressful time for patients, with new information being introduced, both via the healthcare setting and what patients find themselves online. Yet, there is a lack of a consistent, reliable linkage between personal health information, information found within a portal, and other online supporting content. Patients now have ensured access to their medical record via the Health Insurance Portability and Accountability Act, (HIPAA). If such content is to be ultimately comprehensible to patients (e.g., to make informed decisions about their own care), online methods such as patient portals should not simply display medical record content verbatim, but instead make the content understandable to consumers with changes in visual presentation, abstraction level, and vocabulary as necessary to accommodate the lay person.

Professional clinical guidelines provide views of multiple granularities on disease-specific information. However, like individual health records, the content of these guidelines contain medical jargon and typically require the lens of clinical experience and/or knowledge to properly understand the information. By way of illustration, a lung cancer

screening patient with a history of smoking and additional symptoms (e.g., cough, unexplained weight loss) may consult an online professional guideline (e.g., National Collaborating Center for Cancer[12], Figure 2) to learn more about his risk factors and the diagnostic procedures. Yet, as such guidelines outline all possibilities, not just the relevant pathway to the individual's circumstances, the lay reader may become more confused (if not anxious) about his condition. Indeed, professional guidelines take multiple potential symptoms and findings into consideration, and do not represent a patient-specific series of events. For example, while some patients being assessed for non-small cell lung cancer (NSCLC) have an x-ray, others do not. Professional guidelines, although appropriate for practitioners, present information that is sometimes irrelevant, redundant or too complex for the information needs of an individual patient. Given this consideration, patient records cannot simply be linked to professional guidelines as a whole. Depending on the patient, only parts of the practitioner guidelines are applicable. Therefore, guidelines presented to a patient need to be tailored to the individual based on concepts present in their medical record.

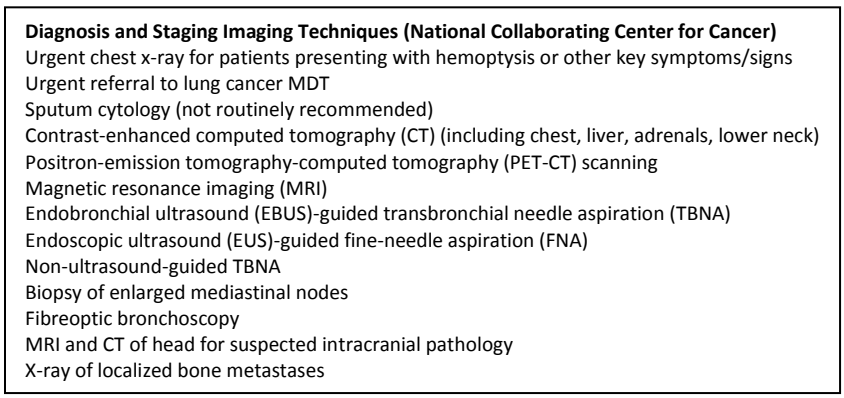


Figure 2. List of imaging techniques from National Collaborating Center for Cancer page on lung cancer screening.

Background

Linking medical record content to supporting information has become commonplace for clinicians, with numerous electronic health record (EHR) systems providing decision support. It has been proposed that patients may be provided with additional information via their PHR in a similar manner[13]. However, the linkage of supporting information for patients via hyperlinks and embedding content is relatively new. MedlinePlus Connect is a web service that accepts ICD-9 codes and returns links to health information[14]. To use this service, institutions must opt in and provide concept unique identifiers (CUIs) from reports. This process also requires the patient to leave the content of their health portal and visit pages from the MedlinePlus website. Numerous healthcare organizations are now using this web service, including Columbia University Medical Center, Sutter Health System and the University of Utah. This system is dependent on correct CUIs in order to retrieve relevant information. However, this pairing of supplemental information can be hindered by incorrect or inexact matches, exclusion of appropriate child matches (e.g., choosing "cough" instead of "chronic cough"), and context dependent definitions[15].

Kaiser Permanente has designed an in-house encyclopedia for patients using their portal[16]. Patients can access the encyclopedia pages while onsite. Each page contains specific information on a medical concept. Yet, the content accessible is not specific to the patient. While patients should not be prevented from viewing additional content that is not necessarily related to the patient, neither MedlinePlus Connect nor the Kaiser Permanente encyclopedia directs patients towards information that focuses on the concepts that have a positive occurrence in their record. In other words, patients can be directed toward information on a biopsy, even if it was decided not to do a biopsy but the word "biopsy" is mentioned in a report. Moreover, while online resources like MedlinePlus Connect can direct patients to relevant supporting information, they do not fully demonstrate how concepts may relate within the framework of a patient's encounter and the narrative of the PHR.

Although there are numerous information models for health information (Heath Level 7's [HL7] Clinical Document Architecture [CDA], Digital Imaging Communications in Medicine [DICOM], etc.), the majority are standards created to support interoperability among healthcare institutions, with practitioners as the end users[17]. To the best of our knowledge, none focus solely on the patients' information needs and how to link patient record content with supporting content. Often, patient information models are the product of using an information model designed for clinicians and then implementing it for a patient view of the information[18-20]. Little study has been done on how patient information needs and preferences should inform information models for patients.

In prior work, we conducted a survey of 41 patients who were undergoing screening or treatment for lung cancer at a clinic in UCLA, regarding their information needs and preferences[21]. Question topics were the result of a literature review of patient information needs and preferences. Results demonstrated that patients were particularly interested in concepts that were important to their diagnosis (90%) and imaging (90%). This earlier work and insight helps guide the development of the proposed information model.

Material and Methods

Class Definition: Literature Review

To test the process of designing the model, the domain of lung cancer screening was chosen. To create an explanatory information model that links guideline information with clinical data for the patient, we first defined a set of classes through a literature review of patient information needs. This literature review is the same utilized to design the survey in McNamara et al. 2014[21] and was performed in March 2013. To conduct this literature review, we used Google Scholar and PubMed search engines to find articles on patient information needs, using the terms "patient information needs", "patient portal" and "patient information preferences". Approximately 5 million articles were returned; however after manually reviewing several pages of ranked results, we observed that the relevance of articles to our study diminished. For example, large numbers of articles focused on other groups' information needs or medical procedures involving the portal vein, both subjects are outside the scope of this study. We reviewed the first 100 articles returned by PubMed and the first 50 returned by Google Scholar ranked by relevance. To be included, an article had to discuss patient information needs and preferences, and be published in the last twenty years. Articles were deemed to discuss patient information needs and preferences based on the content of their title and abstract. In total, 26 pertinent articles were identified. We choose to specify our class set from this literature review, as defining a model relevant to patient information needs is our overarching goal, rather than deriving classes from clinical guidelines, which may be irrelevant to patients. We then analyzed the subset of documents that met our criteria of content and publication years and noted themes that occurred throughout. If a theme was noted in at least three separate articles, the articles in which it was included were added to our focused annotated bibliography. This process resulted in thirteen articles within the focused annotated bibliography[22-34]. Each theme from the bibliography was included as a candidate class. This literature review resulted in the five candidate classes of: Diagnosis[23, 24, 28, 29, 31, 33, 35, 36], Treatment[22-24,28-31,33,35], Common Side Effects of Treatment[22, 23, 29, 35], Symptoms[22, 29, 37, 38], and Diagnostic Test[24,30,34]. As made evident by the literature review, these classes of concepts appear to be popular across patient populations, as such information helps patients to understand their diagnosis and cope with prognosis and treatment. These initial candidate classes could thus serve as the basis for class creation across domains of cancer screening.

Concept Definition: Guideline review

To generate a candidate list of clinical concepts (e.g., normalized instances of nouns found within the diagnosis guidelines) with associated contextual explanations, we reviewed the lung cancer diagnosis guidelines from the National Collaborating Center for Cancer and UpToDate[12, 39]. As seen in Figure 3, these guidelines were visualized as a flowchart, composed of connected nodes.

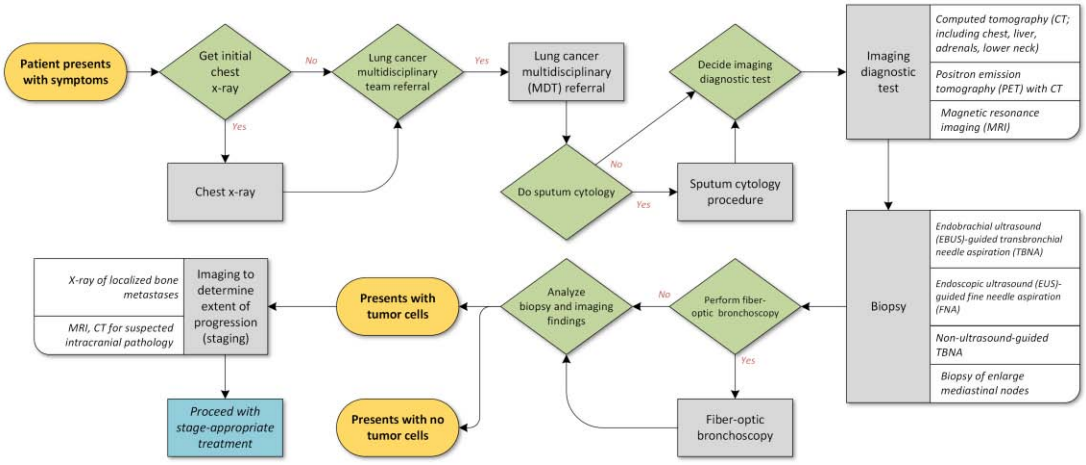


Figure 3. Simplified version of practitioner guidelines for lung cancer diagnosis.

Each node was then considered a candidate concept for the model. With this candidate list, we began to organize the data model, with the constraint that concepts included in the model were representative of classes seen in our literature review on patient information needs, and that the classes included were indicative of the screening process as made evident in our review of the professional guidelines. Table 1 shows our initial model structure of patient information need classes and corresponding concepts.

Table 1. Initial model of concepts and classes.

Class	Tumor	Symptoms	Diagnostic Test
Number of Sources Citing Information Need	8[23,24,28-30,33,35,40]	4[22,27,29]	3[24,30,34]
Guideline Concepts mapped to Class	Tx, T0, Tis, T1, T2, T3, T4	Weight Loss, Fatigue, Chest Pain, Lung Infection, Breathing Trouble, Cough, Hoarse Voice	Sputum Test, Bronchoscopy, Thoracentesis, LDH, PET Scan, Albumin, Chest X-ray, Computed Tomography , Video Assisted Thoracoscopy, Pulmonary Function Test, MRI, Thoracotomy, Fine Needle Aspiration, Mediastinoscopy, Blood Test, Bone Scan

Manual Annotation of Medical Reports

After this initial linking of candidate concepts to classes, we manually annotated pathology, laboratory, oncology, and radiology reports from ten patients to determine the presence of these candidate concepts in reports. We then revised the working list of concepts based on the actual content of reports. This process helped to ensure that the smallest units of information within the model – the concepts – were indeed reflective of the content of reports. During this process, we found that there were concepts indicative of indeterminate nodules, which had not been previously included in the model. This finding required enumeration of several new concepts concerning an indeterminate nodule. We also found that sometimes instead of reporting a TNM stage, a Roman numeral stage was instead reported. Additionally, we found that a patient's smoking history is frequently mentioned in their reports. As smoking history information is relevant to understanding the application of a lung cancer guideline, we decided that these concepts should be incorporated into the model. However, it was found that these concepts were not well-represented by any of the existing classes. Thus, we revised the list of classes to add the classes of Nodule, Stage and Smoking Status. All concepts and classes were then manually matched to their Unified Medical Language System (UMLS) concept unique identifiers (CUIs).

Revision Based on Survey Results

Based on survey results from McNamara et al. 2014[21], we revised the model one more time to include concepts pertaining to the imaging process and concepts relevant to diagnosis. To provide for more detail regarding imaging, the Diagnostic Test class was broken into five classes (Imaging, Biopsy, Excision, Pulmonary Function, and Other Diagnostic Test). To focus on concepts relevant to diagnosis, the Comorbidity class was added. The Comorbidity class contains concepts that are lung disease comorbidities common in smokers.

Manual Annotation of Medical Records

To determine the overlap between the model's concepts and those within patient records, the first author manually annotated an additional unseen 60 patients' oncology, pathology, radiology and laboratory documents. First, a combination of report types (e.g., imaging, pathology, radiology, oncology) and keywords (e.g., "tumor", "smoking", "diagnosis", etc.) were used to filter documents that contained a concept of interest. These reports were then manually reviewed to determine if the concept was present. If, after reviewing these documents there remained unobserved concepts, a patient's entire report set of documents was reviewed to ensure no concept present in a history was unaccounted for.

Results

The finalized information model consists of twelve classes, as seen in Table 2. Nodule is the class that captures the presence of an indeterminate nodule, as well as its location and size. Tumor, while referencing the same nodule, is only utilized when there is confirmation of cancerous cells. Related to this, the class of TNM contains concepts of individual stages of tumor progression, as taken from the TNM staging system of lung cancer[41]. As made evident by

annotations, sometimes Stage I-IV is reported instead of TNM, so the Stage class contains concepts to account for this method of staging. The Smoking class contains a concept that confirms a smoking habit (yes/no), and once confirmed, quantifies it with the concept of pack year history. The Symptom class does not contain every symptom a patient with lung cancer might experience. Rather, it contains only those symptoms commonly experienced by patients as noted in Corliss et al.[9].

The first revision of the model produced a single Diagnostic Test class containing a wide range of diagnostic tests (pulmonary function, imaging, biopsy) that can be used in the diagnosis of lung cancer, as indicated by the National Comprehensive Cancer Network and UpToDate[12, 39]. However due to surveyed patients' interest in imaging, it was decided that the Diagnostic class should be divided into five new classes to allow for a finer granular representation of concepts pertaining to imaging. The five classes are: Biopsy, Imaging, Excision, Pulmonary Function, and Other Diagnostic Test. The Biopsy class is meant to cover all likely types of biopsies associated with the screening process, with the concept "Other Biopsy" to accommodate all other types of biopsy. Similarly, the Imaging, Excision, and Pulmonary Function classes contain concepts reflective of the most common types of concepts associated with them. The Other Diagnostic class contains those concepts that didn't fit into any of the other Diagnostic classes, but were found in the guidelines and reflective of patient information preferences. The Comorbidity class is concerned with smoking related comorbidities, as the majority of lung cancer patients either have smoked or were exposed to second hand smoke.

Out of the 60 patients records annotated for the concepts within these twelve classes, 33 contained the class TNM, 49 contained the class Tumor, and 21 had concepts from the class Stage. 57 patients had concepts from the class Nodule, 35 had concepts from the class Smoking, 56 had concepts from Biopsy class. The Comorbidity class was found in 24 patients' records, 5 patients had concepts from the class Excision. 58 patients had concepts from the class Imaging. Only one patient exhibited concepts from the Pulmonary Function class, one patient had concepts from the Excision class, and one patient had a concept from the Other Diagnostic class. 37 patients had concepts from the Symptoms class.

Table 2. Revised Model of Classes and Concepts.

Symptoms	TNM	Tumor	Stage	Nodule	Smoking Status	Comorbidity	Excision	Imaging	Biopsy	Pulmonary Function Test	Other Diagnostic Tests
Weight Loss	TxNxM*x*	Tumor Present (Y/N)	I	Nodule Present (Yes/No)	Smoker (Yes/No)	COPD	Video-Assisted Thoracic Surgery (VATS)	X-ray	CT-Guided Lung Biopsy	Spirometry	Sputum Test
Fatigue		Tx	II	Nodule Location	Pack Year	Pulmonary Fibrosis	Mediastinoscopy	Computed Tomography	Brochoscopy	Body Plethysmograph	Bone Scan
Chest Pain		T0	III	Nodule Size		Chronic Obstructive Asthma	Thoracotomy	PET Scan	Fine Needle Aspiration	Gas Diffusion	
Lung Infection		Tis	IV	Ground Glass		Chronic Bronchitis		MRI	Thoracentesis		
Breathing Trouble		T1		Multiple Ground Glass		Emphysema			Other Biopsy		
Coughing Blood		T2		Solid							
Hoarse Voice		T3									
		T4									
		Metastases									
		Tumor Location									
		Tumor Size									

Discussion

Based on a literature review of patients' information needs, we designed a method to create an information model for patients that links medical record concepts with guideline content. This method consisted of looking at patient information needs and preferences, and determining which of those also fell within the domains of practitioner guidelines and medical record content. The model is intended to aid in patients' information comprehension. Screening for any disease is complicated and detailed, and can be stressful for the patient, as they encounter and process new information. By providing classes of concepts relevant to the information needs of patients that can be linked to guideline content, the model promotes directing patients to contextualized information relevant to their health. We demonstrated the application of this model within the domain of lung cancer screening by manually annotating 60 lung cancer patient records. While no one class' concepts were present in every record, every class was represented in at least one patient's record. The model's concepts can therefore be used to annotate a patient's record and collect

a set of domain specific concepts. These concepts can then be used to determine what guideline content is relevant to a particular patient's record, based on the concepts within the record. For example, if the concept "CT Guided Lung Biopsy" is found within a patients' record, additional supporting information on the process of getting a CT lung biopsy can be provided to the patient within the same application (i.e. a portal), in which they are accessing their record.

In a prior study, we surveyed 41 patients at a lung cancer clinic at the University of California Los Angeles, to determine how accurately this information model reflected their perceived needs[21]. The survey results support the relevance of our model. 66% wanted information about their health problems, and 90% wanted to know about information on their diagnosis. While less than half (32%) agreed that it was difficult to find information, 61% would like to see terminology from their medical reports defined. This alludes to patients having information available to them, but not necessarily being able to understand it, reaffirming a need to make medical record content more accessible to patients.

Although the concepts of the model are specific to the domain of lung cancer, this method of reviewing patient information needs and preferences, guideline review, and medical record annotation can be used to create an information model for patients undergoing screening in another domain of cancer. In addition, the model produced by the method will likely be similar across cancer domains. Many of the classes produced by this information model method, with perhaps the exception of smoking, would likely be reproduced when the method was implemented in another domain, for instance that of breast cancer. The classes of Nodule, Tumor, and Stage are applicable within breast cancer, as unknown phenomena can be captured by the concept "nodule", and malignant findings are referred as a "tumor" and staged using the same hierarchy of staging. Likewise, while the type of imaging most common varies by type of cancer (mammogram for breast cancer), the Imaging class can easily to altered to focus on those methods most prominent with a particular diagnosis.

Implemented within a portal, this model may allow patients to more effectively learn about concepts within their medical record. While users can currently read their records and then search for information online on concepts found within the record, the proposed model facilitates the automatic linking of record content to educational content by filtering content for an individual patient based on disease presentation within their record. To realize this, natural language processing (NLP) tools may be trained to mine model concepts from patient records, and links from the records to the educational content can be provided via a patient portal. For example, if a patient were to see the concept "CT scan" within their record, they could be provided with links to the definition of a CT scan, images of a CT scanner, example CT scan results, and information on why a CT scan is used in lung cancer screening. This provision of educational content is similar to the personal health record level three as presented in Krist and Woolf 2011[43].

Limitations of this work include: a limited sample size from one institution was used to learn concepts for the model, the assumption that patients' perceptions of what they want to see from their record aligns with their actual information needs, and that annotations in this work were created by one researcher. These limitations may have biased our results to make the model appear to fit better than it will with records from other institutions, or for other annotators.

Our future work includes using this model within a portal as the basis for an information visualization that links patient medical information to guideline content, thus creating personalized guidelines. User studies will be conducted to determine the extent to which the model is successful in helping patients to understand information regarding their lung cancer diagnostic process. We are currently in the process of designing a patient portal that utilizes this model. Once this program is finalized, we will recruit patients from a UCLA lung cancer screening clinic to use the portal and complete a survey based on their perceptions of how well the model meets their information needs. Additional future work may also include the automation of the annotation process using NLP.

Conclusion

We designed and implemented a method to create an information model that links concepts from medical records to relevant information from practitioner guidelines indicative of patient information needs and preferences. Classes for the model were derived from a literature review on patient information needs and populated with concepts found in professional guidelines. Here the application focused on the domain of lung cancer, but is anticipated to be applicable across cancer domains. The revised model was used to annotate 60 patients' records, were it was found that each class was present in at least one patient's record, and that seven of the twelve classes were present in over 30 patients' records.

Acknowledgements

This work was supported by NIH/NCI R01 LM011333, NIH/NIBIB T32 EB016640 and NIH/NLM T15 LM07536.

References

1. Keselman A, Slaughter L, Smith C, Kim H, Divita G, Browne A, et al., editors. Towards consumer friendly PHRs: patients' experience with reviewing their health records. American Medical Informatics Association Symposium 2007.
2. Zielstorff RD. Controlled vocabularies for consumer health. Journal of Biomedical Informatics. 2003;36(4-5):326-33.
3. Trent Rosenbloom S, Daniels T, Talbot T, McClain T, Hennes R, Stenner S, et al. Triaging patients at risk of influenza using a patient portal. JAMIA. 2012;19:549-54.
4. Stroetmann K, Pieper M, Stroetmann V, editors. Understanding patients: participatory approaches for the user evaluation of vital data presentation. ACM Conference on Universal Usability; 2003; Vancouver, British Columbia, Canada.
5. Smith M, R. Saunders, L. Stuckhardt, J.M. McGinnis. Best Care at Lower Cost: the Path to Continuously Learning Health Care in America. Washington D.C.: Institute of Medicine, 2012.
6. Meric F, E.V.Bernstam, N.Q. Mirza, K.K. Hunt, F.C. Ames, M.I. Ross, H.M. Kuerer, R.E. Pollock, M.A. Musen, S.E Singletary. Breast cancer on the world wide web: determinants of web site popularity. Proceeds from the American Society of Clinical Oncology. 2001;20(39b).
7. Ormandy P. Defining information need in health -- assimilating complex theories derived from information science. Health Expectations. 2010;14:92-104.
8. Tse T, D.Soergel, editor Exploring medical expressions used by consumers and the media: an emerging view of consumer health vocabularies. Proceeds from the American Medical Association: AMIA, Symposium; 2003.
9. Corliss J, K. Crowley, D.E. Elbaum. G.J. Long. Patient Information Lung Cancer the Basics 2013 [cited 2013]. Available from: http://www.uptodate.com/contents/lung-cancer-the-basics?source=see_link.
10. Hong W, Tsao A. Lung Cancer: Merck; 2008 [cited 2013]. Available from: http://www.merckmanuals.com/home/lung_and_airway_disorders/cancer_of_the_lungs/lung_cancer.html.
11. Network NCC. NCCN Guidelines for Patients. In: Network NCC, editor. 2012.
12. National Collaborating Center for Cancer. Lung cancer. The diagnosis and treatment of lung cancer. Agency for Healthcare Research and Quality; 2011 [2013]. Available from: http://guideline.gov/content.aspx?id=34282.
13. Kemper D. The Info-Button Standard: Bringing Meaningful Use to Patients 2010 [cited 2013]. Available from: http://thehealthcareblog.com/blog/2010/01/28/the-info-button-standard-bringing-meaningful-use-to-the-patient/.
14. National Library of Medicine. MedlinePlus Connect in Use: National Institutes of Health; 2013 [cited 2013]. Available from: http://www.nlm.nih.gov/medlineplus/connect/users.html.
15. Strasberg HR, G. Del Fiol, J.J Cimino. Terminology challenges implementing the HL7 context-aware knowledge retrieval ('Infobutton') standard. JAMIA. 2013;20:218-23.
16. Silvestre A, Sue V, Allen J. If you build it, will they come? The Kaiser Permanente model of online health care. Health Affairs. 2009:334-44.
17. Blazona B, Koncar M. HL7 and DICOM based integration of radiology departments with healthcare enterprise information systems. International Journal of Medical Informatics. 2007;76S:S425-S32.
18. Winkleman WJ, Leonard KJ. Overcoming structural constraints to patient utilization of electronic medical records: a critical review and proposal for an evaluation framework. JAMIA. 2004;11:151-61.
19. Shea S. The informatics for diabetes and education telemedicine project. Transactions of the American Clinical and Climatological Association. 2002;118:289-304.
20. Sunyaev A, Chornyi D, Mauro C, Krcmar H, editors. Evalution framework for personal health records: Microsoft HealthVault vs. Google Health. Hawaii International Conference on System Sciences; 2010; Hawaii.
21. McNamara M, Arnold C, Sarma K, Aberle D, Garon E, Bui A. Patient portal preferences: perspectives on imaging information. JASIST. 2014;In-Press.
22. Davidson J, Brundage M, Feldman-Stewart D. Lung cancer treatment decisions: patients' desire for participation and information. Psycho-Oncology. 1999;8:511-20.
23. Jenkins V, Fallowfield L, Saul J. Information needs of patients with cancer: results from a large study in UK cancer centers. British Journal of Cancer. 2001;84(1):48-51.
24. Gore J, Brophy C, Greenstone M. How well do we care for patients with end stage chronic obstructive pulmonary disease (COPD)? A comparison of palliative care and quality of life in COPD and lung cancer. Thorax. 2000;55:1000-6.
25. Leydon G, Boulton M, Moynihan C, Jones A, Mossman J. Cancer patients information needs and information seeking behavior: in depth interview study. British Medical Journal. 2000;320:909-13.

26. Murray S, Boyd K, Kendall M, Worth A, Benton T, Clausen H. Dying of lung cancer or cardiac failure: prospective qualitative interview study of patients and their carers in the community. BMJ. 2002;325.
27. Slaughter L, Ruland C, Rotegard A. Mapping Cancer patients' symptoms to UMLS concepts. AMIA. 2005. p. 699-703.
28. Butow P, Maclean M, Dunn S, Tattersall M, Boyer M. The dynamics of change: cancer patients preferences for information, involvement, and support. Annals of Oncology. 1997;8:857-63.
29. Clauser S, Wagner E, Aiello Bowles E, Tuzzio L, Greene S. Improving modern cancer care through information technology. American Journal of Preventive Medicine. 2011;40(5s2):s198-s207.
30. Grant R, Wald J, Poon E, Schnipper J, Gandhi T, Volk L, et al. Design and implementation of a web-based patient portal linked to an ambulatory care electronic health record: Patient Gateway for diabetes collaborative care. Diabetes Technology and Therapeutics. 2006;8(5):576-86.
31. Hess R, Bryce C, McTigue K, Fitzgerakd K, Zickmund S, Olshansky E, et al. The diabetes patient portal: patient perspectives on structure and delivery. Diabetes Spectrum. 2006;19:106-9.
32. Koch-Weser S, Bradshaw YS, Gualtieri L, Gallagher SS. The internet as a health information source: findings from the 2007 health information national trends survey and implications for health communication. Journal of Health Communications. 2010;15(S3):279 -93.
33. Bass S, Ruzek S, Gordon T, Fleisher L, McKeown N, Moore D. The relationship of internet health information use with patient behavior and self efficacy: experiences of newly diagnosed cancer patients who contact the National Cancer Institute's Cancer Information Service. Journal of Health Communications. 2006;11:219-36.
34. Sarkar U, Karter A, Liu J, Alder N, Nguyen R, Lopez A, et al. The literacy divide: health literacy and the use of an internet-based patient portal in an integrated health system—results from the diabetes study of northern California (DISTANCE). Journal of Health Communication. 2010;15(S2):183-96.
35. Leydon G, Boulton M, Moynihan C, Jones A, Mossman J. Cancer patients information needs and information seeking behavior: in-depth interview study. British Medical Journal. 2000;320:909-13.
36. Clauser SB, E.H Wagner, E.J. Aiello Bowles, L. Tuzzio, S.M. Greene. Improving Modern Cancer Care Through Information Technology. American Journal of Preventive Medicine. 2011;40(5s2):s198-s207.
37. Slaughter L, Ruland C, Rotegard A. Mapping cancer patients' symptoms to UMLS concepts. American Medical Informatics Association2005. p. 699-703.
38. Koch-Weser S, Bradshaw Y, Gualtieri L, Gallagher S. The internet as a health information source: findings from the 2007 health information national trends survey and implications for health communication. Journal of Health Communications. 2010;15(S3):279 -93.
39. Deffebach ME, L. Humphrey. Screening for Lung Cancer: UpToDate; 2013 [2013]. Available from: http://www.uptodate.com/contents/screening-for-lung-cancer?source=search_result&search=lung+cancer+screening&selectedTitle=1~19.
40. Hess R, Bryce C, Paone S, Fischer G, McTigue K, Olshansky E, et al. Exploring challenges and potentials of personal health records in diabetes self-management: Implementation and initial assessment. Telemedicine and E-Health. 2007;13(5):509-17.
41. Thomas KW, M.K. Gould. Tumor Metastasis (TNM) Staging System for Non-Small Cell Lung Cancer 2012 [cited 2013]. Available from: http://www.uptodate.com/contents/tumor-node-metastasis-tnm-staging-system-for-non-small-cell-lung-cancer?source=search_result&search=tnm&selectedTitle=1~150.
42. Deffenbach M, Humphrey L. UpToDate Screening for lung cancer 2012 [08/12/2010]. Available from: http://www.uptodate.com/contents/screening-for-lung-cancer?source=search_result&search=lung+cancer+screening&selectedTitle=1~26.
43. Krist A, Woolf S. A vision for patient-centered health information systems. JAMA. 2011;305(3):300-1.

Identification and Management of Information Problems by Emergency Department Staff

Alison R. Murphy, BS, Madhu C. Reddy, PhD
The College of Information Sciences and Technology,
The Pennsylvania State University, University Park, PA

ABSTRACT

Patient-care teams frequently encounter information problems during their daily activities. These information problems include wrong, outdated, conflicting, incomplete, or missing information. Information problems can negatively impact the patient-care workflow, lead to misunderstandings about patient information, and potentially lead to medical errors. Existing research focuses on understanding the cause of these information problems and the impact that they can have on the hospital's workflow. However, there is limited research on how patient-care teams currently identify and manage information problems that they encounter during their work. Through qualitative observations and interviews in an emergency department (ED), we identified the types of information problems encountered by ED staff, and examined how they identified and managed the information problems. We also discuss the impact that these information problems can have on the patient-care teams, including the cascading effects of information problems on workflow and the ambiguous accountability for fixing information problems within collaborative teams.

INTRODUCTION

Hospitals are highly collaborative, information-intensive environments where patient-care teams rely on the accuracy and availability of information to provide safe and effective patient care. However, hospital staff frequently encounter information problems. In this paper, information problems are defined as any wrong, outdated, conflicting, incomplete, or missing information that may interfere with the ability of hospital staff to do their work. Although these information problems have always existed in paper records,[28] there is an increasing need to focus on information problems in electronic records due to the tremendous growth in the use of electronic health record (EHR) systems.

Due to recent U.S. government legislation, there has been an acceleration in the transition from paper-based records to EHR systems. These electronic systems also include the use of other information technology systems that can be integrated with EHRs, such as computerized provider order entry (CPOE) systems.[25,26] Although these EHRs can provide a number of benefits,[3,7,11,24] the use of EHRs do not necessarily eliminate information problems, such as wrong, outdated, conflicting, incomplete, or missing information.[3] In some cases, EHRs can actually introduce new types of information problems,[3,7] such as the unintentional selection of default values[2] and the truncation of data entry fields resulting in the loss of patient data.[23] These information problems can lead to issues among the patient-care team including ambiguity about what treatments/procedures were done to a patient,[3] medical decisions being made based on wrong or outdated information,[16] and even the occurrence of medical errors that could harm patients.[3]

Current medical informatics research focuses primarily on what causes these information problems within hospitals[2,3,6,7,11,16,23,24,29] and the impact that the information problems have on the workflow of hospital staff.[1,3,4,8,9,21] However, there is still limited research that explores how these information problems are identified and managed by the patient-care teams who encounter them during their daily work. If they are not properly identified and managed, there may be serious consequences such as medical errors that harm the patient. Therefore, the goal of this paper is to provide a better understanding of the types of information problems encountered by emergency department (ED) staff, and how the staff identified and managed information problems within a highly collaborative emergency department. We will also discuss the impact that information problems have on collaboration, including the cascading effects of information problems on workflow and the ambiguous accountability for fixing information problems in collaborative teams.

BACKGROUND

Information problems have been extensively studied in the medical informatics community because of the serious impact that they can have on medical errors in patient care. These existing studies primarily focus on the *causes* of information problems, which include information problems caused by the *design* of EHR systems[3] and by the *users*

of EHR systems.[7] Additionally, researchers also discuss the impact that information problems can have on the workflow of hospital staff.

Information Problems Caused by EHR Design

It is a challenge to build information systems for the fast-paced and information-intensive environment of hospitals. Researchers describe how EHR systems tend to be overly structured and designed with rigid rules that encourage data standardization (e.g., drop-down menus, text entry restrictions), which can lead to information problems caused by the design of the EHR.[3] Koppel et al. [11] discuss how system design can cause problems with medication ordering, including: fragmented displays that prevented a coherent view of patients' medications, inflexible ordering formats that led to wrong orders, and separation of system functionality that resulted in double dosing or incompatible orders. Abramson et al.[2] also describe how a medication ordering system automatically selected the default dosing in medication order forms resulting in inaccurate medication requests.

Additionally, Dillion & Lending[6] also identify information problems caused by systems preventing users from entering descriptive data into a record and, instead, forcing them to select values from drop-down menus that are not considered intuitive to the user. EHR design can also prevent the entry of important psycho-social information about patients, which nurses argue, "*provided continuity of care…[and] a richer picture of the patient's situation*" (p. 2065).[29] Furthermore, EHR researchers describe system bugs that cause information problems. This includes EHR systems that truncate data entry fields resulting in lost patient data and EHR systems where two buttons on a screen have the same label but different functionalities, which caused information problems within the system.[23] These studies highlight how EHR design can lead to information problems that users must manage.

Information Problems Caused by EHR Users

Users are also the cause of certain information problems within EHRs. These problems can include information entry errors, delayed information entry, and discrepancies between multiple sources of information. Researchers have often described how users tend to copy-and-paste information from prior notes in the system in order to cut down on their data entry efforts.[3,7,22] However, as Embi et al.[7] highlight, this can lead to outdated or incorrect information being proliferated throughout the system. Siegler & Adelman[22] also discuss how the copy-and-paste function leads to, "*reducing the credibility of the recorded findings, clouding clinical thinking, limiting proper coding, and robbing the chart of its narrative flow and function*" (p. 495).

Other researchers describe how delayed information entry occurs when clinicians are too busy or tightly scheduled to enter patient data into the system directly after seeing the patient, which leads to information in the EHR being outdated or incomplete for extended periods of time.[16] This negatively affects any other patient-care team member who needs access to updated, accurate patient records. As Ash et al.[3] describe, delayed information entry could result in the patient being given the same medication twice by another member of the patient-care team who relies on the patient record for information. Furthermore, information discrepancies are another issue that can occur during the use of EHR systems. Turchin, Shubina, & Goldberg[24] discuss how EHR users encountered situations where medication information provided in the system's structured fields (e.g., medication name, dosage) contradicted information found in the free-text description field. Therefore, these studies highlight how information problems can also be caused by the users themselves.

Impact of Information Problems on Workflow

Within a hospital, a number of clinical and non-clinical staff members must work together to gather, document, and share information in order to provide effective patient care. This workflow of activities is highly collaborative, complex, and prone to interruptions. This is especially true in fast-paced hospital environments like the emergency department where patients require immediate attention. However, it is important that whatever can negatively affect the hospital workflow is minimized in order to reduce impacts to the patient-care process. Medical informatics researchers have studied how information problems can have a negative impact on the hospital workflow.

Researchers discuss how the design of EHR systems can increase the time it takes to perform workflow activities. Holden[8] discusses how issues with information accessibility (e.g., system login, system response time) can negatively impact the workflow of physicians. Additionally, Shachak et al.[21] describe how there are communication interruptions when physicians try to navigate EHR systems to find or enter patient information while also talking with their patients. This can lead to inaccurate information being entered into the system. Other researchers also describe the additional time and effort spent finding and updating information in multiple systems or on different screens.[1,9] The time and effort that users often do not have.

The existing medical informatics literature on information problems describes what causes information problems and the impacts that information problems can have on workflow. In this paper, we seek to extend this understanding to include how patient-care teams *identify* and *manage* the information problems that they encounter during their daily work.

METHODS

Research Site

We conducted this study in the emergency department of a large teaching hospital in northeastern United States. The ED has approximately 55,000 visits per year. The first author conducted 54 hours of observations and 4 hours of semi-structured interviews with 7 clinical and non-clinical staff in the ED. We observed approximately 85 ED staff members (Table 1).

Table 1. Observed ED Staff

ED Staff Role	***Number Observed***
Nurses (including Charge Nurses)	18
Registration Assistants	14
Physicians	12
Residents	10
Emergency Medical Technicians (EMTs)	6
ED Technicians	4
Transporters	4
ED Volunteers	4
Chaplains (spiritual advisors)	3
Sanitation/Cleaning	3
Care Coordinators	2
Social Workers	2
Maintenance	2
Pharmacists	1
TOTAL	**85**

The ED staff communicated and received information using a variety of sources, including: verbal communication, cell phones, pagers, desktop computers, laptops, computers-on-wheels (COWs), paper documents, white boards, and mounted electronic screens that displayed the "tracking board" (i.e. a chart of ED rooms with patient information and the assigned ED staff). The ED staff primarily used an electronic health record (EHR) system to document patient information. This system included the patients' medical records, laboratory and medication orders, laboratory results, clinical team narratives, registration information, and tracking board. One group of users, registration assistants, also used an Admissions Discharge and Transfer (ADT) system that was integrated with the EHR to document information when patients first arrived in the ED. This ADT system created the patients' unique identification number and included their name, date of birth, social security number, zip code, and complaint (i.e. symptoms, why they were there).

Data Collection

In this study, we used qualitative data collection methods including observations and semi-structured interviews. We used this methodological approach because qualitative research includes an immersion into a field site, which allows the researcher to observe the naturalistic processes and activities of the participants.[12] Our ED observations resulted in detailed descriptions of participants' behaviors and interactions related to information problems. These problems were situated within the context of the participants' busy ED setting and everyday work activities.[13] These qualitative methods have also been used in other informatics studies for a similar purpose.[3,4,19]

The first author conducted 54 hours of observations in two ED areas that had the largest amount of communication and information exchange among the staff: the registration area and the main nurses' station. We also observed the ED staff in hallways and at smaller nurses' stations. Detailed field notes were taken about the workflow,

communication, collaboration, and technology used by both clinical and non-clinical staff. During the observations, the first author also had short, informal discussions with participants in order to clarify assumptions or gain additional information about specific situations. The field notes were then transcribed into an electronic document for analysis (Table 2).

The first author also conducted 7 semi-structured interviews in order to better understand the ED staff's roles and responsibilities, as well as their use and perceptions of the EHR system. The first author interviewed 1 physician, 1 registered nurse, 3 registration assistants, 1 care coordinator, and 1 social worker. The interview data was then transcribed into an electronic document for analysis (Table 2).

Table 2. Data Collection Method, Participants, and Data

Method	***Focus***	***Hours***	***Participants***	***Transcribed Data***
Observations	ED Workflow	54 hours	85 staff	175 pages
Semi-structured Interviews	ED Staff	4 hours	7 staff	28 pages

Data Analysis

The transcriptions resulted in 203 pages of data. This data was analyzed by the first author using Braun & Clarke's[5] six-phase thematic analysis approach (Table 3). This approach facilitates the process of becoming familiar with the data, systematically identifying codes and themes, and then defining and naming the common themes found across the entire data set. The analysis resulted in three main themes, as described in the results section.

Table 3. Braun & Clarke Six-Phase Thematic Analysis Approach

Phase	***Description***
(1) Familiarizing ourselves with the data	Transcribe the notes taken during observations/interviews and read through the transcriptions to ensure a general understanding of the data.
(2) Generating initial codes	Label segments of data in a systematic way across all of the data.
(3) Searching for themes	Review individual codes and identify preliminary themes.
(4) Reviewing themes	Review preliminary themes to ensure that they make sense across the entire data set.
(5) Defining and naming themes	Continuously refine each theme, identify a specific name for each theme, and define the boundaries of the theme.
(6) Producing the report	Present themes with interesting examples from the data that illustrate the individual themes.

RESULTS

Through our data analysis, we identified three main findings: the types of information problems encountered by ED staff, how the staff identified the information problems, and how the staff managed the information problems.

Types of Information Problems

We identified the following information problems in the ED: wrong, outdated, conflicting, incomplete, and missing information (Table 4). These information problems identified in the ED are similar to information problems described by other researchers.[3,11,16,29]

Table 4. Types of Information Problems in the ED

Information Problem Type	***Description***	***Example from Data***
Wrong	Information is not accurate.	"*For example, if you put allergies in the wrong spot, they're put in the problems list instead of the allergies list – which happens – [and it led to] another physician making patient care decisions off of the wrong information.*" [physician interview]
Outdated	Information is no longer accurate and has not been updated.	"*We had a patient who was in here because of abuse injuries and the boyfriend was listed as the emergency contact from a previous visit. The boyfriend was called, but the boyfriend was the abuser. So this created a potentially dangerous situation for the patient and the hospital because the information was not updated.*" [social worker interview]
Conflicting	Information that should be the same but it is not the same.	*In the check-out area, a physician tells the registration assistants that his patient's insurance had been stolen. He says that while scanning the patient's registration information he realized something was wrong because the record said the patient's race was listed as African American and the man was Caucasian. The patient then confirmed that the [EHR] record did not have his correct social security number.* [observational field notes]
Incomplete	Only part of the information is provided.	*"Some of my information about my patient does not get into the patient's record because I can't edit [in the system]. And there's no place for me to put some of my notes about the patient, which could help others, like the next shift."* [social worker interview]
Missing	Information is not available.	*In the registration area, a doctor is talking to an EMT and the family of his patient. He says that the patient is refusing to tell him anything about his medical conditions, medications, and what happened in his accident. The doctor says there is no existing record for this patient and he needs any information they can provide so he can proceed with his patient care plan.* [observational field notes]

These information problems occurred quite frequently in the ED. Even though participants discussed the importance of ensuring information accuracy in their work, they also stated that certain problems, such as data entry errors or incompleteness, are often an inherent aspect of the highly dynamic hospital environment. A participant described this: "*There's always the possibility of errors, that's just the nature of entering information. And you only know what you know. Sometimes you don't have the whole story*" [social worker interview].

Identification of Information Problems

The ED staff recognized the information problems when the problems interfered with their ability to do work, and they were subsequently identified as information problems. Any of the information problems described in the previous section could have persisted within the EHR system if no one had identified the problem. Therefore, it is important to understand how ED staff identifies information problems because the information problem must first be identified before it can be managed or fixed by the team. The ED staff identified information problems in two ways: comparing the EHR system's information to other information sources, and trying to make sense of information that did not "look right."

Comparing Sources: The ED staff frequently identified information problems when they were comparing the EHR system information to other sources of information. These other sources included other ED staff members, patients, visitors, or paper records. For example:

A registration assistant identifies that a patient's first and last names are switched in the system while she is talking with a visitor who requested to see the patient [observational field notes].

The registration assistant was comparing the EHR system to the information that the visitor was verbally providing about the patient when she identified this as an information problem. Additionally, a social worker identified an information problem when comparing the information in the EHR system to a patient's statement:

> *"The other problem with the system is people checking the wrong information. So the nurses will ask [patients] questions, like "has anyone ever forced you to have sex against your will?" One time a nurse checked the wrong box for that question, then I am called in [through a notification in the EHR system] to talk to the patient about it. When I addressed them about the sensitive topic, the patient became irate saying that they never said that and demanding that their record be changed"* [social worker interview].

Furthermore, the participants also identified information problems when comparing the EHR information to something they were physically seeing. For instance, the care coordinator identified an information problem when she noticed that the patient's pill color was not the same color that the patient normally received:

> "*I had a patient who was elderly that had an INR of 16 and it should have been 3.5 at most. I went in to talk to the patient and she said, "the color changed" in reference to her pills. So I looked in the records and found that the drug she was taking was 5 times the amount she should have been taking. So [comparing the pill to] the record in the system helped me identify that there was an issue*" [care coordinator interview].

***Sensemaking*:** Participants also described identifying information problems when they encountered information that did not look right, which then caused them to think about why that information did not look right – or to make sense of the information. Sensemaking is "*the process of encoding information into external representations to answer task-specific questions*" (p. 322).[17] When the ED staff tried to make sense of information that did not look right, they were questioning the accuracy of information based on their previous experiences, medical knowledge, or other training or skills that they have. The staff described this as a "*gut instinct*" or simply that the information just "*didn't look right*." The following two examples illustrate the identification of information problems while making ED staff was trying to make sense of the information:

> The care coordinator described a time when the EHR system showed that the diabetes patient was using a glucometer, but noticed that her "*[glucose] numbers were off.*" The care coordinator stated: "*A lot of times I have a gut instinct, you know? It didn't look right. So I talked to the patient and she said, 'I had a glucometer but my son stepped on it' so she wasn't using one at home, even though her record said she was [using a glucometer]*" [care coordinator interview].

> *A registration assistant (RA) questions another RA about something she had entered into the EHR system. The second RA said she could not read her handwriting from her notes when she was transcribing the information into the system, so the information in the system may be wrong. The first RA said: "Yeah, it just didn't look right. So I had to check with you"* [observational field notes].

In both of these examples, the social worker and the RA identified the information problem by making sense of the information and discussing it with a patient or other ED staff member.

Management of Information Problems

After the ED staff identified information problems, they then had to respond to the problem. The participants managed information problems in one of three ways: by fixing the problem, finding the right person to fix the problem, or finding a way to work around the problem.

***Fixing the problem*:** When the participants identified an information problem and knew how to correct the problem, they typically just fixed the issue in the EHR system themselves. This is exemplified in the following two scenarios:

> *In the check-out area, a physician tells the registration assistants that his patient's insurance had been stolen. He says that while scanning the patient's registration information he realized something was wrong because the record said the patient's race was listed as African American and the man was Caucasian. The patient then confirmed that the [EHR] record did not have his correct social security number. The physician says he then asked the patient for his correct information and entered it into the EHR system so that the patient could be billed correctly for the visit"* [observational field notes].

A registration assistant identifies that a patient's first and last names are switched in the system while she is talking with a visitor who requested to see the patient. She then asks the visitor additional questions about the patient (date of birth, home address) to verify the patient's identity and asks other registration assistants about the patient's correct name. The registration assistant realizes that the first and last names of the patient are mistakenly switched, so she fixes the information error in the system [observational field notes].

***Finding someone to fix the problem*:** There were other times where the participant who identified the information problem did not have the knowledge or the editing rights to fix the problem in the EHR system. For instance, in the previous example, a social worker realized that a nurse had checked the wrong box in the system after the patient verified that the system was incorrect. However, the social worker could not change the error in the system so she had to find the nurse to fix it:

"And the issue with the system is that the person who entered it has to change it. So I have to track down the nurse to change it" [social worker interview].

Additionally, the social worker also described a time when the patient's contact information was outdated and she needed to find someone who could fix the problem:

"Since I can't update it as a social worker, I try to call others to update it…I talk to registration, but they say that it's the nurse's responsibility, but when I talk to the nurse, they say it's not their responsibility." When asked if it was unclear whose responsibility it was to fix the problem, the social worker responded: "*Well, it's clear to them that it's* ***not*** *their responsibility [laughs]! But it's not clear whose responsibility it is"* [social worker interview].

This highlights how the participants managed the information problem by finding someone who could fix the problem since the EHR system restricted her from being able to fix it. This example also highlights the ambiguity or conflicting opinions around who is accountable for the accuracy and completeness of patient information in an environment where information is collectively managed by a patient-care team.

***Working around the problem*:** The participants also found ways to work around the information problem so that they could continue with their task without having to fix the problem in the EHR system. For example:

A registration assistant (RA) discusses with another RA how she reads the "comments" field before letting a visitor back to see a patient, since the nurses occasionally write notes in this field about whether a patient can have visitors. However, she also mentions that the field is not always reliable, so they usually call the nurse to double-check [observational field notes].

By calling the nurse, the registration assistant found a way to work around the unreliable "comments" field in order to continue with her task of directing visitors to patient rooms. In addition, the social worker described how the system does not allow her to enter all of her notes about the patient, so she described using a paper-based workaround:

"Some of my information about my patient does not get into the patient's record because I can't edit [in the system]. And there's no place for me to put some of my notes about the patient, which could help others, like the next shift. So I usually just write it on the [paper] face sheet or the next shift takes notes during our hand-off." [social worker interview].

Therefore, these workarounds allowed the ED staff to continue with their work, even though it did not fix the information problem in the system.

DISCUSSION

The results describe the various types of information problems that ED staff encountered during their work activities, as well as some of the approaches that they took to identify and manage those information problems. Given the highly collaborative nature of hospital work, it is important to consider how these information problems affect the collaboration of patient-care teams. In this section, we discuss how the management of information problems can have a serious impact on collaborative hospital teams including the cascading workflow effects of information problem management and the ambiguous accountability for fixing information problems in collaborative teams.

The Cascading Workflow Effects of Information Problems

Patient-care teams are highly collaborative and team members rely on each other to provide accurate and up-to-date information about a patient. EHR systems provide a centralized view of a patient's history and current status, which creates a shared awareness of that patient across the patient-care team. Therefore, when there is an information problem – wrong, outdated, conflicting, incomplete, or missing information – it does not just affect the one team member who identifies the problem. It affects the entire patient-care team who relies on that information to do their work. Therefore, the way in which hospital staff manage information problems can affect the entire team's collaborative understanding of a patient.

After identifying an information problem, hospital staff may manage the information problem by using a workaround.[15] These workarounds are temporary solutions that allow users to adapt technologies or processes in order to minimize interruptions.[27] Our study's results describe how ED staff used workarounds as one way to manage information problems that they encountered during their work. Other researchers have also described how clinicians frequently perform workarounds when the EHR design interferes with their work[18,20,30] or when there is a problem with the information in the system (e.g., missing, incomplete, outdated).[2,10,11,23] Since the primary concern of patient-care teams is to take care of patients, the staff will find ways to perform their activities by working around EHR design or information problems. Additionally, the user may not go back to the system and fix the known issue.[10] Not fixing the problem results in an information problem persisting in the record for an extended period of time. This is common for hospital staff to not immediately fix an information problem in the system because they are responsible for a number of patients and have a variety of urgent tasks.[3,16] However, this can cause problems in highly collaborative environments because the effects of workarounds can directly impact other members of the patient-care team. This is especially true if the known information problem is not fixed and other members of the team are not made aware that the information problem exists.

Kobayashi et al.[10] discuss the "cascading effects" of workarounds where working around one issue can lead to the need for other workarounds. The authors' discussion of cascading effects describes the impact that one person's workaround can have on other team members' work. Similarly, Saleem et al.[20] state that the use of EHR workarounds, "*introduces the potential for gaps in documentation as well as the unintentional propagation of errors*" (p. 662). Therefore, the persistence of an information problem and lack of making others aware of the problem can affect the next team member who uses that information for his/her own work. So, although there may be perfectly valid reasons for hospital staff to enact workarounds to complete their own patient-care tasks, the effects of the workaround on the overall collaborative team should also be considered. For example, our study's results described how the system restricted a social worker from editing the patient's record, so she created a workaround by writing her patient notes on a printed patient report. Although the workaround allowed the social worker to record patient notes for herself, it also prevented any of the other patient-care team members from being able to see her notes in the EHR system. Her notes could help provide a more comprehensive view of the patient's condition to other members of the clinical team, as seen in prior studies.[29]

Ambiguous Accountability When Managing Information Problems in Collaborative Teams

In hospitals, patient information is co-owned and co-managed by multiple members of the patient-care team. This means that there is a shared responsibility for entering, updating, and maintaining the accuracy of the information. However, members of a collaborative patient-care team may have an ambiguous understanding and, at times, a conflicting opinion about who is responsible for managing the information. This can lead to the persistence of information problems because the team members either assume or believe that someone else is responsible for entering the information or fixing information problems.

To help address this issue, collaborative organizations frequently enact formal policies to address accountability. Hospitals create formal policies to outline accountability guidelines for how staff should handle a variety of situations, such as managing information problems. Hospital staff are required to complete training on these policies and comply with the policies during their daily work. However, while some of the formal policies include very specific instructions, others are more general or vague in order to account for variances across different areas of the organization.[14] The issue with these general policies is that the staff may have different interpretations of the policies when conducting their daily work. This can lead to an ambiguous understanding or a conflicting opinion about responsibilities, as discussed in the study's results. Researchers have described how individuals interpret formal policies in their everyday work practices can vary from person-to-person, which causes conflict or tension within

highly collaborative teams.[14] Therefore, understanding how well aligned the formal policies are to the everyday work activities of collaborative teams may provide insight into the persistence of information problems.

RESEARCH LIMITATIONS

A limitation of this study is that the data was collected and analyzed by a single researcher. One researcher in a large ED setting is not able to observe all instances of information problems that occur. Being a single observer required the first author to make choices about who to observe and where to focus when many different activities were occurring simultaneously. However, since the goal of this work was not to present an exhaustive list of information problems, but rather to identify common information problems in the ED and how they were identified and managed by the hospital staff, we believe that we still were able to collect and analyze the data appropriately.

CONCLUSION

Information problems frequently occur in hospitals and can negatively impact the patient-care workflow and potentially cause harm to patients. This study classifies and describes the types of information problems that ED staff encountered during their work activities. Additionally, the paper addresses a limitation in existing medical informatics research by discussing how collaborative patient-care team members identify and manage these information problems. By better understanding and addressing the challenges of identifying and managing information problems within complex collaborative patient-care teams, hospitals could reduce the negative impact that information problems can have on hospital workflow and the patient-care process.

Furthermore, as hospitals transition from paper records to EHR systems, it is important to consider the impact that EHR design has on information problems. EHR systems dynamically display the most recent patient information in a digital format that can be viewed from a variety of distributed locations (e.g., computers, laptops, mobile devices). This differs from paper records that include static information with visual cues indicating who updated the record (e.g., signature, initials, handwriting) and when it was updated (e.g, date, change in ink, new sheet of paper).[28] This shift from paper to electronic documentation changes the way that patient-care teams enter, view, and share information and how they do their work. Therefore, given the serious effects that information problems can have on patient care, it is important to more closely examine the design of EHR systems and the impact that they can have on the identification and management of information problems.

ACKNOWLEDGEMENTS

We would like to thank the hospital's ED staff for their willingness to participate in this study, as well as the AMIA reviewers for their informative feedback. This work is supported by the U.S. National Science Foundation under grant IIS-1017247. Any opinions, findings, and conclusions or recommendations expressed herein are those of the researchers and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

1. Abraham J, Kannampallil TG, Reddy MC. Peripheral activities during EMR use in emergency care: A case study. In Proc. of the American Medical Informatics Association (AMIA) Annual Symposium. 2009:1-5.
2. Abramson EL, Patel V, Malhotra S, Pfoh ER, Osorio SN, Cheriff A, et al. Physician experiences transitioning between an older versus newer electronic health record for electronic prescribing. International Journal of Medical Informatics. 2012;81(8):539-548.
3. Ash JS, Berg M, Coiera E. Some unintended consequences of information technology in health care: The nature of patient care information systems-related errors. Journal of the American Medical Informatics Association. 2004;11(2):104-112.
4. Bardram JE, Hansen TR. Why the plan doesn't hold: A study of situated planning, articulation, and coordination work in a surgical ward. In Proc. of the Computer-Supported Cooperative Work (CSCW) Conference. 2010:331-340.
5. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology. 2006;3(2):77-101.
6. Dillon TW, Lending D. Will they adopt? Effects of privacy and accuracy. Journal of Computer Information Systems. 2010;50(4):20-29.
7. Embi PJ, Yackel TR, Logan JR, Bowen JL, Cooney TG, Gorman PN. Impacts of computerized physician documentation in a teaching hospital. Perceptions of faculty and resident physicians. Journal of the American Medical Informatics Association. 2004;11(4):300-310.

8. Holden RJ. Physicians' beliefs about using EMR and CPOE: In pursuit of a contextualized understanding of health IT use behavior. International Journal of Medical Informatics. 2010;79(2):71-86.
9. Horsky J, Allen MB, Wilcox AR, Pollard SE, Neri P, Pallin DJ, et al. Analysis of user behavior in accessing electronic medical record systems in emergency departments. In Proc. of the American Medical Informatics Association (AMIA) Symposium. 2010:311-315.
10. Kobayashi M, Fussell SR, Xiao Y, Seagull FJ. Work coordination, workflow, and workarounds in a medical context. In Proc. of the Association for Computing Machinery CHI Extended Abstracts on Human Factors in Computing Systems. 2005: 1561-1564.
11. Koppel R, Metlay JP, Cohen A, Abaluck B, Localio AR, Kimmel SE, et al. Role of computerized physician order entry systems in facilitating medication errors. Journal of the American Medical Informatics Association. 2005;293(10): 1197-1203.
12. Lincoln YS, Guba EG. Naturalistic Inquiry. Newbury Park, CA: Sage Publications; 1985.
13. Maxwell, JA. Qualitative research design: An interactive approach. Thousand Oaks, CA: Sage Publications, Inc.; 2005.
14. Murphy AR, Reddy MC, Xu H. Privacy practices in collaborative environments: A study of emergency department staff. In Proc. of the Computer-Supported Cooperative Work (CSCW) Conference. 2014:269-282.
15. Park SY, Chen Y. Adaptation as design: Learning from an EMR deployment study. In Proc. of the Association for Computing Machinery CHI Conference. 2012:2097-2106.
16. Park SY, Lee SY, Chen Y. The effects of EMR deployment on doctors' work practices: A qualitative study in the emergency department of a teaching hospital. International Journal of Medical Informatics. 2012;81(3):204-217.
17. Paul SA, Reddy MC. Understanding together: Sensemaking in collaborative information seeking. In Proc. of the Computer Supported Cooperative Work (CSCW) Conference. 2010:321-330.
18. Poissant L, Pereira J, Tamblyn R, Kawasumi Y. The impact of electronic health records on time efficiency of physicians and nurses: A systematic review. Journal of the American Medical Informatics Association. 2005;12(5):505-516.
19. Reddy MC, Shabot MM, Bradner E. (2008). Evaluating collaborative features of critical care systems: a methodological study of information technology in surgical intensive care units. Journal of Biomedical Informatics. 2008;41(3): 479-487.
20. Saleem JJ, Flanagan M, Militello LG, Arbuckle N, Russ AL, Burgo-Black AL, et al. Paper persistence and computer-based workarounds with the electronic health record in primary care. In Proc. of the Human Factors and Ergonomics Society Annual Meeting. 2011;55(1):660-664.
21. Shachak A, Hadas-Dayagi M, Ziv A, Reis S. Primary care physicians' use of an electronic medical record system: a cognitive task analysis. Journal of General Internal Medicine. 2009;24(3):341-348.
22. Siegler EL, Adelman R. Copy and paste: a remediable hazard of electronic health records. The American Journal of Medicine. 2009;122(6):495-496.
23. Sittig DF, Singh H. Defining health information technology-related errors. Archives of Internal Medicine. 2011;171(14):1281-1284.
24. Turchin A, Shubina M, Goldberg S. Unexpected effects of unintended consequences: EMR prescription discrepancies and hemorrhage in patients on warfarin. In Proc. of the American Medical Informatics Association (AMIA) Annual Symposium. 2011:1412-1417.
25. US Department of Health and Human Services. Benefits of EHRs. Retrieved from http://www.healthit.gov/providers-professionals/benefits-electronic-health-records-ehrs.
26. US Department of Health and Human Services. HITECH Act Enforcement Interim Final Rule. Retrieved from http://www.hhs.gov/ocr/privacy/hipaa/administrative/enforcementrule/hitechenforcementifr.html.
27. Vogelsmeier AA, Halbesleben JR, Scott-Cawiezell JR. Technology implementation and workarounds in the nursing home. Journal of the American Medical Informatics Association. 2008;15(1):114-119.
28. Weir CR, Hammond KW, Embi PJ, Efthimiadis EN, Thielke SM, Hedeen AN. An exploration of the impact of computerized patient documentation on clinical collaboration. International Journal of Medical Informatics. 2011;80(8):e62-e71.
29. Zhou X, Ackerman MS, Zheng K. I just don't know why it's gone: maintaining informal information use in inpatient care. In Proc. of the Association for Computing Machinery CHI Conference. 2009:2061-2070.
30. Zhou X, Ackerman M, Zheng K. CPOE workarounds, boundary objects, and assemblages. In Proc. of the SIGCHI Conference on Human Factors in Computing Systems. 2011:3353-3362.

An Evaluation of Two Methods for Generating Synthetic HL7 Segments Reflecting Real-World Health Information Exchange Transactions

Thomas S. Mwogi, MBChB MMed[1,2], Paul G. Biondich, MD MS[1,2], Shaun J. Grannis, MD MS[1,2]
[1]Regenstrief Institute, Indianapolis, IN
[2]Indiana University Purdue University (IUPUI), Indianapolis, IN

Abstract
Motivated by the need for readily available data for testing an open-source health information exchange platform, we developed and evaluated two methods for generating synthetic messages. The methods used HL7 version 2 messages obtained from the Indiana Network for Patient Care. Data from both methods were analyzed to assess how effectively the output reflected original 'real-world' data. The Markov Chain method (MCM) used an algorithm based on transitional probability matrix while the Music Box model (MBM) randomly selected messages of particular trigger type from the original data to generate new messages. The MBM was faster, generated shorter messages and exhibited less variation in message length. The MCM required more computational power, generated longer messages with more message length variability. Both methods exhibited adequate coverage, producing a high proportion of messages consistent with original messages. Both methods yielded similar rates of valid messages.

Introduction

Health information is gathered within and among different organizations and technical systems, stored in different formats with different identifiers. This situation requires technology that integrates disparate data, and such technology must be validated through testing. To validate software's ability to meet industry standards including interoperability with pre-existing software, the use of real world data during the testing phase is necessary.

Access to real world data for testing software is however very challenging for various reasons.[1] First, technology developers often lack rights to access identifiable protected health information (PHI). Second, when such access is permitted, strict regulations governing the use of PHI may impose de-identification and data management burdens limiting the efficiency and effectiveness of the testing process. Further, when representative data is unavailable, software developers may substitute simplistic test data that lacks the vagaries, complexities and idiosyncrasies of real-world data. Subsequently, the lack of access to representative transactional testing data may hinder software development and testing, potentially impacting software quality.

The work in this paper was motivated by a real world problem. OpenHIE is a global, open-source collaborative initiative emerging to assist in the strengthening of national health information exchanges for underserved and resource poor settings.[2] It combines several open source components necessary to accommodate large volumes of health information exchange transactions. To test OpenHIE's integrative functionality, we require representative HL7 messages from 'real-world' settings that must be made available to the multiple open source development communities that comprise OpenHIE.

Although general approaches for generating synthetic data have been published.[3-5] Specific methods for generating large volumes of representative synthetic messages approximating real world HL7 transactions are currently not well-described, nor are we aware of any freely available tools to support this need. The HL7 Messaging Workbench tool allows an array of functionality most prominently allowing creation of conformance profiles and standards conformant message segments based on HL7 version 2.[6] However, functionality to insert synthetic content into HL7 version 2 messages is pending.

Consequently, we sought to evaluate the feasibility, efficiency and effectiveness of two methods that generate synthetic HL7 messages using de-identified HL7 data from the Indiana Network for Patient Care, the nation's largest and longest running health information exchange. We chose these two methods based on their ability to generate new data from the original messages.

Our process for creating synthetic messages involved specific phases. First, we create a message segment framework consisting of valid HL7 message segments. Once the message segment framework is created, we then inject instance data (simulated patient data, clinical data, etc.) into the appropriate fields in the message. We focus our analysis in

this paper on the first phase, message segment creation. We compare the two methods using various metrics, in order to determine the most effective method for generating synthetic HL7 message segment frameworks that accurately approximate the original real-world messages. The comparison metrics included computational effort, level of compliance with the HL7 messaging standard, variability of messages generated, and conformance to the original 'real-world' data.

Segment generation in this paper is focused on HL7 version 2. This was largely dictated by the fact that both the HIE source transaction data (from the INPC) is HL7 version 2, and the target software to be tested (OpenHIE) consumes HL7 version 2 messages. HL7 version 2 is one of the most broadly used messaging standards across the world, and in developing countries where OpenHIE is likely to have widespread use, HL7 version 2 offers the advantage of small data footprint, which is well adapted to slow network infrastructures. For this reason, HL7 version 2 has a big role to play and will likely co-exist with other newer standards like HL7 version 3 and Clinical Document Architecture.

Methods

Source Data

We extracted HL7 messages received by the Indiana Network for Patient Care (INPC) during a continuous 24-hour period. The messages were de-identified and stripped down to include only HL7 segments together with the timestamp, source facility, and the message header components including the message type and event type. The figure below illustrates the example source HL7 message segment framework data from the INPC.

```
201311112257|ORU^R01|MSH|PID|PV1|ORC|OBR1|OBX1|OBX2|OBX3|OBX4|OBX5|OBX6|OBX7|OBX8|OBX9|OBX10
201311112257|ORU^R01|MSH|PID|PV1|ORC|OBR1|OBX1
201311112257|ORU^R01|MSH|PID|PV1|ORC|OBR1|OBX1|OBX2|NTE|NTE|NTE|NTE|NTE|NTE|NTE
201311112258|ORU^R01|MSH|PID|PV1|ORC|OBR1|OBX1|OBX2|OBX3|OBX4|OBX5|OBX6|OBX7|OBX8
201311112258|ORU^R01|MSH|PID|PV1|ORC|OBR1|OBX1|NTE|NTE|NTE|NTE|NTE|NTE|NTE|NTE|NTE|NTE
201311112359|ADT^A08|MSH|EVN|PID|ROL|NK1|NK1|PV1|PV2|ROL|ROL|GT1|IN1|IN2|ZIN
201311112358|ORU^R01|MSH|PID|PV1|ORC|OBR1|OBX1
```

Figure 1: Example HL7 message segments extracted from source messages exchanged within the INPC.

We developed a python script to analyze the proportion of HL7 event types in the dataset. The event-type proportions were used to ensure a similar proportion of messages in the synthetic output. We implemted two methods for synthesizing new HL7 message segment frameworks, labelled the Markov Chain Model[7] (MCM) and the Music Box Model (MBM), respectively.

Two Approaches to HL7 message synthesis

Markov Chain Model: A python script was written based on the principles of the markov chain model[7], which is an approach that models nodes (states within a system) and transition probabilities among nodes. In our HL7 segment framework generating process, nodes represent HL7 segments and the transition probabilities represent the probability of transitioniing from one HL7 segment to the next. We developed a transition matrix using the original HL7 source messages. The transition matrix defined the probability of each segment transitioning to the next segment. Transition matrices were generated for each specific HL7 message event type. We then used the transition matrix to generate random HL7 message frameworks that approximated the original messages.

Music Box Model: This model's name was derived from the music box, a 19th/20th century music instrument that produces sound using a set of pins placed on a revolving cylinder so as to pluck the tuned teeth in what would seem like random plucking movement. For this method we generated HL7 message segment frameworks by choosing messages from the original de-identified source data using simple random sampling with replacement and then adding the chosen message segment to a new generated pool. Using the precalculated proportions for each event type HL7 messages matching the event type were picked at random from the original messages until a desired number of messages were generated.

The two models were evaluated using several parameters to determine the most suitable model that can be used to generate data that were consistent with messages from INPC. These were the parameters of interest:

1. Time required to generate message segment frameworks.
2. Number of segments generated per message.

3. Proporation of message segment frameworks that conform to the HL7 messaging standard rules.
4. Proportion of message segment frameworks that match original messages.

Determining conformance with HL7 standard rules: In order to validate the segment transitions that were generated, we developed a collection of valid segment transitions based on the HL7 message specification. Each HL7 message type contains an abstract message with a collection of segments including rules describing features such as optionality and repetition. Below is an example of an incomplete ADT^A01 abstract message with its rules.

```
MSH             Message Header
EVN             Event Type
PID             Patient Identification
 [PD1]          Additional Demographics
[ { NK1 } ]     Next of Kin / Associated Parties
PV1             Patient Visit
[ PV2 ]         Patient Visit - Additional Info.
[ { DB1 } ]     Disability Information
```

Figure 2: ADT^A01 abstract message with rules

We used a python script to generate all possible valid segment transitions from the information contained in the abstract messages of each trigger event. The above ADT^A01 abstract message would generate the following valid segment transitions separated by commas and so on.

`MSH|EVN,EVN|PID,PID|PD1,PID|NK1,PID|PV1,PD1|NK1,PD1|PV1,NK1|PV1,NK1|NK1,PV1|PV2,PV1|DB1,PV2|DB1,DB1|DB1…`

Using the set of valid segment transitions, we could then determine the percentage of valid segment transitions generated by each of the two models.

Results

Our original INPC data source contained 627,329 representative HL7 messages. The proportion of each message type stratified by trigger event is shown below.

Table 1: Original HL7 message proportions per event type

	TYPE	FREQUENCY		TYPE	FREQUENCY		TYPE	FREQUENCY
1	ORU^R01	0.669666	**12**	ADT^A06	0.001153	**23**	ADT^A13	0.000072
2	ADT^A08	0.148348	**13**	ADT^A09	0.000607	**24**	ADT^A38	0.000056
3	ADT^A04	0.063141	**14**	ADT^A18	0.000575	**25**	ADT^A25	0.000032
4	ADT^A03	0.026297	**15**	ADT^A16	0.000561	**26**	ADT^A28	0.000018
5	ADT^A31	0.023431	**16**	ADT^A44	0.000453	**27**	ADT^A23	0.000014
6	ORM^O01	0.019760	**17**	ADT^A07	0.000348	**28**	ADT^A14	0.000006
7	MDM^T02	0.013127	**18**	ADT^A15	0.000268	**29**	ADT^A32	0.000005
8	ADT^A05	0.010650	**19**	ADT^A11	0.000204	**30**	ORU^R03	0.000003
9	BAR^P01	0.009563	**20**	ADT^A26	0.000159	**31**	ADT^A12	0.000003
10	ADT^A01	0.006517	**21**	ADT^A34	0.000123	**32**	ADT^A33	0.000003
11	ADT^A02	0.004757	**22**	ADT^A10	0.000080	**33**	ADT^A17	0.000002

The most frequent message triggers in the source data set were ORU^R01, ADT^A08, ADT^A04, ADT^A03 and ADT^A31, which comprised 93.1% of the total messages. These proportions were maintained when generating message segment frameworks.

The two models were used to generate increasingly larger number of HL7 message segment frameworks – from 100 to 1,000,000. The two models were compared with respect to the amount of time required to generate messages, total

segments generated and the number and percentage of valid segment transitions in all the messages generated. Table 2 below shows the results.

Table 2: Markov Chain vs Music Box model in generating sequential increasing HL7 messages

	Markov Chain Model vs Music Box Model in generating sequentially increasing messages					
	Markov Chain Model			Music Box Model		
Number of HL7 Messages	Time (sec)	Total Segments	Valid Segments (%)	Time (sec)	Total Segments	Valid Segments (%)
100	0	784	741 **(94.5)**	9	742	718 **(96.7)**
250	1	2,470	2,397 **(97.0)**	8	1,681	1,628 **(96.9)**
500	2	4,893	4,762 **(97.3)**	8	3,982	3,842 **(96.5)**
750	3	8,233	7,956 **(96.6)**	8	5,723	5,538 **(96.8)**
1,000	4	9,791	9,443 **(96.5)**	9	7,755	7,504 **(96.8)**
2,500	10	27,547	26,619 **(96.6)**	9	20,449	19,890 **(97.3)**
5,000	17	46,271	44,465 **(96.1)**	10	38,403	37,259 **(97.0)**
7,500	28	74,587	71,999 **(96.5)**	11	54,993	53,300 **(96.9)**
10,000	35	94,401	90,825 **(96.2)**	12	76,626	74,332 **(97.0)**
25,000	93	247,556	238,643 **(96.4)**	18	191,678	185,853 **(97.0)**
50,000	177	481,597	463,539 **(96.3)**	27	375,491	363,900 **(97.0)**
75,000	271	719,318	692,848 **(96.3)**	42	585,645	568,228 **(97.0)**
100,000	355	956,191	920,368 **(96.3)**	48	766,797	743,159 **(97.0)**
250,000	880	2,414,645	2,325,899 **(96.3)**	123	1,933,244	1,874,977 **(97.0)**
500,000	1769	4,860,340	4,682,388 **(96.3)**	222	3,834,268	3,717,745 **(97.0)**
750,000	2752	7,246,021	6,978,775 **(96.3)**	323	5,739,783	5,565,519 **(97.0)**
1,000,000	3695	9,656,978	9,301,474 **(96.3)**	425	7,655,476	7,423,464 **(97.0)**

On average the Markov chain model generated more segments per message when compared with the Music box model. The percent of valid HL7 segment transitions in both models were comparable ranging between 94% and 98%. In both models the proportion of valid message segments generated was independent of the number of messages generated.

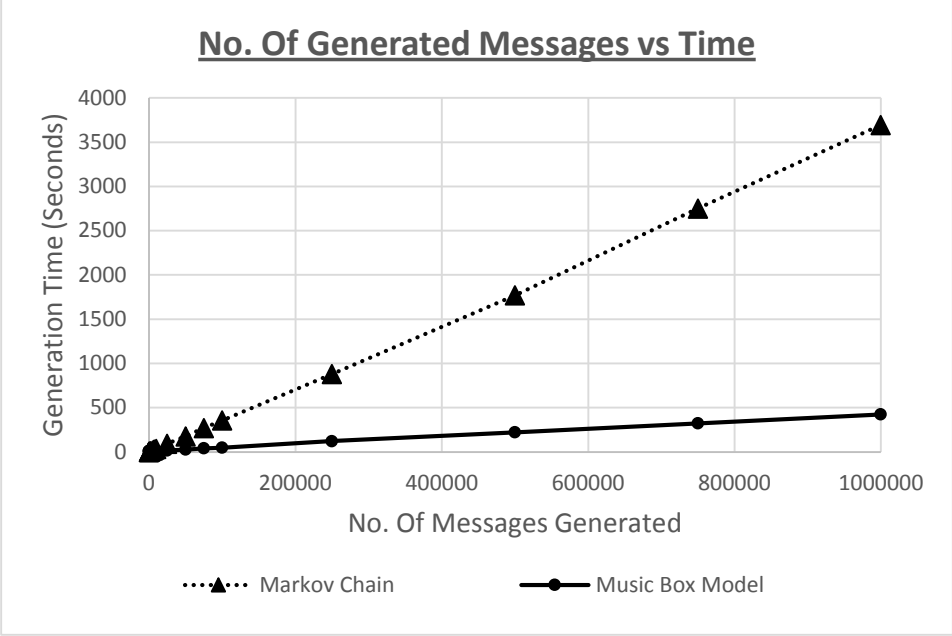


Figure 3: Comparisons of time taken to generate increasing no. of messages by the two models

Because the Music Box method must load all source messages into memory before generating new messages, the Markov Chain method was faster when generating a small numbers of messages (< 2,500). Overall, the music box method was faster. The Markov Chain method generated 270 message segment frameworks per second and the Music Box method generated 1,100 message segment frameworks per second. The message generating rates were nearly constant, as shown in Figure 3.

Comparing coverage of the two models: We assessed the degree to which each method generated message segment frameworks corresponding to one or more of the original source messages, a measure we defined as "coverage". Table 3 below shows the data obtained when the two models generated sequentially increasing number of messages.

Table 3: Comparisons of the two models in terms of coverage of original messages

	Markov Chain Model vs Music Box Model Coverage Of Original HL7 Messages		
	Markov Chain Model		**Music Box Model**
Number of HL7 Messages	**No. of messages identical to original HL7 messages (%)**		**No. of messages identical to original HL7 messages (%)**
100	281,630 **(44.8)**		350,997 **(55.9)**
250	312,527 **(49.8)**		388,518 **(61.9)**
500	356,187 **(56.8)**		418,368 **(66.7)**
750	361,454 **(57.6)**		431,500 **(68.8)**
1,000	396,876 **(63.3)**		456,437 **(72.8)**
2,500	427,513 **(68.2)**		502,382 **(80.1)**
5,000	453,976 **(72.4)**		519,882 **(82.9)**
7,500	463,736 **(73.9)**		527,091 **(84.0)**
10,000	470,447 **(75.0)**		549,014 **(87.5)**
25,000	496,201 **(79.1)**		570,735 **(91.0)**
50,000	506,989 **(80.8)**		584,566 **(93.2)**
75,000	522,169 **(83.2)**		590,466 **(94.1)**
100,000	530,474 **(84.6)**		595,614 **(94.9)**
250,000	540,282 **(86.1)**		606,252 **(96.6)**
500,000	553,554 **(88.2)**		612,750 **(97.7)**
750,000	560,124 **(89.3)**		615,214 **(98.1)**
1,000,000	566,515 **(90.3)**		616,733 **(98.3)**

Overall the Music Box method exhibited greater coverage than the Markov chain method. Both models showed that with an increasing number of HL7 message segment frameworks generated, there was a corresponding increase in coverage of the original messages. The rate of coverage increase plateaued after 75,000 messages. The graph below reflects the data above. Note that the initial 100 messages in both methods corresponded with a more than 40% coverage:

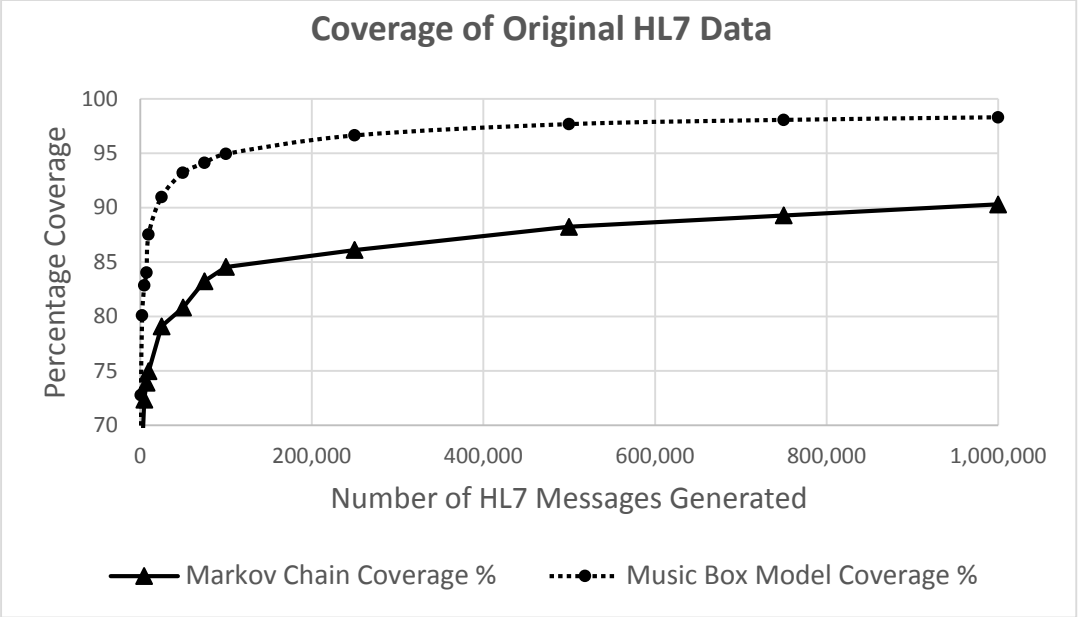


Figure 4: Graph of percentage coverage of original HL7 messages with increasing number of messages generated

The Music Box model had a higher coverage than the Markov Chain for every level of any number of HL7 messages generated. The Music Box model increased to 98.3% compared with Markov Chain at 90.3% when both generated 1 million messages.

Shown below is the comparisons based on segment length per message:

Table 4: Analysis of the number of segments per HL7 message of the two models

	Original Data	Music Box Model	Markov Chain Model
No. of HL7 Messages	627,329	750,000	750,000
Mean	15.67	12.65	14.66
Median	10	10	10
Mode	9	9	9
Std. Deviation	28.24	17.91	34.92
Skewness	14.62	21.32	26.39
Std. Error of Skewness	.003	.003	.003
Minimum	6	6	6
Maximum	1,145	980	3,612

We compared the average number of segments per message in the original 627,329 source messages to 750,000 messages generated by the Markov Chain model and the Music Box model. The most common number of segments per message (mode) generated was 9 in both the original data and both methods. All the data in the three arms were positively skewed with a median of 10 segments. Mean segment length was 15.67 in the original data closely matched by the Markov Chain model at 14.66 segments per message. The spread of segment length per message was also closely correlated between the original data and the Markov Chain as shown by the standard deviation in comparison to that of the Music Box mode. However, the Markov Chain model was prone to generating messages with significantly large number of segments (up to 3,612) compared to the largest number of segments in the original data (1,145).

Discussion

Overall the Music Box model required less time to generate messages when compared with the Markov Chain model. The Music Box model was 5 times faster than the Markov Chain model in generating the same number of messages. This is primarily explained by the fact that the Markov Chain model uses a complex algorithm based on transitional matrix probabilities, which require more computational effort. The Music Box model, on the other hand, implements a straight forward simple random sample of messages, and is therefore preferable when messages must be generated with maximum speed.

Both models generated messages with similar rates of valid segments as gaged by compliance with the HL7 standard rules for each message trigger type. Both topped 97% which reflected to proportion of valid message segments in the Indiana Health Information Exchange source. This rate of valid messages was independent of the number of messages to be generated.

There was a high proportion of identical message segment frameworks in the original dataset. This was reflected by the high coverage by the generated messages. As few as 100 messages generated by both models corresponded to approximately 300,000 (50%) of the original messages. Overall, the Music Box model generated a higher proportion of message segment frameworks corresponding original messages. This is explained by the fact that this method was randomly selecting from the pool of original messages. The Markov Chain method attained a coverage of 90.3% after generating a million messages.

The Markov Chain model generated relatively longer messages when compared to both the original messages and the Music Box model. The Markov Chain also generated greater variability with respect to the number of segments per message compared to the Music Box model. In most situations where software is being tested, this wider variability of messages may be desirable.

Alternatively, the Music Box model generated relatively shorter messages compared with the original messages. This is not necessarily a reflection of the weakness of this method but more a reflection of selecting the most common messages out of the original data. The messages with the largest number of segments were much less common and therefore were less likely to be selected by the Music Box model. However, since these messages are used in generating the transition matrix, the Markov Chain model did reproduce these messages.

Using the mean and standard deviation, the Markov Chain model appeared to more closely correlate with the original data than the Music Box Model. The pattern of the length of messages generated by both models was skewed to the right. Both models showed a statistically significant standard error of skew at 0.03. This means most messages were short with a median segment length of 10. However, messages of up to 3,000 segments per message existed on the Markov Chain model compared to the maximum segment length of 1,145 in the original data. This may have contributed to the Markov Chain method's higher mean and larger standard deviation that correlated more closely with the original data. It is possible that if we constrained number of segments allowed by the Markov Chain method, and thus minimized its outliers, the MCM may have less closely correlated with the original data.

Either method can be used to generate HL7 segments in other settings tailored to the local context. In this paper, we used source data from the Indiana Network for Patient Care. The source data was in the form of de-identified HL7 data stripped down to segment level. Informaticians interested in generating segments using either method in their local context can user our software (made available upon request) to analyze the proportion of HL7 message types in their dataset and then generate transition matrices for each message type. Synthetic message segments for each message type can then be generated according to the transition matrices. For the MBM method, new segments can be generated by randomly selecting each message type in keeping to the analyzed proportions.

There are limitations in making conclusions about HL7 message generation from this study. The synthetic process in this paper focused on generation of HL7 message segment frameworks only and doesn't involve generation of individual field-level data within the segments. Our future work includes developing pragmatic methods for populating fields in each segment with synthetic data that reflect underlying characteristics of the original data. The two methods used to generate message segments can be similarly applied to generate data for individual fields in each segment. The MBM method is being used to randomly select simulated patient identifiers to complete the HL7 patient ID (PID) segment. The MCM will be used to generate simulated address data by first generating transition matrices based on real-world address data. Therefore, both methods are useful not only for generating the segments but also for generating the entire HL7 message.

The MCM's use of transitional matrices introduces additional complexities beyond simple random sampling that require more computational cycles regardless of the language of code or optimization level, thus we believe that the fundamentally different characteristics of the two approaches accounted for much of the difference observe time to completion. However implementation-specific software inefficiencies also likely contributed to a portion the observed time differences. Thus, different implementations of these approaches may yield different computational efficiency results.

Conclusion

In summary, both methods represent effective approaches to generating message segment frameworks, which are necessary precursors to creating fully realized synthetic HL7 messages. The Music Box model was faster, generated shorter message segment frameworks and had less variability in message segment framework length. The Markov Chain required more computational power, generated longer message segment frameworks with some outliers and had more variability in message segment framework length. Both models demonstrated adequate coverage, generating message segment frameworks corresponding to a high proportion of the original messages. The data generated by both models also had a high compliance with the HL7 standard rules.

The work in this paper forms an important first phase in evaluating important models that can be used to generate HL7 messages that reflect real-world data.

References

1. McHale JV. Using anonymized NHS data without consent: a step too far? British Journal of Nursing. 2012;21(1):54-5.
2. Grannis S, Biondich P, editors. OpenHIE: Helping underserved environments better Leverage their electronic health information through standardization. Proceedings of the 2014 PHI Conference; Apr 29-May 1, 2014; Atlanta, GA.
3. Barse EL, Kvarnstrom H, Jonsson E, editors. Synthesizing test data for fraud detection systems. Computer Security Applications Conference, 2003 Proceedings 19th Annual; 2003 8-12 Dec. 2003.
4. Lin P, editor Development of a synthetic data set generator for building and testing information discovery systems. Proc 3rd Int'l Conf Information Technology; 2006: IEEE CS Press.
5. Houkjaer K, Torp K, Wind R, editors. Simple and Realistic Data Generation. Proc 32nd Very Large Databases; 2006: VLDB Endowment.
6. Workbench M. Developed by Peter Rontey at the US Veterans Administration (VA) in conjunction with the HL7 Conformance Special Interest Group.
7. Markov AA. Extension of the limit theorems of probability theory to a sum of variables connected in a chain. Dynamic Probabilistic Systems. 1971;1(Markov Chains).

Scalable and High-Throughput Execution of Clinical Quality Measures from Electronic Health Records using MapReduce and the JBoss® Drools Engine

Kevin J. Peterson, MS[1,2] and Jyotishman Pathak, PhD[1]
[1] **Dept. of Health Sciences Research, Mayo Clinic, Rochester, MN**
[2] **Dept. of Computer Science & Engineering, University of Minnesota, Twin Cities, MN**

ABSTRACT

Automated execution of electronic Clinical Quality Measures (eCQMs) from electronic health records (EHRs) on large patient populations remains a significant challenge, and the testability, interoperability, and scalability of measure execution are critical. The High Throughput Phenotyping (HTP; `http://phenotypeportal.org`) project aligns with these goals by using the standards-based HL7 Health Quality Measures Format (HQMF) and Quality Data Model (QDM) for measure specification, as well as Common Terminology Services 2 (CTS2) for semantic interpretation. The HQMF/QDM representation is automatically transformed into a JBoss® Drools workflow, enabling horizontal scalability via clustering and MapReduce algorithms. Using Project Cypress, automated verification metrics can then be produced. Our results show linear scalability for nine executed 2014 Center for Medicare and Medicaid Services (CMS) eCQMs for eligible professionals and hospitals for >1,000,000 patients, and verified execution correctness of 96.4% based on Project Cypress test data of 58 eCQMs.

1 INTRODUCTION

Secondary use of electronic health record (EHR) data is a broad domain that includes clinical quality measures, observational cohorts, outcomes research and comparative effectiveness research. A common thread across all these use cases is the design, implementation, and execution of "EHR-driven phenotyping algorithms" for identifying patients that meet certain criteria for diseases, conditions and events of interest (e.g., Type 2 Diabetes), and the subsequent analysis of the query results. The core principle for implementing and executing the phenotyping algorithms comprises extracting and evaluating data from EHRs, including but not limited to, diagnosis, procedures, vitals, laboratory values, medication use, and NLP-derived observations. While we have successfully demonstrated the applicability of such algorithms for clinical and translational research within the Electronic Medical Records and Genomics (eMERGE)[1] and Strategic Health IT Advance Research Project (SHARPn)[2, 3], an important aspect that has received limited attention is scalable and high-throughput execution of electronic Clinical Quality Measures (eCQMs) using EHR data. Briefly, eCQMs are Center for Medicare and Medicaid Services (CMS) defined quality measures for eligible professionals and eligible hospitals for use in the EHR Incentive program for electronic reporting[4]. The e-specifications include the data elements, logic and definitions for that measure in an Health Level Seven (HL7) standard - Health Quality Measures Format (HQMF) - which represents a clinical quality measure as an electronic XML document modeled using the Quality Data Model (QDM). While significant advancements have been made in the clarity of measure logic and coded value sets for the 2014 set of eCQMs for Meaningful Use (MU) Stage 2, these measures often required human interpretation and translation into local queries in order to extract necessary data and produce calculations. There have been two challenges in particular: (1) local data elements in an EHR may not be natively represented in a format consistent with HQMF/QDM including the required terminologies and value sets; and (2) an EHR typically does not natively have the capability to automatically consume and execute measure logic. In other words, work has been needed locally to translate the logic (e.g., using an approach such as SQL) and to map local data elements and codes (e.g., mapping an element using a proprietary code to SNOMED). This erodes the advantages of the eMeasure approach, but has been the reality for many vendors and institutions in the first stage of MU[5].

In our prior work[3], we have devised solutions for EHR data normalization and standardization, including generation of structured data templates using natural language processing. In this work, we address the second challenge for scalable and high-throughput execution of eCQMs using MapReduce, the open-source JBoss® Drools Rules engine, and Common Terminology Services 2 (CTS2)[6, 7]. Specifically, we have developed a set of architectural principles that guide the functional and non-functional requirements of our proposed system that supports interoperability, testability, and scalability for automated execution of eCQMs. To evaluate and test our system for performance, scala-

bility, and correctness of measure result calculations, we leveraged test datasets from Project Cypress which provides a rigorous and repeatable testing tool of EHRs and EHR modules in calculating MU Stage 2 CQMs, and is the official testing tool for the 2014 EHR Certification program supported by the Office of the National Coordinator for Health IT (ONC). Our evaluation results show that the system implemented can not only replicate Cypress benchmarks for calculated measure results (accuracy of 96.4%), but also linearly scales to a large number of patient records (>1,000,000 patients). The source code of our system, released under Apache License Version 2.0, is available via: `http://api.phenotypeportal.org`.

2 MATERIALS AND METHODS

In order to evaluate a candidate architecture, we first present a series of Architectural Quality Attributes that outline the critical high-level functional and non-functional requirements of the system. These guide design decisions and technology choices, and must hold for the candidate architecture to be viable. Next, we examine the individual components of the system and their interconnections, along with an examination of the general architectural style. To reinforce the system architectural goals, we will investigate one of the eCQMs, CMS MeasureID CMS163v1 (*Diabetes: Low Density Lipoprotein (LDL) Management and Control*), as a use case at various stages of transformation and ultimately computation. Finally, to test our architecture and implementation viability, we quantitatively measure system performance in terms of measure computation time and execution "correctness," or verification of the results.

2.1 Tools, Standards, and Models

Project Cypress. Project Cypress[1] is a testing and verification framework for eCQMs and is the official testing tool for EHR measure calculation provided by the ONC. It provides a set of 36 de-identified test patients, along with expected execution results for each quality measure calculation. Both the patient dataset and the expected results are distributed in JSON format[8], which is easily parsed and analyzed for testing. Project Cypress also includes several table-based graphical views of expected results, along with in-depth debugging information.

QDM/HQMF. The Quality Data Model (QDM)[9, 10] is a standardized electronic model for representing quality criteria. These models are represented using the Health Quality Measures Format (HQMF)[11] specification, which is an HL7 Version 3 Draft Standard for Trial Use (DSTU). For any given eCQM, the QDM defines, via logical operators and temporal comparisons, criteria used to classify patients into several categories, or *Populations*. These include the *Initial Patient Population* (IPP), *Numerator* (NUMER), *Denominator* (DENOM), *Denominator Exclusion* (DENEX), and the *Denominator Exception* (DENEXCEP)[12].

CTS2. Common Terminology Services 2 (CTS2) is an Object Management Group® terminology services standard[6], as well as an accepted HL7 Normative Service Functional Model.[2] The CTS2 standard defines a common data model and interface for interacting with a terminology service, including read/write access and various maintenance, content loading, and querying functionality.

2.2 Architectural Quality Attributes

In order to better understand the intent and function of the system, we have identified several guiding concepts, or Architectural Quality Attributes[13], that are necessary for a successful system.

Quality Attribute 1: Interoperability
Statement: Communications, interfaces, and data exchanges should be based on industry standards and specifications.
Rational: Standards-based design allows for effective componentization of the system, as component interfaces and data contracts can be linked to published specifications. This promotes low coupling and improved separation of concerns. A standards-based approach also enables implementation and information hiding[14], which allows greater opportunities for substituting component implementations based on use case.

[1] http://projectcypress.org/
[2] Publishing pending at time of writing.

Implications: First, a standardized measure definition format is needed. This will enable interoperability and provide a common, shared grammar for communicating measure intent. Second, as structured vocabularies play a major role in the measure logic, a standard terminology service interface is required. Finally, standard protocols are needed to integrate these components and to interact with the system.

Quality Attribute 2: Testability
Statement: System results should be verifiable, repeatable, and comparable to benchmarked and curated datasets.
Rational: The identification of patient cohorts in many instances remains a manual, human-driven process[15]. This presents a system testing problem, as it may be difficult to determine acceptable system outputs without significant effort. It is important, therefore, that a subject matter expert curated test dataset be available, and that the system architecture incorporate such datasets in order to allow for automated and repeatable testing.
Implications: Testing infrastructure will need to accommodate the automation of benchmarked and curated test datasets. This also will require reporting processes to clearly demonstrate test results.

Quality Attribute 3: Scalability
Statement: System architecture will allow for increasing workloads and expanding datasets.
Rational: As shown in the eMERGE[1] project, efficiently processing large datasets is important to research. High scalability is necessary for both present and future viability.
Implications: Processing algorithms and deployment models must enable horizontal scalability and parallel execution.

2.3 Architectural Style

The architectural style of our platform is a combination of *pipe and filter* transformations and a *MapReduce* computational algorithm. In the pipe and filter phase, input data is transformed via several discrete components in a streaming fashion, creating an information flow through the system which ultimately leads to the desired output[16]. Output of this phase is a machine-executable representation of the quality measure. One of the main benefits of this style is loose coupling, as components need not know implementation details of each other. This approach is possible via our Architectural Principle 1 – *Interoperability* and standards-based design – as components will have explicit and standardized contracts. These contracts are implemented as Representational State Transfer (REST) based services wherever possible, giving them a consistent interface[17]. Measure execution logic is then partitioned by the MapReduce computational algorithm, allowing for parallelization and horizontal scalability.

2.4 System Components

The overall system architecture and scope is shown in Figure 1. Functional software modules of the execution system are enumerated below, while other components or infrastructure shown can be assumed to be external dependencies.

HQMF/QDM XML to JSON Transformer. The HQMF/QDM XML to JSON transformation component parses the HQMF/QDM XML representation and transforms it into easily parsable JSON. This functionality is exposed via a REST service wrapping the Project Cypress health-data-standards[3] project. This component is Ruby[4] based, and represents the only non-Java Virtual Machine (JVM) component of the architecture.

JSON to JBoss® Drools Compiler. The QDM JSON representation is then programmatically transformed into executable JBoss® Drools Rules via a compilation process. The result of this transformation is a single Drools Rule file per measure that is executable by the *Drools Execution Engine* or other JBoss® Drools environment. The compilation component is based on the Groovy[5] language, as Groovy provides templating and string interpolation functionality that is difficult to replicate in Java.

Drools Execution Engine. Utilizing the JBoss® Drools Rule engine, the generated Drools Rules are executed given an input patient set. The result of this execution is the classification of patients into populations (such as NUMER,

[3] https://github.com/projectcypress/health-data-standards
[4] https://www.ruby-lang.org/en/
[5] http://groovy.codehaus.org/

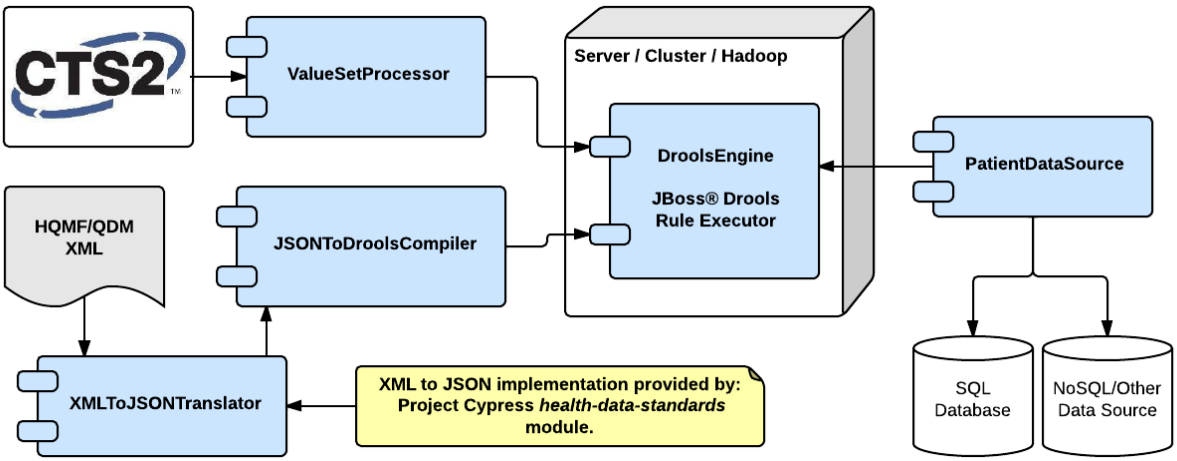


Figure 1: Scalable and High-Throughput Phenotyping Architecture for Clinical Quality Measure Calculation

DENOM, etc). A thin interface façade was created over the "JBoss® Drools version 5.6.0.Final" library, hiding the implementation details of the engine. There are, in fact, no Drools-related interfaces exposed to the client – meaning the Drools-based implementation is decoupled from the user-facing contracts.

Patient Data Source. Data is provided to the execution engine via an abstraction layer utilizing the Iterator pattern[18]. This approach neither provides nor specifies a specific data source, but facilitates custom data extraction implementations tailored to the specific database and environment.

Value Set Processor. QDM Data Elements semantically define their intended criteria using *Value Sets*, or enumerated sets of coded concepts from standard vocabularies. Take, for instance, the criteria: *"Diagnosis, Active: Acute Pharyngitis" using "Acute Pharyngitis Grouping Value Set (2.16.840.1.113883.3.464.1003.102.12.1011)."* In this example, in order for the QDM Data Element to be satisfied, a patient record must contain a Diagnosis entry coded with concepts contained in the Value Set uniquely identified by the given HL7 Object Identifier (OID)[19].

To facilitate this, a mechanism for resolving the concepts of a Value Set is required. This functionality is described in two procedures shown in Algorithm 1, and defines the capability of the *Value Set Processor*. The procedure `EXISTS_IN_VALUESET` defers explicit full resolution in favor of an *exists*-type functionality, which tests membership of a given $coded\_entry$ within a Value Set. In this procedure, the set of concepts C denotes each concept c defined by the Value Set identified by the given OID. The boolean existence of $coded\_entry$ in C is returned. Furthermore, it often is necessary to find the subset of coded entries that align semantically with a defined Value Set. In the procedure `FIND_MATCHES`, instead of checking existence in C as above, the intersection of C and the $coded\_elements$ set is returned. In both procedures, the function $resolve(oid)$ returns the set of all concepts in a Value Set given an OID.

Algorithm 1 Value Set Processor Functionality

1: **procedure** EXISTS_IN_VALUESET(oid, coded_entry)
2: $C \leftarrow \{c | c \in resolve(oid)\}$
3: **return** $coded\_entry \in C$
4: **end procedure**
5: **procedure** FIND_MATCHES(oid, coded_elements)
6: $C \leftarrow \{c | c \in resolve(oid)\}$
7: **return** $C \cap coded\_elements$
8: **end procedure**

2.5 Processing Algorithms and System Deployment

MapReduce. MapReduce is a computational abstraction designed to allow for easier parallelization and distribution of large datasets for processing[20]. The computation is a two-step process, consisting of a *map* function followed by a *reduce* function. The *map* function, given an input, divides this input into smaller sub-problems which then are consolidated by the *reduce* function.

We can apply this concept in the context of patients and populations. In the *map* function, each patient is classified into the appropriate populations given an eCQM. We then *reduce* each population (and its given set of patients) into some analytic result – for instance, analyzing demographic characteristics or creating patient lists for the quality measure.

Algorithm 2 MapReduce for Clinical Quality Measures

```
1: map(String measure, List patients)
2:     for each patient in patients:
3:         for each population in getPopulations(measure, patient):
4:             emit(population, patient)
5: reduce(String population, Iterator patients)
6:     emit(postprocess(population, patients))
```

Algorithm 2 illustrates the general flow of information through the measure execution logic. The *getPopulations* function encompasses the actual measure execution logic – in our case, the *Drools Execution Engine* component. The *postprocess* function in the *reduce* portion is left to be implemented by the client, as this will be use case dependent. Note that the current code base includes a basic demographics analytics post processor, but it is expected that clients will extend it to meet their needs, or substitute their own processing entirely.

2.6 Measure Execution using the JBoss® Drools Rules Engine.

The JBoss® Drools Rule Engine provides the execution framework for the clinical quality measures. Given the input HQMF/QDM XML, the intermediary JSON representation is then compiled into Drools Rules. Figure 2 shows a representation of CMS MeasureID CMS163v1 and the translation of one example criteria to a Drools Rule. In a Drools Rule execution, decisions are made based on data elements, or *Facts*, that have been placed into working memory either from an external data source or as the result of other rule executions. In this example, a `MeasurementPeriod` is first checked – this is a parameter that controls the valid start/end date range of the measure. A `Patient` is then retrieved from the working memory, and the `Encounters` of that patient are examined for a *semantic* match (a match in Value Set 2.16.840.1.113883.3.464.1003.101.12.1001), and a *temporal* match (expressed by the `during` operator in reference to the `MeasurementPeriod`). Finally, if this result is satisfied, a `PreconditionResult` is inserted into working memory. The Drools execution engine, using the Rete[21] algorithm, assures that dependent rules will then fire, as the new `PreconditionResult` is now available in working memory.

2.7 Quantitative Evaluation Methods

Performance (Execution Time). Scalability and performance are important aspects of the architecture, and the ability to efficiently process large patient sets is critical. Examining the architecture for these aspects requires several components - first, a large de-identified patient set (>1,000,000 patients with EHR data) is needed. Next, a mechanism to verify the execution results is required to ensure a valid test environment. Finally, as shown in our prior work[22], computational complexity is variable across different quality measures, with examples ranging from simple logical comparisons, to large nested decision trees with temporal relationships between events. We must account for this variability by measuring complexity of the measures under study.

To meet these requirements, the previously mentioned Project Cypress test dataset was re-purposed for execution time performance testing. In order to simulate >1,000,000 patients, the Project Cypress' 36 patients were "cloned"

Initial Patient Population =

- AND: "Diagnosis, Active: Diabetes" starts before or during "Measurement Period"
- AND: "Patient Characteristic Birthdate: birth date" >= 18 year(s) starts before start of "Measurement Period"
- AND: "Patient Characteristic Birthdate: birth date" <= 75 year(s) starts before start of "Measurement Period"
- AND:
  - OR: "Encounter, Performed: Office Visit"
  - OR: "Encounter, Performed: Face-to-Face Interaction"
  - OR: "Encounter, Performed: Preventive Care Services - Established Office Visit, 18 and Up"
  - OR: "Encounter, Performed: Preventive Care Services-Initial Office Visit, 18 and Up"
  - OR: "Encounter, Performed: Home Healthcare Services"

```
rule "EncounterPerformedOfficeVisit_precondition_21"
   dialect "mvel"
   no-loop
when
   $measurementPeriod : MeasurementPeriod()
   $p : Patient ( )
   $event : edu.mayo.qdm.patient.Encounter(
       this during $measurementPeriod
   ) from droolsUtil.findMatches("2.16.840.1.113883.3.464.1003.101.12.1001"), $p.getEncounters())
   then
       insertLogical(new PreconditionResult("EncounterPerformedOfficeVisit_precondition_21", $p, $event ))
   end
end
```

Figure 2: Drools Rule Representation of Initial Patient Population for CMS MeasureID CMS163v1

and re-created to form a larger cohort. Execution time was then measured for 12 different patient set sizes, each set size equaling the Project Cypress test patient set size multiplied by an increasing multiple of 2500. Specifically, given the Project Cypress test patient set P, we can define each test patient multiset T_j as $T_j = P \otimes (2500j)$ where $1 <= j <= 12$. For each set of test patient data T_j, the execution time necessary to compute the test set was measured and recorded.

The complexity of eCQMs under study is an important criteria to quantify, especially when comparing execution times. Several efforts provide guidance in this area[12, 23], from which we derive our complexity attributes: *Max Depth, Boolean Operators, Negations,* and *Temporal Operators*. *Max Depth* measures the maximum amount of nesting in the boolean clauses. If thought of as a boolean decision tree, this measurement would equal the longest path length to the root node. *Boolean Operators* is a count of every `AND`/`OR` operator in the measure, while *Negations* measures how many of these operators include a negation. Finally, *Temporal Operators* counts the number of temporal logic calculations specified in the measure – either as relative temporal comparisons between events, or comparisons to the start and/or end time of the measurement period.

Due to brevity, a subset of the complexity measurements is shown in Table 1 (See full details at: `http://docs.phenotypeportal.org/figures/htp_complexity.png`). To test the computation time of the execution engine under a variety of complexity scenarios, 9 measures representing different complexities were executed. The chosen measures represent a broad spectrum of complexity, ranging from the most complex to relatively straightforward. Execution tests were performed on an Amazon 2xlarge EC2 instance with 8 vCPUs and 30 GiB RAM.

Table 1: CMS Clinical Quality Measure Algorithm Complexity Metrics

CMS ID	Max Depth	Boolean Operators	Negations	Temporal Operators
CMS159v1	8	41	16	56
CMS160v1	6	46	12	63
CMS153v1	4	110	27	112
CMS145v1	4	84	14	114
CMS126v1	3	57	9	64
CMS134v1	3	26	6	29
CMS163v1	2	10	1	12
CMS127v1	2	8	0	10
CMS148v1	2	4	1	8

Performance (Verification). Automating the verification of the measure is in many ways as important as the automation of the execution itself. Quality Attribute 2 describes the need for a robust testing environment, and the ability to measure verification performance is key to our architectural evaluation. Verification was performed using the Project Cypress verification framework, following guidance from the *Test Procedure for §170.314© Clinical Quality Measures* document[24]. Project Cypress was used as the testing oracle for the verification process, providing a curated test dataset along with expected outcomes. Automation of this verification process also allowed for Test Driven Development[25] practices to be used. To automate this verification process, first the Project Cypress JSON Patient Data was transformed into the target HTP data representation format. Next, the HQMF/QDM XML representation of a given eCQM was selected, and execution of that measure against the test dataset was started. When complete, the results were compared against the expected outcomes, which were loaded from a provided Project Cypress results file.

3 Results

Performance (Execution Time). Execution times, shown in Figure 3, show the total elapsed execution time over the test patient sets. Execution times tended to scale linearly as the patient set size increased. The execution time of CMS MeasureID CMS145v1 *Beta-Blocker TherapyPrior Myocardial Infarction (MI) or Left Ventricular Systolic Dysfunction (LVEF* $< 40\%$*)* was anomalous, as shown in the results. Although similar in measured complexity compared to other executed measures, it exhibited a significantly higher execution time.

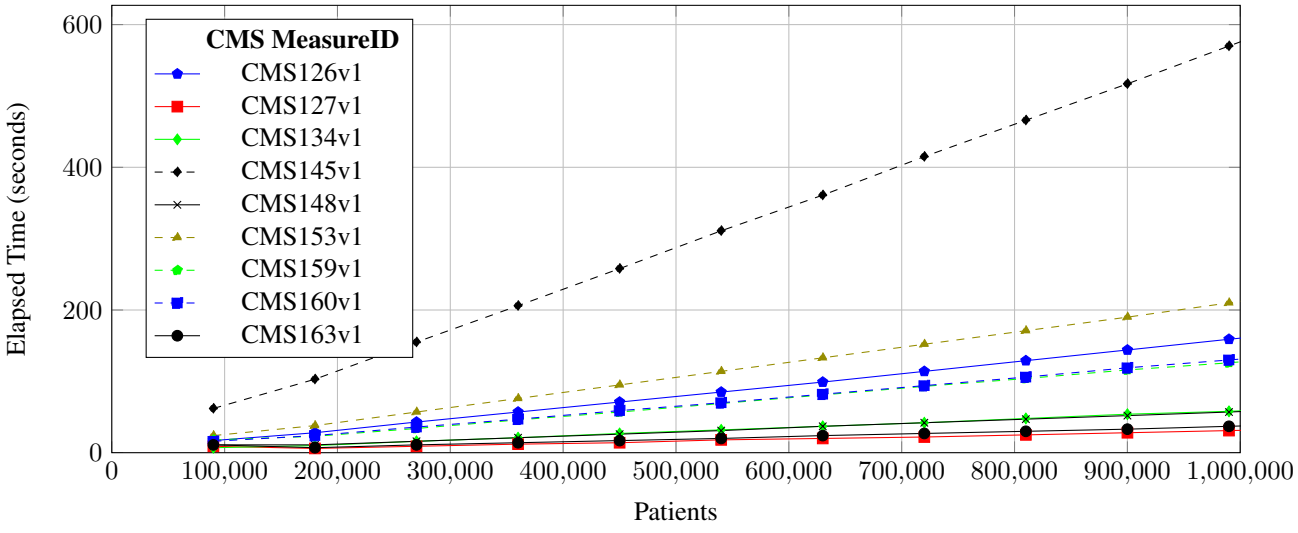


Figure 3: CMS Clinical Quality Measure Algorithm Execution Time

Table 2: Project Cypress Verification Results - Patients Classified per Population *(actual/expected)*

CMS ID	IPP	NUMER	DENOM	DENEX	DENEXCEP	Result
CMS122v1	**2/2**	**2/1**	**2/2**	**0/0**		**Failure**
CMS126v1	3/3	1/1	3/3	1/1		Success
CMS127v1	5/5	1/1	5/5			Success
CMS134v1	2/2	1/1	2/2	0/0		Success
CMS145v1	2/2	1/1	2/2		0/0	Success
CMS148v1	2/2	1/1	2/2	0/0		Success
CMS153v1	4/4	1/1	4/4	1/1		Success
CMS159v1	**2/2**	**0/1**	**2/2**	**0/0**		**Failure**
CMS160v1	3/3	2/2	3/3	0/0		Success
CMS163v1	2/2	1/1	2/2	0/0		Success

Performance (Verification). A subset of the Project Cypress validation results are shown in Table 2. For each measure, results are displayed in the form of *(actual/expected)* patients classified for the given population. Of the 58 total measures analyzed, 56 matched the expected results, while 2 measures failed to match the expected results in one or more of the population criteria. The two anomalous measures, CMS MeasureIDs CMS159v1 and CMS122v1, both failed to correctly identify the *Numerator* (NUMER) population. Due to brevity, see detailed results here: `http://docs.phenotypeportal.org/figures/htp_cypress.png`.

4 Discussion

Summary. The evaluation of the candidate architecture begins with ensuring that the Architectural Quality Attributes still hold. By using standards such as QDM/HQMF and CTS2, along with the uniform interface of REST, we were able to achieve a degree of *Interoperability* in the system. Also, as shown by the Project Cypress integration, the architecture was able to achieve a high degree of *Testability*, while the MapReduce processing algorithm enables a solid platform for high system *Scalability*. Our quantitative analysis results also verify that the Quality Attributes hold. Achieving 96.4% Cypress verification (56 out of 58 measures) shows high system "correctness," while the execution time results show predictable and scalable results for the selected measures given datasets of $>$1,000,000 patients.

Anomalous Results. CMS MeasureIDs CMS159v1 and CMS122v1 both failed to produce the correct results when executed against Project Cypress test data. For CMS159v1, the numerator population is defined by *"Risk Category Assessment: PHQ-9 Tool (result $<$ 5)" $<=$ 13 month(s) starts after end of "Occurrence A of Risk Category Assessment: PHQ-9 Tool."* Cypress verification results identify test patient `BH_Adult_D` for inclusion, and test data for this patient shows two qualifying assessments, the first on *Sat, 24 Sep 2011 12:30:00 GMT* and the second on *Thu, 01 Nov 2012 12:30:00 GMT*. Our results exclude this patient via the $<=$ 13 month(s) temporal constraint. It is unclear as to whether this is a misinterpretation of the temporal constraint or an otherwise unknown system defect.

With regard to the verification failure of CMS122v1, the test patient `Diabetes_Adult_A` was erroneously added to the numerator population due to a miscalculation of the the constraint: *"Occurrence A of Laboratory Test, Result: HbA1c Laboratory Test (result $>$ 9 %)."* This patient does have an HbA1c Laboratory Test result, but with a value of only 8%. Inclusion of this test patient indicates a system defect in analyzing the value of this laboratory test result.

The anomalous execution time of CMS145v1 is not sufficiently explained by complexity calculations in Table 1, as it does not differ significantly from other observed measures. One possible explanation could be the large number of individual Drools Rules produced by the compilation process of this measure. To explore this further, Figure 4 shows the number of generated Drools Rules as compared to execution time for each analyzed measure. The results indicate a possible correlation between the two, and shows the execution time of CMS145v1 degrades predictably given the large number of discrete Drools Rules needed to express it.

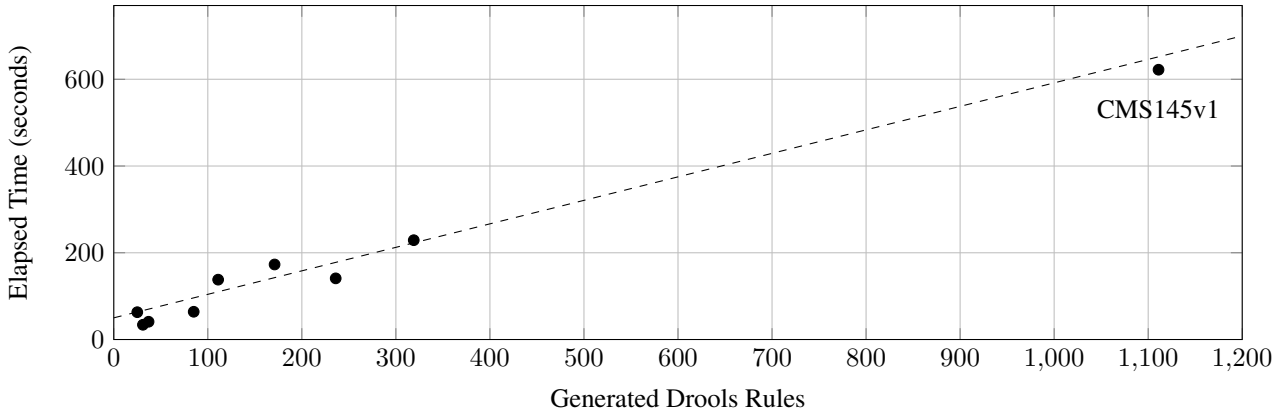


Figure 4: Measure Execution Time per 1,000,000 Patients vs. Number of Generated Drools Rules

Limitations. Compilation of the QDM XML into JBoss® Drools Rules has several significant opportunities for improvement. First, Drools Fusion and Complex Event Processing[7] was only partially utilized for the temporal relationships. Utilizing these modules would cause a significant increase in rule readability and conciseness. Next, the generated rules are prone to combinatorial explosions of rule activations – especially in measures that heavily utilize `OR` clauses. The `exists` Drools operator would usually alleviate this, but a unique aspect of the QDM measures called *Specific Occurrences* makes this difficult. A *Specific Occurrence* is the notion of assigning an identifier to a specific event (such as a Diagnosis or Encounter), and referring to that specific event elsewhere in the measure. To account for this, we follow the *Specific Occurrence* processing algorithm proposed by the MITRE Corporation[26]. Execution time performance metrics are also limited by the small Project Cypress patient set size, and as such, a larger test patient set would be more appropriate for scalability and performance evaluations.

Future Work. In general, optimizing the Drools Rule compilation has the potential to increase scalability and performance to even higher levels, as well as to provide more readable and concise rules. Also, the relationship between Complexity in Table 1 and Execution Time in Figure 3 is not fully understood. If a correlation exists, it has not been statistically analyzed. Furthermore, asserting relative complexity rankings between measures is error-prone, as algorithms and metrics for capturing true measure "complexity" are not known. More analysis is needed to explore how this complexity can be measured. Along these lines, Figure 4 suggests that the number of generated Drools Rules – not complexity – may be a more accurate model of execution time. If the correlation does indeed hold, future work in Drools Rule compilation optimizations may be approached with that model in mind.

Acknowledgement. This work has been supported in part by funding from the National Institutes of Health (R01-GM105688).

References

[1] Omri Gottesman, Helena Kuivaniemi, Gerard Tromp, W Andrew Faucett, Rongling Li, Teri A Manolio, Saskia C Sanderson, Joseph Kannry, Randi Zinberg, Melissa A Basford, et al. The Electronic Medical Records and Genomics (eMERGE) network: past, present, and future. *Genetics in Medicine*, 15(10):761–771, 2013.

[2] Jyotishman Pathak, Kent R Bailey, Calvin E Beebe, Steven Bethard, David S Carrell, Pei J Chen, Dmitriy Dligach, Cory M Endle, Lacey A Hart, Peter J Haug, et al. Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium. *Journal of the American Medical Informatics Association*, 20(e2):e341–e348, 2013.

[3] Susan Rea, Jyotishman Pathak, Guergana Savova, Thomas A Oniki, Les Westberg, Calvin E Beebe, Cui Tao, Craig G Parker, Peter J Haug, Stanley M Huff, et al. Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: The SHARPn project. *Journal of Biomedical Informatics*, 45(4):763–771, 2012.

[4] Electronic Specifications for Clinical Quality Measures. `http://www.cms.gov/`

Regulations-and-Guidance/Legislation/EHRIncentivePrograms/Electronic_Reporting_Spec.html, April 2011.

[5] Paul C Fu Jr, Daniel Rosenthal, Joshua M Pevnick, and Floyd Eisenberg. The impact of emerging standards adoption on automated quality reporting. *Journal of Biomedical Informatics*, 45(4):772–781, 2012.

[6] Common Terminology Services 2 (CTS2). http://www.omg.org/spec/CTS2/1.0, 2012.

[7] Michal Bali. *Drools JBoss Rules 5. X Developer's Guide*. Packt Publishing Ltd, 2013.

[8] Douglas Crockford. The application/json media type for Javascript Object Notation (JSON). 2006.

[9] Diana Behilng, Didi Davis, Gennady Sherman, J Michael Fitzmaurice, Jeffrey Klann, Julie Steele, Jyothi Mallampalli, and Lang. Quality Data Model based Health Quality Measures Format (eMeasure) implementation guide, release 1 (US realm) based on HL7 HQMF release 2.0.

[10] Quality Data Model technical specification. http://www.qualityforum.org/Projects/n-r/Quality_Data_Model/QDM_3_0_Technical_Specification.aspx, April 2011.

[11] HL7 Version 3 Standard: Representation of the Health Quality Measure Format (eMeasure) DSTU, Release 2E. http://www.hl7.org/implement/standards/product_brief.cfm?product_id=97#ImpGuides, 2014.

[12] Dingcheng Li, Cory M Endle, Sahana Murthy, Craig Stancl, Dale Suesse, Davide Sottara, Stanley M Huff, Christopher G Chute, and Jyotishman Pathak. Modeling and executing electronic health records driven phenotyping algorithms using the NQF Quality Data Model and JBoss® Drools engine. In *AMIA Annual Symposium Proceedings*, volume 2012, page 532. American Medical Informatics Association, 2012.

[13] Liam O'Brien, Paulo Merson, and Len Bass. Quality attributes for service-oriented architectures. In *Proceedings of the International Workshop on Systems Development in SOA Environments*, page 3. IEEE, 2007.

[14] Kevin J Sullivan, William G Griswold, Yuanfang Cai, and Ben Hallen. The structure and value of modularity in software design. In *ACM SIGSOFT Software Engineering Notes*, volume 26, pages 99–108. ACM, 2001.

[15] Cui Tao, Dingcheng Li, Feichen Shen, Zonghui Lian, Jyotishman Pathak, Hongfang Liu, and Christopher G Chute. Phenotyping on EHR data using OWL and semantic web technologies. In *Smart Health*, pages 31–32. Springer, 2013.

[16] David Garlan and Mary Shaw. An introduction to software architecture. *Advances in Software Engineering and Knowledge Engineering*, 1:1–40, 1993.

[17] Roy T Fielding and Richard N Taylor. Principled design of the modern web architecture. *ACM Transactions on Internet Technology (TOIT)*, 2(2):115–150, 2002.

[18] Erich Gamma, Richard Helm, Ralph Johnson, and John Vlissides. *Design patterns: Abstraction and Reuse of Object-Oriented Design*. Springer, 1993.

[19] Steven J Steindel. OIDs: How can I express you? Let me count the ways. *Journal of the American Medical Informatics Association*, 17(2):144–147, 2010.

[20] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

[21] Charles L Forgy. Rete: A fast algorithm for the many pattern/many object pattern match problem. *Artificial Intelligence*, 19(1):17–37, 1982.

[22] William K Thompson, Luke V Rasmussen, Jennifer A Pacheco, Peggy L Peissig, Joshua C Denny, Abel N Kho, Aaron Miller, and Jyotishman Pathak. An evaluation of the NQF Quality Data Model for representing electronic health record driven phenotyping algorithms. In *AMIA Annual Symposium Proceedings*, volume 2012, page 911. American Medical Informatics Association, 2012.

[23] Jessica Ross, Samson Tu, Simona Carini, and Ida Sim. Analysis of eligibility criteria complexity in clinical trials. *AMIA Summits on Translational Science Proceedings*, 2010:46, 2010.

[24] Project Cypress test procedures. http://www.healthit.gov/sites/default/files/cypress_test_procedure_11272013.pdf, February 2014.

[25] David S Janzen and Hossein Saiedian. Test-driven development: Concepts, taxonomy, and future direction. *Computer Science and Software Engineering*, page 33, 2005.

[26] Clinical quality measure logic and implementation guidance. http://projectcypress.org/documents/2014_eCQM_Specific_Occurrence_v14.pdf, 2013.

Knowledge Crystallization and Clinical Priorities: Evaluating How Physicians Collect and Synthesize Patient-Related Data

Ari H Pollack, MD[1,2], Carolyn G. Tweedy, BS[2], Katherine Blondon[3], MD, PhD, Wanda Pratt[2], PhD.
[1]Seattle Children's Hospital, Seattle, WA; [2]University of Washington, Seattle, WA; [3]University Hospitals of Geneva, Geneva, Switzerland.

Abstract

Information seeking and synthesis are time consuming processes for physicians. Although systems have the potential to simplify these tasks, future improvements must be based on an understanding of how physicians perform these tasks during clinical prioritization. We enrolled 23 physicians in semi-structured focus groups discussing simulated inpatient populations. Participants documented and discussed their data gathering and prioritization processes. Transcripts were coded to identify themes and generalized process flows. Results indicate that data are collected to categorize and prioritize patients according to expected clinical course. When data do not support these expectations, or when categorization indicates potential for morbidity, physicians increase efforts to act or re-categorize patients. Unexpected clinical changes have a significant impact on the decision-making and prioritization by clinicians. A modified version of the Knowledge Crystallization Framework helps to frame this work laying a foundation to advance information displays and facilitate information processing by physicians in clinical care environments.

Introduction

As increasing amounts of data become available within the medical record, Electronic Medical Record (EMR) systems need to prioritize and visualize that data to help clinicians process it in their time-constrained schedules. Clinicians perform many different tasks throughout the day—some are urgent and require immediate attention, while others can be postponed. The task of identifying priorities occurs continuously, and clinicians constantly re-evaluate their list of activities as new information surfaces. Ensuring that clinicians have all the data they need to make these decisions is a significant challenge. Today, clinicians must wade through volumes of information to find meaning, assign value, and take action. It has been estimated that physicians use approximately two million pieces of data when caring for patients.[1] This data includes both the acquired scholastic knowledge as well individual data for each patient. Some have said that "much of the art of medicine lies in gathering this information."[2] This practice needs to move away from artistry and self-discovery to a more rigorous process that limits the potential for error. The artistry within medicine should present itself in the interpretation and application of data, not the process of finding it.

It seems clear that access to additional data can lead to more informed prioritization and improved clinical outcomes; however, the costs associated with gathering and processing these data impact both clinicians and patients. The manual processes of searching for information and analyzing it to understand trends or see patterns involves too much time, especially as patients and their problems increase in complexity. Ideally, all the relevant data required to provide clinical care should be available at the time of need. In reality, the most critical data might not be readily accessible. Therefore, a clinician must use their judgment to decide when they have enough data to proceed with analysis and decision-making, as the cost of continual searching becomes too great.[3,4]

In this exploratory study, we examined how physicians collect, process, and utilize data during the clinical prioritization process. We sought to model this process and examine how our model compares to the knowledge crystallization framework previously described by Card et al., which was proposed as a means for understanding and amplifying cognition during information-seeking activities.[5] Our goal was to identify opportunities to improve the delivery and utilization of information in the clinical care environment and future medical information systems.

Methods

Recruitment & Focus Group Structure

This study was reviewed and approved by the Institutional Review Board at Seattle Children's Hospital (SCH). Our target subject enrollment was 16 to 24 physicians. In order to be included in this study, we required that physicians be credentialed at SCH at either the attending (supervising physician) or fellow (sub-specialty trainee) level, and that they provided acute clinical services to hospitalized patients within the four months leading up to their scheduled

focus group session. Resident physicians commonly work in shifts at SCH and their prioritization process is heavily influenced by the handoff from the previous resident on duty, hence they were excluded from our recruitment efforts. We also excluded surgical providers given differences in workflows between the medical and surgical specialties.

Recruitment letters were sent to individual medical staff members. Respondents were scheduled to participate in one of five focus groups, each with 4 to 5 participants. Each subject participated in only one focus group, and we ensured that each focus group had a diverse group of participants, not limited to a single medical specialty.

All participant interactions, with the exception of recruitment, occurred during the focus groups, which took place between September and December 2013. To ensure balanced participation by all participants, each focus group was led by a focus group facilitator from our research team. Each session was 90 minutes in length and leveraged a semi-structured format divided into two parts.

During the first half of each session, participants received a series of six fictional cases (Table 1) which represented common clinical presentations and were designed to simulate a typical inpatient population. Cases were written in purposefully vague language to allow for the potential of a large differential diagnosis. Using the cases to stimulate thought, the participants wrote responses to a series of questions (Table 2) on individual worksheets. Participants were asked to provide basic demographic information on this worksheet, including age, gender, professional level (attending or fellow), medical specialty, and year of completion in medical school and graduate medical education.

Table 1. Focus group cases with admitting or presentation diagnosis

1. **Asthma**: Allison is 3 year old female with a history of 2 days of upper respiratory symptoms, and 1 day of increased work of breathing. At presentation to the emergency department had a significantly elevated respiratory score, and partial response to inhaled bronchodilators.
2. **Hypertension and headaches**: Ben is a 15 year old previously healthy male with 1 month of progressively worsening headaches and general malaise, who is found to have a blood pressure of 192/120 mmHg.
3. **Cellulitis**: Alex is a 6 year old female previously healthy who sustained an insect bite 3 days ago on her lower leg. She presents with 1 day of increasing redness, swelling, and tenderness of her calf and now has streaking extending up to her thigh and a temperature up to 39.5 C.
4. **Gastroenteritis**: Sara is a 3 year old previously healthy female with 4 days of vomiting and diarrhea, afebrile, unable to tolerate any oral intake, and no appreciable urine output over the past 12 hours and lethargy.
5. **Failure to thrive:** Isaac is a 4 week old infant male likely full term (exacts dates unknown, due to little prenatal care) weighing 3.28 kg and at admission weighs only 3.3 kg.
6. **Abdominal pain:** Matt is an 11 year old previously healthy male with acute onset abdominal pain, nausea, and fever.

Table 2. Focus group worksheet writing prompts

1. Describe the process you would use to prioritize this group of patients. In your description please consider the data that is important in the process, where this data comes from, and how you acquire this data. Please do not rank the patients in these cases.
2. Please describe how data in different situations may lead to different prioritizations. For example a fever in a previously healthy 6 year old with upper respiratory symptoms, is different from a fever in a 6 year old on hemodialysis with a central line, which is different from a fever in 6 year old with acute lymphoblastic leukemia and neutropenia. Think about how you contextualize and bundle data elements to help with the prioritization process.
3. Please list any challenges you might experience during the process you described above.

In the second part of the focus group, individual participants presented and discussed their written responses with the focus group. We recorded each focus group session with a digital audio recorder. The audio files and worksheets were transcribed for later processing and analysis.

Data Analysis

All participant data was de-identified prior to analysis. We coded focus group transcripts to identify themes and map out individual process flows described in the focus groups. Coding was done iteratively, and focused on participants' description of their prioritization processes. We made reference to the worksheets to verify patterns and process flows, and sought to explain differences in said flows using the demographic information provided by participants.

Results

Participants

We recruited 23 physicians—19 attending physicians and four fellows—representing six different medical specialties (Table 3). The participants from sub-specialty services provide care to patients on their own medical service and to other services in a consultative fashion, with the exception of five participants who were Infectious Disease physicians that provide only consultative professional services.

Table 3. Provider Characteristics

	Fellow	Attending
Number	4	19
Female n (%)	1 (25)	8 (42)
Age Range	30 -39	30 -59
Average number of years since completing medical school (range)	4.5 (4-5)	18.1 (7-31)
Average number of years since completing graduate medical education (range)	NA	10.3 (0-26)
Medical specialties (n)	General Pediatrics (4) Gastroenterology (3) Infectious Disease (5) Nephrology (7) Pulmonology (1) Rheumatology (3)	

After mapping out the process described by each participant, it became clear that physicians consistently perform the same activities during their prioritization processes. Two distinct, though similar, prioritization workflows emerged from this analysis (Figure 1).

Workflows

The two workflows we identified contain similar elements: categorizing each patient, gathering data, and testing the fit of the collected data with the working diagnosis. The distinction between the workflows seems to correlate with a physician's identification as either one with primary patient responsibilities or one with only consultative services. Both groups employ a tacit understanding of their patients and categorize patients based on this understanding. However, providers with primary responsibility for their patients prioritize their patients based on level of concern, which determines the order of data collection for these patients. Providers collect data, test the fit of that data within their existing mental model, and then reprioritize their work based on this test of fit. It appears as though consulting providers (without primary patient responsibilities) do not prioritize their patients based on their initial clinical categorization, but gather data first, then prioritize patients based upon the fit of data within their working diagnosis.

Here we describe the two workflows using scenarios that capture the essence of participants' processes, followed by a series of representative quotes from the focus groups.

Workflow 1: Primary providers

When I come into the hospital in the morning, I begin with the patients on my service for which I have the most concern. Typically these are the sickest patients, but not always. Sometimes I am most worried about the patients who I just haven't figured out or who are not behaving as expected, based on the information I have at the time I start my day. This data comes from a variety of sources, including communications received from other providers on the care team as well as the medical record. However, at this point I am mostly prioritizing my patients based on my own tacit understanding and categorization.

Once I have identified the patients I am most concerned about, I begin collecting additional data to investigate how that patient has changed since I last saw them. I am less concerned with absolute values (for example, an elevated creatinine level) since many were abnormal in the first place. The new data either confirms my working diagnosis or identifies inconsistencies that need further exploration and explanation. I am reassured by patients who are progressing as expected, even if they are clinically sicker, but receiving appropriate therapy. When I identify a patient who doesn't seem to fit within my mental model or categorization, I become more concerned and this patient escalates on my prioritization list. This leads me to gather additional data for this patient in an attempt to identify the correct diagnosis and ultimately provide appropriately-targeted therapy. Their response to therapy provides additional data that either supports or conflicts with my understanding of their disease process and accordingly affects my level of concern and ultimately my prioritization. Once I believe my patient to be receiving the correct therapy, I usually move on to the next patient.

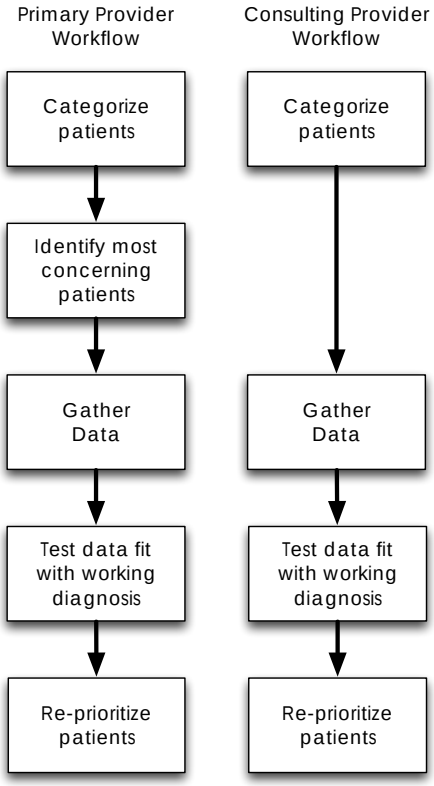


Figure 1. Workflow elements and order for primary and consultant providers.

Supporting examples from focus groups:

Participant 3.2: "…I prioritize first based on if the patient's on my service because then I'm primarily responsible as opposed to just being a consultant, and then if I'm primarily responsible, then I guess I kind of think of … the least stable or the most sick…. [the sickest patient is] unpredictable – the patient has been primarily unpredictable while they've been there, they're not following a course we would like them to follow…"

Participant 3.1: "…we inherently are constantly thinking about the sickest patient more than anybody else, so when I come into the door, I automatically will open their chart up first…"

Participant 4.2: "…if anything catches my eye that would be worrisome, I might do a little bit more investigating."

Participant 5.5: "I usually want to know exactly what interventions have been tried already, because that also modifies my probability of what's going on…and kind of like just how certain we are about things, and certainly there [are] vital sign changes or [as] abnormal lab results come back, then that changes my diagnostic impression."

Workflow 2: Representative of physicians with only consultative service responsibilities

Since I am not the primary responsible provider for the patients I am providing care to, I am less focused on their clinical acuity as I expect their primary service to address any urgent needs. Therefore, I typically begin my day with my list of patients, sorted alphabetically by last name or by location within the hospital, and then start at the top of that list and work my way down collecting new data for each patient, for example microbiology results or fever curves. As I collect these data, I am mentally processing to ensure that it fits with my understanding of the patient. I am reassured when the data fits my mental model, and more concerned when it does not. When the data doesn't fit, I spend time searching for and collecting data to help me identify the correct diagnosis. I am also concerned when I feel that the patient isn't receiving appropriate therapy, and I work with the primary team to explore why.

Supporting example from focus groups:

> *Participant 5.3: "I think a couple of things that are really different about an ID consultant. One is this idea that – ah, someone else will take care of the patient. I'm totally not interested in: "are they coding at the moment"… there's other people to deal with that, so I'm expecting that if they have questions for us, they'll call us. So I'm really more concerned with more bigger picture stuff…And so trying to keep track of all of that can be one of the bigger challenges…What I tend to do when I come in…I look at everything, because what I found was that unless I looked at seven different sources of information, there were gaps."*

Prioritization: Acuity vs. Change

When asked how they prioritize their patients, almost all focus group participants discussed invoking the concept of "sick vs. not sick" during their patient care regimes. This concept is thought to have been first used to quickly triage injured soldiers during the French Revolution, and is widely attributed to Dominique Jean Larrey, a surgeon who attended to Napoleon's Imperial Guards.[6] Participants in our focus groups had a difficult time defining this concept. When pushed to explain what they meant by "sick vs. not sick," participants ultimately agreed

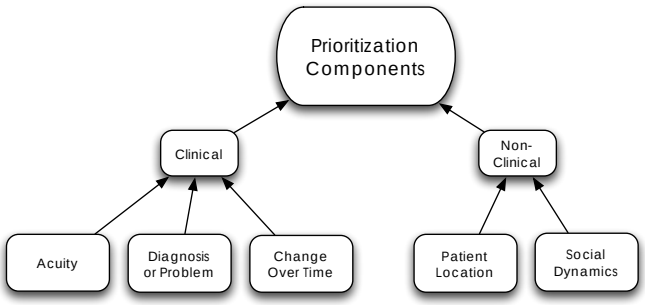


Figure 2. The components used by clinicians in the prioritization process contain both clinical and non-clinical elements.

upon a process rather than a definition; a process performed by experts with expert knowledge. Physicians often use experience and intuition in conjunction with objective data to decide who is sick. Most participants cited clinical acuity as the basis for the distinction, including a variety of data points most commonly including vital signs and laboratory values. Clearly, clinical acuity plays a role in the prioritization process, but changes or trends seem to play a vital role in this process, as well (Figure 2). Changes in either objective or subjective data elements—both expected changes and unexpected changes— appear to facilitate the distinction between patients who are sick, and those who are not. Quotes from the focus group participants explain this best:

> *Participant 2.2: "So I'm looking for patterns that I recognize that I'm supposed to be seeing in patients in general based on experience. I'm looking for how the patterns are not following what I'm expecting based on my knowledge of the type of patient … And I'm looking for a change in the pattern or in the trend which maybe I wasn't expecting…And that stops me in my tracks, why is this not making sense? … That's when everything stops and that's when I really want to be able to have the ability to grab around all the data, the numbers, the written words, the commentary which is hard to find, the imaging, the everything, and just sort of say why does this not make sense?"*
>
> *Participant 2.4: "So if you're sure of the diagnosis and you're sure they're on the right therapy even if they're worse, not such a big deal. If all of a sudden you say, 'Wait, this doesn't make sense anymore.' You have to go back to square one…Are we treating the wrong thing?"*
>
> *Participant 2.2: "… We have plenty of patients who get worse, but sometimes that's part of pattern. And that actually doesn't freak you out anywhere near as much as when somebody gets even just a little bit worse, but that's not what they're supposed to do. It's when they do what they're not supposed to do that you think something is weird going on, you have to regroup."*
>
> *Participant 4.1 "Change, yeah, a new vital sign change, new critical lab value. I'm trying to get at that decompensation potential…"*

Our findings suggest that physicians express elevated levels of concern and anxiety for patients that do not follow an expected clinical course. Upon encountering this scenario, participants indicate a heightened level of concern, which leads to increased vigilance for those patients who have the potential to acutely decompensate. In order to uncover this type of discrepancy, participants indicated that they rely on changes in clinical outcomes and did not find absolute values as helpful.

Challenges

Participants expressed many concerns and difficulties, relating to both data collection and processing, when trying to prioritize their patients (Table 4 and Table 5). Participants identified missing data as a significant problem. A variety of factors contribute to missing data—it might not have been collected in the first place, or it might have been collected but stored in an alternative location (e.g. lab result in an outside record). Many participants echoed this sentiment and expressed frustration with data sources that are inaccessible or hard-to-access, especially data retained by a nurse, consultant, or patient.

In addition to missing data, another major challenge identified during the focus groups related to the time required to collect relevant data. Participants reported the need to search across multiple sources, oftentimes in a variety of physical locations. The need to search through voluminous data sources and ultimately filter out non-relevant segments significantly contributed to the amount of time participants required to complete the data collection process. Other challenges encountered related to data quality as well as technical challenges with interacting with

Table 4. Challenges with Data Collection

1. Missing data
   - Not collected
   - Collected but not resulted
   - Collected but not available or not found
     - Data in outside record
     - Source (i.e. resident, nurse, consultant, patient, care provider) not accessible
     - Most recent results not available
   - Available data not recognized as valuable
2. Communication between care team members did not occur
3. Noise
   - Filtering unimportant data
4. Time Consuming
   - Searching
   - Waiting (related to availability of resources)
   - Need to check multiple sources in a variety of locations
5. Quality concerns
   - Data gatherers have varying skill levels
   - Language barriers
6. EHR
   - Speed
   - Interface

Table 5. Challenges with Data Processing

1. Bias
   - Anchoring
   - Bandwagon
   - Confirmation
   - Premature Closure
   - Presentation
   - Source value
2. Time Consuming
3. Multitasking
4. Different focuses between stakeholders (physician, nurse, patient, caregiver)
5. Unexpected results
6. Discovering trends especially across data types
7. Framing, putting results in context
   - Normal vs abnormal
   - Better or worse
   - Over time

the electronic medical record.

During the focus group discussions, participants also indicated that they encounter difficulties processing information after it has been collected. Many of these challenges concerned cognitive biases that interfere with the participant's ability to appropriately process and interpret available data. For example, some participants expressed challenges with incorporating new data into existing mental models, especially when the new data did not fit or was unexpected. Despite the quantity or quality of data that may be available within the medical record, a clinician may not always think to gather all potentially relevant data, or may even miss an important test result, due to an error in their clinical reasoning or potential retrieval issues.

Other themes that emerged around the discussion of data processing were difficulties with trending data over time, across data types, and putting it in the appropriate context.

Discussion

Clinicians spend much of their time working with information: searching, acquiring, processing, framing, and acting on information with the goal of providing the best patient care. Most clinicians in a hospital setting are responsible for providing care to multiple patients at once. Given the multitude of competing needs, clinicians must constantly re-prioritize their work to ensure the most urgent needs are met first. This process might best be described as a "knowledge crystallization" task, which was first described by Card et al. as a task "in which a person gathers information for some purpose, makes sense of it by constructing a representational framework (referred to as a schema), and then packages it into some form for communication or action."[3]

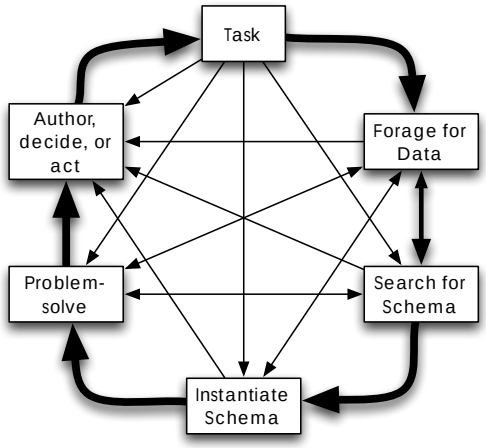


Figure 3. Knowledge Crystallization framework as proposed by Card, Mackinlay, and Shneiderman in 1999. This model aids information seekers by providing structure as well as context to their search process ultimately reducing the cognitive burden needed to accomplish a task. Typically, one moves around the outside in a counterclockwise fashion, but can jump to other steps as appropriate.

Knowledge Crystallization

Within the knowledge crystallization framework (Figure 3), a knowledge seeker typically moves from a task to gathering data to finding and instantiating an appropriate schema that allows the individual to address the task and act appropriately. Card et al. describe this process in detail, and specify three required elements for knowledge crystallization: data, tasks, and schemas.[3] The value of the knowledge crystallization model lies in its ability to improve the efficiency of acquiring and processing data in order to accomplish a stated task. Initially envisioned as a tool to create schemas for information visualization models, use of knowledge crystallization extends to most information seeking activities.

At a high level, the prioritization process used by clinicians during patient care activities seems to model the knowledge crystallization framework. However, the process revealed by physicians in our focus groups indicates a slight—though notable—difference.

When caring for patients, clinicians have a primary task: provide the best and most appropriate therapy in a timely fashion. To accomplish this task, clinicians systematically collect targeted data. To be effective, clinicians begin with a mental model or schema they have deemed appropriate to their current situation, and they ask questions and seek answers that test the validity of this schema (in the field of clinical reasoning, this is analogous to the hypothetico-deductive approach, where experts use co-selection in order to test a hypothesis).[7,8] This process continues until the clinicians reach a point where they are comfortable enough to act and deliver appropriate care.

Therefore, we cannot accept the knowledge crystallization framework wholesale without making modifications. Recognizing and understanding the difference between the knowledge crystallization framework and a physician's prioritization process allows us to propose a new model which highlights the importance of the schema for clinicians while providing patient care. The new model resulting from this research (Figure 4) incorporates the concept of the schema and how the prioritization process is influenced by the collection and processing of data.

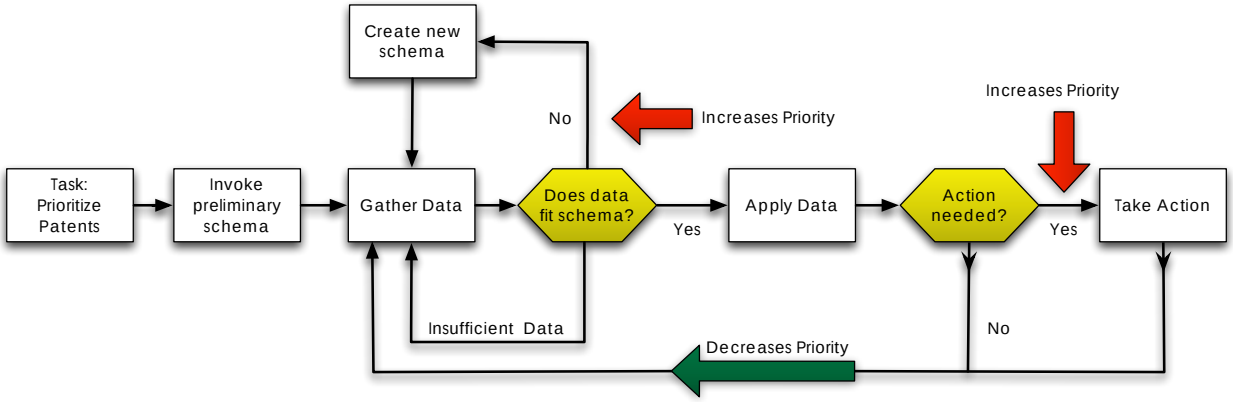


Figure 4. Modified knowledge crystallization model addresses the task of prioritizing patients. Once clinicians have invoked their preliminary schema, categorizing each patient, they gather data and asses the fit of the data with their schema. Poor-fitting data requires a new schema, increasing the priority of that patient.

Beginning with a schema prior to gathering data allows a clinician to proceed with purpose and direction, what some would refer to as the artistry in medicine.[9,10] When a physician determines the need to take action and provide care, therapeutic interventions appear to be based on their diagnostic categorization (i.e. schema). Depending on the acuity and urgency of the situation, clinicians may choose to act with a low level of diagnostic certainty to prevent further clinical deterioration. However, regardless of the degree of certainty with the working schema or diagnosis, instantiating a schema allows for constant evaluation and testing, generating refinements, increasing accuracy, and ultimately leads to the appropriate intervention for the patient.

Our model identifies the value of the schema and how clinicians use data to test its validity. Clinicians place high importance on the validity of their schema: when they uncover data that contradicts that schema, their level of concern for the patient, and therefore the patient's priority increases. We expected clinical acuity (i.e. sick vs. not sick) to be the major driver in the prioritization process; but this research suggests that classification errors leading to poor schema formation have equal, if not greater importance. Recognizing the importance of patient categorization and ultimately schema formation has the potential to influence future information delivery systems.

Implications

Tools designed with this new model in mind will have a greater capacity to dramatically improve a clinician's ability to provide care for their patients, and streamline their work processes. We first propose design principles to help guide the development of new clinical information system tools followed by specific recommendations where use of our model has the potential for positive change: workflow improvements to assist in patient categorization, and potential integrations with the burgeoning field of information visualization.

Design Principles

Our model provides multiple opportunities to improve clinical information systems. Designers should focus their development efforts highlighting information that contradicts the pre-formed schemas utilized by clinicians in the care process. These tools have the potential to highlight subtle and early changes in the clinical course of a patient, perhaps even before a clinician has recognized them. In addition, our research clearly demonstrated the importance and value clinicians put on expected and unexpected changes in the clinical courses their patients follow. Therefore, the development of new tools should focus on highlighting these types of changes, making it easier for clinicians to detect, process and act. Designers have the opportunity to focus their efforts on concepts such as change from expected, change over time, as well as relative changes and absolute changes.

Patient Categorization

Large data repositories—a result of the implementation of electronic health records at many hospitals—have tremendous potential for machine learning applications. Machine learning involves the process of applying an algorithm to accomplish a task, where "learning" is achieved through recursively repeating the task and improving upon an empirical metric. Clustering, a typical machine learning problem, sets out to group objects such that the objects within a group or cluster have similar traits. The process of assigning a diagnosis to a patient is analogous to cluster analysis, in that a physician identifies specific traits of an individual (e.g. fever, cough, chest pain) and identifies potential similar groups or diagnoses for that patient (e.g. upper respiratory infection, pneumonia or asthma). As the physician gains more information she adjusts the grouping to improve the fit, just as we have described in our model. There are examples of using cluster analysis to better understand the variability within a specific disease, but we are not aware of its use in clinical systems to help in real-time patient classification.[11] The physicians we interviewed agreed on the importance of recognizing when patients begin to deviate from their expected course, and it is clear that an automated cluster analysis tool built into the electronic medical record with the potential to highlight these changes would be tremendously valuable to physicians. Furthermore, given the biases that interfere with the medical decision-making process these tools have the potential to provide more objective evidence to providers during times of great need.[12]

Information Visualization

One could argue that the results of cluster analysis achieve their greatest potential when expressed visually. The human visual system quickly allows us to find patterns and associations within data sets when organized in a visual fashion.[13] Cluster analysis exploits this trait to discover patterns and groupings based on specified attributes.

Imagine a clinical system that visually represents groups or clusters of patients based on historical clinical attributes from within the medical record. Patients of an individual physician could be overlaid on this display, allowing the physician to quickly identify appropriate groupings for his or her patients. It would be possible to show how a patient's affinity towards different groupings changes over time as new data becomes available. Our results demonstrate the importance of categorization and schema formation within clinical care, specifically how inaccurate categorization of patients leads to increased concern from medical providers. Visual presentation of cluster analysis has incredible potential to help physicians recognize subtle changes in a patient's clinical course more quickly, reducing this anxiety.

We recognize that this vision for cluster analysis within the medical record will be challenging to implement and may be in the future but information visualization still has great potential with much simpler tools using our model as a basis for presenting information to clinicians. For example, visualizations that identify changes in patient variables over time would be easier to implement, and still offer tremendous benefit to physicians. Wang et al. indicate that tools have been developed to visualize event time sequences that provide clinicians and administrators insight into how the order of events affect overall patient care.[14,15] Leveraging tools to highlight changes in clinical outcomes across different data points, like changes from baseline, changes from normal, or changes from an expected course, support the task of clinical prioritization. Most electronic systems provide this basic capability via result trending over time, but it is easy to imagine how more intricate displays spanning disparate data sources could help physicians identify patients with more urgent needs. As indicated in our focus groups, physicians tend to sort their patient list by demographic information. A patient list with sort options encompassing the degree of change over the last 6 or 12 hours based on vital signs and lab values would help physicians quickly identify patients with the highest degree of change among their population of patients. Combined with a physician's tacit knowledge of the expected course for an individual patient, this tool could help identify patients with unexpected deviations in their clinical course.

These examples leverage Shneiderman's mantra for information visualization: overview first, zoom and filter, and then details on demand, which would provide clinicians with a quick, but highly interpretable snapshot of their patients' statuses.[16] Although clinicians often create mental models to help process this information, visualization techniques could help lighten cognitive processing required by physicians as well as provide new insights into data available within the medical record.[17,18] Chittaro describes a series of techniques and goals for visualizing information and applying them to the medical domain.[19] However, none of these tools help to prioritize information within a single record, nor across multiple patient records. Leveraging information visualization to highlight the points in our model that increase a patient's priority has the potential to dramatically improve physicians' ability to prioritize all of their patients.

Strengths and Limitations

Too often the evaluation of healthcare information technology occurs late in the software development cycle, making changes much more difficult and costly to implement.[20] In addition these late evaluations base their system design on assumptions of stakeholder needs and workflows. Conversely, our work explored understanding the fundamentals of information seeking and processing for practicing physicians. This knowledge should allow clinical information system designers to develop systems that better target the needs of clinicians. Because our study based its findings on a diverse sample of medical providers, the results should be generalizable across a broad population of providers.

Despite the diversity of medical providers within our sample, all of the participants were pediatric providers at a single organization, which might diminish its generalizability. Nonetheless, we would argue that pediatric providers approach the task of clinical prioritization in the same manner that internal medicine or family practitioners do. In addition, the scenarios utilized in the focus groups only included hospitalized patients, excluding the outpatient setting, though most of the participants in the study provide care in both settings. One could easily repeat this study and include an expanded group of clinicians in a variety of settings to ensure the broader applicability, especially for ambulatory patients where care extends over weeks to months and even years.

Conclusion

Physicians expend a significant amount of time and mental effort working with data, finding context, and ultimately transforming information into knowledge, which proves to be a tremendously costly endeavor. Understanding the cost structure of information seeking has broad implications in healthcare information systems. Development of these systems should be based on a thorough understanding of clinical knowledge acquisition and processing as well as the costs of these procedures.

The first step in reducing these costs requires us to understand how physicians approach the journey of knowledge discovery and identify points or situations that have the potential to increase these costs. We present a model that describes the information seeking behavior used by physicians to prioritize patients in a hospital setting, highlighting the importance of accurate clinical categorization in this process. Our goal for this research was to inform the design and creation of novel tools or systems that focus on the data and information required during the prioritization process and improve their delivery. Through machine learning applications and information visualization, our model has great potential to reduce the costs of knowledge acquisition in healthcare settings and as a result improve clinical outcomes.

Acknowledgments

We would like to thank all participants and funding sources (NIH T32 DK007662) and Dr. David Hendry for his support in this endeavor.

References

1. Pauker SG, Gorry GA, Kassirer JP, Schwartz WB. Towards the simulation of clinical cognition. The American Journal of Medicine. 1976 Jun;60(7):981–96.
2. Smith R. What clinical information do doctors need? BMJ : British Medical Journal. BMJ Group; 1996 Oct 26;313(7064):1062.
3. Card, S. K., Pirolli, P., & Mackinlay, J. D. (1994). The cost-of-knowledge characteristic function: display evaluation for direct-walk dynamic information visualizations, 238–244. doi:10.1145/191666.191753
4. Pirolli, P., & Card, S. (1999). Information foraging. Psychological Review, 106(4), 643–675. 3. Card SK, Mackinlay JD, Shneiderman B. Readings in Information Visualization: Using Vision to Think. 1999.
5. Card SK, Mackinlay JD, Shneiderman B. *Readings in Information Visualization: Using Vision to Think*. 1999.
6. Shapiro N, Howell MD. Sick? Or, not sick?*. Critical Care Medicine. 2005 May 1;33(5):1151.
7. Schmidt HG, Norman GR, Boshuizen HP. A cognitive perspective on medical expertise: theory and implication [published erratum appears in Acad Med 1992 Apr;67(4):287]. Academic Medicine. 1990 Oct 1;65(10):611.
8. Elstein AS, Schwarz A. Evidence Base Of Clinical Diagnosis: Clinical Problem Solving And Diagnostic Decision Making: Selective Review Of The Cognitive Literature. BMJ : British Medical Journal. BMJ; 2002 Mar 23;324(7339):729–32.

9. Schon DA. Educating the reflective practitioner: toward a new design for teaching and learning in the professions. 1st ed. San Francisco: Jossey-Bass; 1987. 1 p.
10. Montgomery K. How Doctors Think: Clinical Judgment and the Practice of Medicine. Oxford University Press; 2005. 1 p.
11. Haldar P, Pavord ID, Shaw DE, Berry MA, Thomas M, Brightling CE, et al. Cluster analysis and clinical asthma phenotypes. Am J Respir Crit Care Med. 2008 Aug 1;178(3):218–24.
12. Dawson NV, Arkes HR. Systematic errors in medical decision making: judgment limitations. J Gen Intern Med. 1987 May;2(3):183–7.
13. Tufte ER. The visual display of quantitative information. Second. Cheshire, Conn.: Graphics Press; 2001.
14. Wang TD, Plaisant C, Shneiderman B, Spring N, Roseman D, Marchand G, et al. Temporal Summaries: Supporting Temporal Categorical Searching, Aggregation and Comparison. IEEE Trans Visual Comput Graphics. 2009;15(6):1049–56.
15. Wang TD, Wongsuphasawat K, Plaisant C, shneiderman B. Visual information seeking in multiple electronic health records. New York, New York, USA: ACM Press; 2010. pp. 46–55.
16. Shneiderman B. The Eyes Have It: A Task by Data Type Taxonomy for Infromation Visualizations. Visual Languages. 1996;:336–43.
17. Kang Y-A, Görg C, Stasko J. How Can Visual Analytics Assist Investigative Analysis? Design Implications from an Evaluation. IEEE Trans Visual Comput Graphics. 2010 Jun 2;:570–83.
18. Liu Z, Stasko JT. Mental models, visual reasoning and interaction in information visualization: a top-down perspective. IEEE Trans Visual Comput Graphics. 2010 Nov;16(6):999–1008.
19. Chittaro L. Information visualization and its application to medicine. Artif Intell Med. 2001 May;22(2):81–8.
20. Yen P-Y, Bakken S. Review of health information technology usability study methodologies. J Am Med Inform Assoc. 2012 May;19(3):413–22.

Providing Hospital Patients with Access to Their Medical Records

Jennifer E. Prey, MS[1], Susan Restaino, MD[2,3], David K. Vawdrey, PhD[1]
[1]Department of Biomedical Informatics, Columbia University, New York, NY; [2]NewYork-Presbyterian Hospital, New York, NY; [3]Department of Medicine, Columbia University, New York, NY

Abstract

Being a hospital patient can be isolating and anxiety-inducing. We conducted two experiments to better understand clinician and patient perceptions about giving patients access to their medical records during hospital encounters. The first experiment, a survey of physicians, nurses, and other care providers (N=53), showed that most respondents were comfortable with the idea of providing patients with their clinical information. Some expressed reservations that patients might misunderstand information and become unnecessarily alarmed or offended. In the second experiment, we provided eight hospital patients with a daily copy of their full medical record—including physician notes and diagnostic test results. From semi-structured interviews with seven of these patients, we found that they perceived the information as highly useful even if they did not fully understand complex medical terms. Our results suggest that increased patient information sharing in the inpatient setting is beneficial and desirable to patients, and generally acceptable to clinicians.

Introduction

Being in the hospital has been called "one of the most dis-empowering situations one can experience in modern society."[1] Patients often feel isolated, anxious, and that they do not have control over their care. They often do not know what medications they are taking, their treatment plans, or even the names of members of their care teams.[2–4] Therefore, it is important to try to engage hospital patients and facilitate greater participation in their care.

Engaging patients has received significant attention in recent years. In contrast to the traditionally paternalistic doctor-patient relationship, consumers increasingly expect direct and fast access to their health records.[5] This engagement of patients not only helps the patients feel more involved, but also can lead to improved health outcomes.[6,7] The journal *Health Affairs* recently referred to patient engagement as equivalent to the next "blockbuster drug".[8] The Institute of Medicine (IOM) recommended that individuals receive opportunities to have access to medical information and clinical knowledge which will enable the patients to be the "source of control".[9] The U.S. Meaningful Use financial incentive program stipulates requirements for patients having access to clinical summaries, electronic messaging with their providers, patient-specific educational resources and online access to their personal health information (including information like care team members, medication lists and history, laboratory results, and problem lists).[10] Further, the Consumer Assessment of Healthcare Providers and Systems (CAPHPS) Hospital Survey, which collects patients' perspectives of hospital care, impacts hospital reimbursement from the Centers for Medicare and Medicaid.[11]

Despite a growing interest in patient engagement, little work has been conducted to study patient engagement in the inpatient setting.[12–15] As noted in a recent New England Journal of Medicine article, access to information by patients is becoming more common, but providing access in the inpatient setting is particularly complex.[16] To our knowledge, no one has studied the impacts of providing patients in the hospital with full access to their medical record. This includes data like medications, laboratory results, radiology and pathology reports, clinical summaries, operative/procedure reports and progress notes. We wanted to understand the clinician perspective on increased patient access to information as well as study the patient experience in receiving greater access to their information while in the hospital. Our hope was that by providing patients with access to their full clinical chart, they would feel more engaged in their care.

Methods

This research consisted of two parts; a survey administered to clinicians as well as a complementary field study conducted with patients in the hospital. In the field study, patients received daily printouts of their medical chart and then participated in semi-structured interviews to discuss their experiences. The study was conducted at Columbia

University Medical Center (CUMC), a large urban academic center that is part of NewYork-Presbyterian Hospital. This research was approved by the medical center's human subjects institutional review board.

Clinician Survey

A one-page survey was distributed to attendees of a staff meeting in a cardiac step-down unit, mostly comprised of nursing staff and physician assistants. It was also distributed at a monthly meeting of CUMC's Housestaff Quality Council, a group of resident physicians representing each clinical department at the medical center. Additional surveys were collected from clinical staff on the cardiology floor.

The survey consisted of two sections, one that focused on clinicians' perceptions of sharing different types of information with patients (e.g. medication information, clinical notes), and a second section that asked about potential consequences of increased patient access to information. Each of these items was evaluated on a 5-point Likert-type scale from 'Strongly Disagree' to 'Strongly Agree'. Questions used negatively and positively worded stems to guard against acquiescence.[17] The survey also provided additional space for participants to provide comments and encouraged them to elaborate on potential concerns, burdens, or benefits of increased patient access to information.

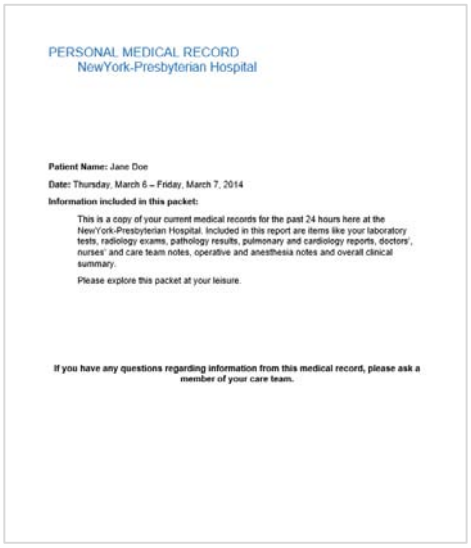
Responses to the survey were tabulated and analyzing using both Microsoft Excel 2010 and the R Statistical package.[18] A descriptive summary of the data consisted of calculating frequencies, medians, and inter-quartile ranges (IQRs). Kruskal-Wallis analysis was completed, and a Bonferroni correction used for post-hoc analysis.[19]

Inpatient Field Study

We provided cardiology patients with a daily printout of their hospital care, printed from the institution's electronic health record (EHR).[20] After four consecutive days of receiving this printout, we conducted in-situ semi-structured interviews with patients to elicit their feedback. Interviews were audio recorded, and transcripts of the audio were thematically coded.

Medically stable patients on a cardiology floor were approached for potential participation based on their ability to speak English and their anticipated date of discharge. In line with our institutional policies, their attending physician described the study and referred interested patients to the primary researcher (JP), who obtained informed consent.

After providing consent, the patient received a daily printout of his/her medical record, which included new information added to the EHR in the previous 24 hours. Specifically, the printout contained: laboratory test results, physician progress and consult notes, radiology reports, pathology reports, cardiology test results, discharge planning materials, operative reports, nutrition notes, the medication administration record, and other nursing documentation.


PERSONAL MEDICAL RECORD
NewYork-Presbyterian Hospital

Patient Name: Jane Doe

Date: Thursday, March 6 – Friday, March 7, 2014

Information included in this packet:

This is a copy of your current medical records for the past 24 hours here at the NewYork-Presbyterian Hospital. Included in this report are items like your laboratory tests, radiology exams, pathology results, pulmonary and cardiology reports, doctors', nurses' and care team notes, operative and anesthesia notes and overall clinical summary.

Please explore this packet at your leisure.

If you have any questions regarding information from this medical record, please ask a member of your care team.

Figure 1. Cover page for patient printout.

Minor editing was performed to remove phone numbers and social security numbers, and to delete unnecessary white space. No additional changes were made to the documents. A cover page describing the potential contents of the printout was added (Figure 1). The report was printed and hand-delivered each morning to each of the participants. The reports ranged from 5–40 pages each; 10–15 pages was typical.

After receiving daily reports for four days, a semi-structured interview was conducted. In addition to basic demographic questions, patients were asked to describe various experiences related to receiving their clinical information. Topics that were discussed included:

- Patients' use of the printout (e.g., frequency, time spent)
- Patients' perceived capacity to understand the information
- Preferences regarding structure/format of the information
- Patients' desire to receive such information in the future
- Behavior changes as a result of having the information (e.g., asking more questions)

Interviews were audio-recorded and transcribed. As done in previous studies by this research team,[13] transcripts were then thematically coded by the primary researcher (JP).[21]

Results

Clinician Survey

Approximately 65 surveys were distributed and 53 clinicians responded: 21 physicians (40%), 20 nurses (38%), 7 physician assistants (13%), and 5 allied health professionals (e.g., nutrition specialists, social workers) (9%). Responses are shown in Table 1 along with frequency, median, and IQRs in ordinal number format (1=Strongly Disagree; 5=Strongly Agree). Only the first question (sharing some information in general) had an IQR equal to zero. Six questions had an IQR of two. The Kruskal-Wallis test analyses differences between the groups (Physicians vs. PAs vs. Nurses vs. Other). Only the question on note-writing behavior was statistically significantly different between groups (p=0.033); however, post-hoc pairwise analysis of between group differences with a Bonferroni correction did not find any significance. The survey results are also presented using diverging stacked-bar graphs in Figures 2–4.[22]

Table 1. Summary statistics for clinician survey on attitudes toward inpatient information sharing.

	Disagree	Strongly Disagree	Neutral	Agree	Strongly Agree	No Response	Range	Q1	Median	Q3	IQR	Kruskal-Wallis
I am comfortable with patients having access to:												
Some information from their EHR (in general)	0 (0%)	0 (0%)	6 (13%)	28 (62%)	11 (24%)	9 (20%)	3-5	4	4	4	0	0.458
Medication Information	0 (0%)	0 (0%)	1 (2%)	23 (45%)	27 (53%)	3 (6%)	3-3	4	5	5	1	0.336
Care Team Profiles	3 (0%)	0 (6%)	8 (15%)	23 (44%)	18 (35%)	2 (4%)	2-5	4	4	5	1	0.419
Lab Results	2 (2%)	1 (4%)	3 (6%)	23 (44%)	23 (44%)	2 (4%)	1-5	4	4	5	1	0.511
Radiology Results	3 (2%)	1 (6%)	5 (10%)	21 (40%)	22 (42%)	2 (4%)	1-5	4	4	5	1	0.088
Pathology Results	2 (2%)	1 (4%)	3 (6%)	24 (46%)	22 (42%)	2 (4%)	1-5	4	4	5	1	0.211
Operative/Procedure Reports	1 (2%)	1 (2%)	6 (11%)	27 (51%)	18 (34%)	1 (2%)	1-5	4	4	5	1	0.345
Progress Notes	10 (6%)	3 (19%)	11 (21%)	19 (36%)	10 (19%)	1 (2%)	1-5	2	4	4	2	0.092
Consultation Notes	7 (4%)	2 (13%)	12 (23%)	20 (38%)	12 (23%)	1 (2%)	1-5	3	4	4	1	0.233
In my opinion, providing patients increased access to their medical information will result in:												
Increased time at bedside	9 (4%)	2 (17%)	15 (29%)	20 (38%)	6 (12%)	2 (4%)	1-5	3	3	4	1	0.970
Increased patient misunderstanding of information	15 (2%)	1 (28%)	11 (21%)	20 (38%)	6 (11%)	1 (2%)	1-5	2	3	4	2	0.541
Increased patient anxiety	14 (10%)	5 (27%)	13 (25%)	14 (27%)	6 (12%)	2 (4%)	1-5	2	3	4	2	0.638
Increased legal liability	11 (2%)	1 (21%)	16 (31%)	16 (31%)	8 (15%)	2 (4%)	1-5	3	3	4	1	0.784
More work for me as a caregiver	16 (0%)	0 (31%)	14 (27%)	18 (35%)	4 (8%)	2 (4%)	2-5	2	3	4	2	0.333
Increase in patient questions	5 (2%)	1 (10%)	17 (33%)	25 (49%)	3 (6%)	3 (6%)	1-5	3	4	4	1	0.198
Improved health behaviors	16 (4%)	2 (31%)	6 (12%)	26 (50%)	2 (4%)	2 (4%)	1-5	2	4	4	2	0.140
Enhanced patient-clinician communication	1 (0%)	0 (2%)	18 (35%)	30 (58%)	3 (6%)	2 (4%)	2-5	3	4	4	1	0.301
Increased patient engagement in decision making	2 (0%)	0 (4%)	16 (31%)	28 (54%)	6 (12%)	2 (4%)	2-5	3	4	4	1	0.570
I would change the way I write my clinical notes if I knew a patient could view them:	15 (4%)	2 (31%)	12 (25%)	15 (31%)	4 (8%)	6 (13%)	1-5	2	3	4	2	0.033*

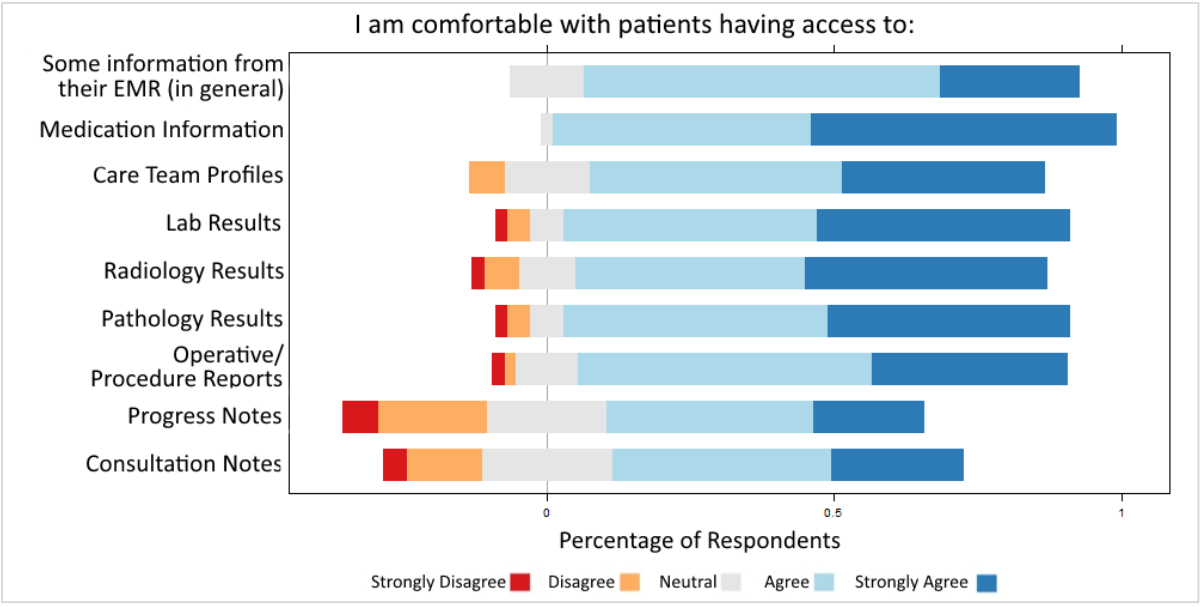


Figure 2. Clinician perspectives on patient information sharing.

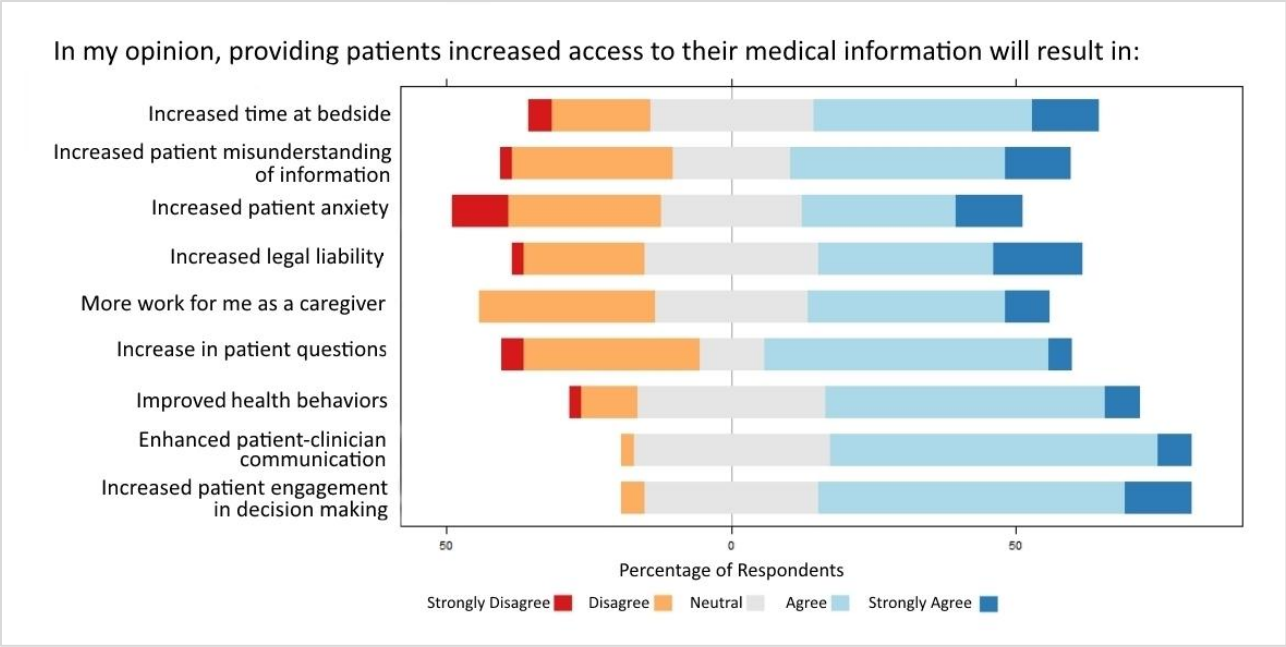


Figure 3. Clinician perspectives on consequences of increased patient information sharing.

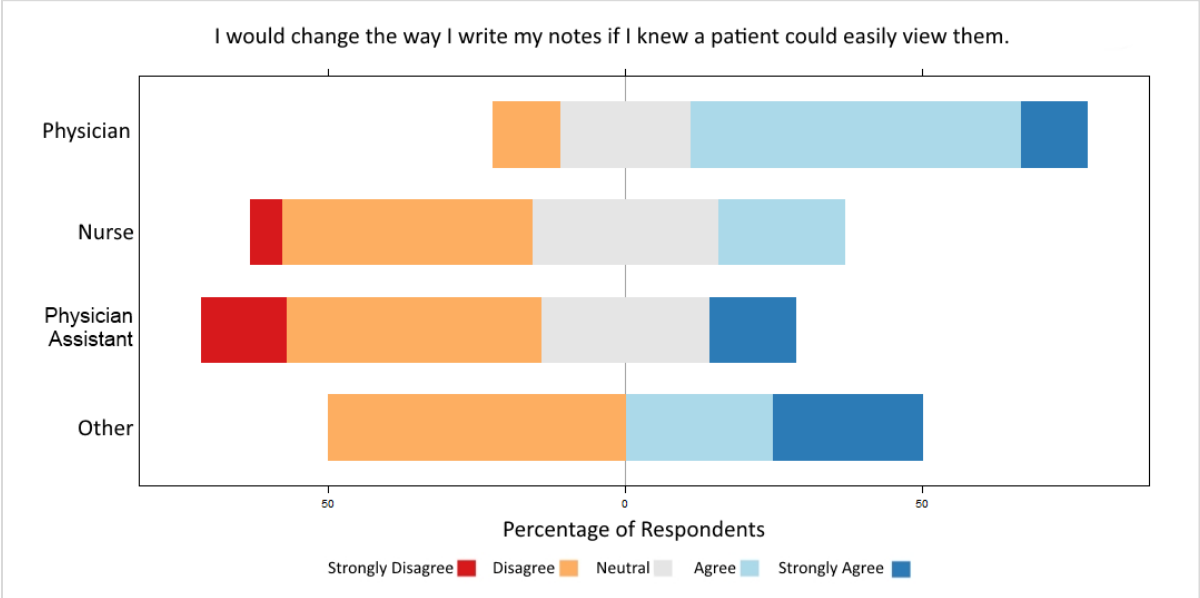


Figure 4. Clinician perspectives on expected changes in note-writing with increased patient information sharing.

Inpatient Field Study

Eight patients participated in the field study (Table 2). Seven completed the follow-up interview after receiving printouts of their hospital record for four days. The eighth patient was discharged prior to completing four days of receiving his records.

Table 2. Participant Demographics of Inpatient Field Study.

	Sex	*Age*	*Admission Reason*	*Presenting Chief Complaint*	*LOS*[*]	*Highest Level of Education*	*Reports Using Internet for Health Information*
P1	Female	87	Congestive heart failure exacerbation	Shortness of breath	6 days	High School	No
P2	Male	56	Diarrhea, atrial fibrillation	Diarrhea, shortness of breath	22 days	Some College	Yes
P3	Male	65	Septic Shock	N/A[†]	8 days	Graduate School	Yes
P4	Male	43	Myocardial infarction	Chest pain	4.5 months	College Graduate	No
P5	Male	35	Hemoptysis	Hemoptysis, shortness of breath	5 days	College Graduate	Yes
P6	Male	54	Awaiting Heart Transplant	N/A[‡]	5 weeks	Associates Degree	Yes
P7	Male	54	Ventricular tachycardia	Shortness of breath	5 weeks	Some College	No

[*]Length of stay at time of interview, [†]Transferred because of fevers from inpatient acute rehab, [‡]Admitted for heart transplant listing

Overall, patients' feedback on receiving the daily printouts was positive. Out of the seven patients we interviewed, six patients indicated that they would want to receive this type of information again if they had to return to the hospital. All six expressed a desire to continue receiving the printouts for the remainder of their current stay. Key themes that emerged from the semi-structured interviews were: variations in use, difficulty understanding medical jargon, appreciation of having the information, suggestions for improvement, changes in interaction with clinicians, and changes in engagement.

Variation in Use

There was a range in use of the printout. Two patients stated that they did not look at the printout (P5 and P7), P4 looked at it on two of the four days, and the remaining patients reported looking at it every day. The scope of use also varied. Some patients reported only skimming the information, while others read it in detail and even added their own annotations. For example, P2 commented: "I take a look, and if I see something that's not [in range], I just mark it."

Difficulty Understanding Medical Jargon

The printouts were created from the EHR without any substantive changes made to the content or format. Patients discussed having particular difficulty with certain medical terms and acronyms. P3 stated, "I read until I found technical stuff, and then I would jump over it…the parts I read, and I understood, were interesting and beneficial." P2 said, "Most of the things I understand, but some of the other things…like letters in combination, 'WT' or whatever… [were] not of use to me."

Several patients reported looking for specific items which they knew about. P1 specifically asked about her creatinine values and searched for the results of a sonogram performed the previous day. P5 expressed difficulty with understanding some material, but was planning to use the Internet to supplement his knowledge: "I'm familiar with some stuff," he said, "but…some of the terms I'm not familiar with, so once I have my laptop I'm probably going to take another look at it."

Appreciation in Having the Information

Patients expressed appreciation for having access to information about their care that they wouldn't normally receive. P7, who did not review the printouts at all, reported that it was important to him to "have a file." P2 said it was "very, very neat…because they give you all the details," and P5 stated: "Notes help a lot too, to see if the doctors are interpreting what I say… and they're pretty right on. And just [to see] what they thought as far as the possible diagnosis."

Suggestions for Improvement

In addition to wanting more explanation around medical terms and acronyms, patients expressed ideas regarding changes to the process that would make the intervention more helpful. For instance, P5 requested that there be a

more clear delineation of the author of notes: "Sometimes I see a bunch of comments, and I don't see exactly which doctor said it." P3 expressed the desire for more descriptive information: "If you could give me more information describing my condition and my situation…more than the technical stuff…more information as far as what the doctors are finding, and what the doctors are doing."

We also discussed the possibility of presenting the information to patients via a computerized interface. Most patients (four of the five who viewed the printouts) expressed an interest in viewing their information on a tablet computer or laptop.

Changes in Interaction with Clinicians

We asked patients if having access to their information prompted them to ask follow-up questions. Only one of the seven patients (P5) mentioned asking anything related to information in his printout. Several patients acknowledged the busy clinical workload of the providers and did not want to further impose on their limited time. P2 said, "[My doctors] have one or two minutes, [then] they're running all the time."

Changes in Engagement

Patients expressed feeling more informed about their health as a result of receiving the daily printouts. P2 mentioned that "if you don't ask, you don't receive [information], but with that paper, I am very on top of the situation." P5 said, "Instead of just listening to the doctor…I should pay attention to myself too" and believed the printouts could help him to do so.

In addition to feeling more informed, multiple patients mentioned the ability to act as a 'fact checker' of the information. P1 reported finding a prior procedure that was recorded with an incorrect date. P5 said it was, "neat to see the notes because sometimes during the interviews with the doctors, they'll write things down that I said, and well I didn't really mean it exactly that way." P5 also said, "I see the hospital medications…that's important I guess, just in case there are any discrepancies…between what stuff I think I should be taking, and [that] which they didn't give me here, and I'm like why, not?" P2 and P6 mentioned noticing discrepancies between when medications were documented as administered and when they believed them to have actually been administered.

Discussion

While patient engagement has recently been a topic of considerable interest[23–29], research in the inpatient setting is lacking.[12,14,15] In *The Patient Checklist*, Elizabeth Bailey observed that "once a patient enters a hospital for treatment of any kind, what he or she needs most of all is knowledge – what is happening to him, and why."[30] We believe our study may be the first to evaluate the effect of providing patients with access to their entire charts, including physician notes, during their hospital stay. The most important findings of our study were that information sharing: 1) was perceived as desirable to patients and acceptable to clinicians, 2) allowed patients to more actively participate in their hospital care, and 3) may impact clinician behavior in terms of workload, communication, and note-writing practices.

Our results suggest that greater information sharing with hospital patients can be beneficial. Of the seven patients we interviewed, six of them requested to continue receiving the daily printouts. Patients appreciated seeing the details of their hospital care even in the raw, unfiltered format that came from our EHR system. We believe that a more tailored format, though difficult to actualize, could have even greater potential to increase patient engagement. The survey results indicated that clinicians also viewed information sharing with patients favorably. For clinicians, the sharing of objective data in particular seemed to provoke little controversy. The sharing of more subjective data, such as progress notes and consultation notes, was less agreeable to survey respondents, but still, the majority were comfortable with sharing this information.

Clinicians' expectations of the consequences of inpatient information sharing were mixed. Most were optimistic that sharing information would enhance communication and increase patient engagement in decision-making. However, some survey responses reflected concerns about increased anxiety and misunderstanding of information. Though our sample was small, none of the patients interviewed expressed that they felt additional anxiety from viewing their records. On the contrary, several patients commented that receiving the daily printout made them feel more informed and better able to understand the details of their care. Additionally, having access to their information allows patients to act as another 'line of defense' to identify incorrect information and potentially decrease medical errors.

In the outpatient setting, the OpenNotes project explored the effects of giving patients the ability to read their doctor's office visit notes.[31] The project has been successful enough that the three institutions involved continued sharing notes with patients after the study ended. Additionally, the project has expanded to other institutions, including the Department of Veterans Affairs (VA).[32] Although further study of information sharing in the hospital setting is warranted, our results suggest that the OpenNotes concept may also be applicable to inpatient care.

OpenNotes investigators reported that increased access to physician notes by patients may change note-writing behavior.[33] Comments from our clinician survey suggested that this was a concern at our institution. For example, patients may be offended by reading certain descriptions in their charts (e.g., "obese", "disheveled"). Moreover, notes often contain expressions of uncertainty, frightening differential diagnoses, and findings that lack interpretation. If notes are readily accessible to patients, clinicians might avoid using certain terms, or omit clinically relevant details to protect patients from unnecessary anxiety. On the other hand, because so much of what clinicians currently document is not easily understandable by patients, they may be inclined to write in a manner that is more intelligible to patients (e.g., by using fewer acronyms).

Design Considerations for Inpatient Engagement

Beyond a pilot study, printing and hand-delivering daily reports as we did in this study is probably not a feasible option for most hospitals. However, a recent study from a neonatal intensive care unit described delivering a one-page paper handout to parents each day of their baby's stay.[34] The handout included information on the baby's care team, respiratory status, nutritional status, medications, most recent lab results, and care plan. Similarly, the VA began providing inpatients with access to limited information through the Daily Plan project, which delivered a daily printout of the patient's medications, appointments and diagnostic tests which nurses would then review with the patients each day.[35] We[13] and others[14,15] are working to deploy online patient engagement solutions for hospital patients. So far, these projects have provided tablet computers to patients in the hospital on a small scale and with limited information—excluding clinical notes, for example.

Future work should investigate how to present patient-specific medical information in a more patient-friendly manner. Appealing to diverse populations of patients, including those with low health literacy, low literacy in general, and non-native English speakers is a particular challenge. Using resources such as the Consumer Health Vocabulary and MedlinePlus may help to make complex medical terminology more understandable.[36,37] Additionally, use of visualizations has been shown to help patients better understand health information, especially for individuals with low health literacy.[38–40] Research on the creation of visualizations that explain inpatient data is warranted.

Another challenge to providing increased access to patient information is addressing privacy and security concerns. A recent article by Bates and colleagues discusses the need for family, caregiver, and care partner access to patient data.[41] Restricting access to certain types of data may be a feature necessary for privacy of sensitive topics like testing for drug use, or pregnancy in teenagers.

Limitations

This study was limited by a relatively small sample size for the clinician survey (n=53), and similarly by a small sample size of hospital patients who participated in the inpatient field study (n=8). Of the eight patients in the inpatient field study, one was discharged prior to the interview and two additional patients had not read through their records. Thus, the sample size was insufficient to achieve thematic saturation. Nevertheless, we believe our results shed light on the understudied topic of engaging hospital patients in their care and provide a foundation for future research. The study was conducted at a single site, an academic medical center in an urban setting, and our findings may not generalize to other settings. More specifically, our study was focused primarily in the domain of cardiology, and different results may be found in other patient populations. For example, patients with cardiac disease may have higher baseline engagement levels than patients with less chronic diseases, and thus may be more interested in seeing their data. Future work should explore whether differences in settings or patient populations exist with respect to inpatient engagement.

Conclusion

Healthcare delivery organizations are moving towards greater sharing of information with patients. Consumers, who are the "sole subject matter expert on themselves,"[42] are increasingly expecting to have access to their data. Our study found that clinicians were mostly comfortable with increased information sharing, and patients benefitted from

receiving daily printouts of their hospital care record. Increased access to information will enable patients to more actively participate in their hospital care.

Acknowledgments

This research was supported by grants from the National Library of Medicine (T15LM007079) and the Agency for Healthcare Research and Quality (R01HS21816). The authors thank our study participants, Dr. Suzanne Bakken for her feedback on creating the survey instrument, and the rest of our departmental colleagues for their continued support.

References

1. Bickmore TW, Pfeifer LM, Jack BW. Taking the time to care. SIGCHI Conference on Human Factors in Computing Systems [Internet]. ACM Press; 2009 [cited 2013 Feb 27]. p. 1265. Available from: http://dl.acm.org/citation.cfm?doid=1518701.1518891

2. Cumbler E, Wald H, Kutner J. Lack of patient knowledge regarding hospital medications. J Hosp Med Off Publ Soc Hosp Med. 2010 Feb;5(2):83–6.

3. Makaryus AN, Friedman EA. Patients' Understanding of Their Treatment Plans and Diagnosis at Discharge. Mayo Clin Proc. 2005 Aug;80(8):991–4.

4. O'Leary KJ, Kulkarni N, Landler MP, Jeon J, Hahn KJ, Englert KM, et al. Hospitalized patients' understanding of their plan of care. Mayo Clin Proc. 2010;85(1):47–52.

5. Collins SA, Vawdrey DK, Kukafka R, Kuperman GJ. Policies for patient access to clinical data via PHRs: current state and recommendations. J Am Med Inform Assoc. 2011 Sep 7;18(Suppl 1):i2–i7.

6. Hibbard JH, Greene J. What The Evidence Shows About Patient Activation: Better Health Outcomes And Care Experiences; Fewer Data On Costs. Health Aff (Millwood). 2013 Feb 4;32(2):207–14.

7. Hibbard JH, Greene J, Overton V. Patients With Lower Activation Associated With Higher Costs; Delivery Systems Should Know Their Patients' "Scores."Health Aff (Millwood). 2013 Feb 4;32(2):216–22.

8. Dentzer S. Rx For The "Blockbuster Drug" Of Patient Engagement. Health Aff (Millwood). 2013 Feb 4;32(2):202–202.

9. Institute of Medicine, ebrary I. Crossing the quality chasm [Internet]. Washington, D.C.: National Academy Press; 2003. 38 p. Available from: http://site.ebrary.com/lib/uvalib/Doc?id=10056947

10. HealthIT.gov [Internet]. [cited 2014 Mar 8]. Available from: http://www.healthit.gov/providers-professionals

11. Buhlman N, Matthes N. The time to prepare for value-based purchasing is now. White Pap Hosp. 2011;

12. Prey JE, Woollen J, Wilcox L, Sackeim AD, Hripcsak G, Bakken S, et al. Patient engagement in the inpatient setting: a systematic review. J Am Med Inform Assoc [Internet]. 2013 Nov 22 [cited 2014 Jan 16]; Available from: http://jamia.bmj.com/cgi/doi/10.1136/amiajnl-2013-002141

13. Vawdrey DK, Wilcox LG, Collins SA, Bakken S, Feiner S, Boyer A, et al. A tablet computer application for patients to participate in their hospital care. AMIA Annual Symposium Proceedings. American Medical Informatics Association; 2011. p. 1428.

14. Dykes PC, Carroll DL, Hurley AC, Benoit A, Chang F, Pozzar R, et al. Building and Testing a Patient-Centric Electronic Bedside Communication Center. J Gerontol Nurs. 2013 Jan 1;39(1):15–9.

15. Greysen SR, Khanna RR, Jacolbia R, Lee HM, Auerbach AD. Tablet computers for hospitalized patients: A pilot study to improve inpatient engagement. J Hosp Med. 2014;n/a–n/a.

16. Walker J, Darer JD, Elmore JG, Delbanco T. The Road toward Fully Transparent Medical Records. N Engl J Med. 2014 Jan 2;370(1):6–8.

17. Barnette JJ. Effects of Stem and Likert Response Option Reversals on Survey Internal Consistency: If You Feel the Need, There is a Better Alternative to Using those Negatively Worded Stems. Educ Psychol Meas. 2000 Jun 1;60(3):361–70.

18. Team RC. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2014. Available from: http://www.R-project.org/

19. Cabin RJ, Mitchell RJ. To Bonferroni or Not to Bonferroni: When and How Are the Questions. Bull Ecol Soc Am. 2000 Jul 1;81(3):246–8.

20. Wilcox AB, Vawdrey DK, Chen Y-H, Forman B, Hripcsak G. The Evolving Use of a Clinical Data Repository: Facilitating Data Access Within an Electronic Medical Record. AMIA Annu Symp Proc. 2009;2009:701–5.

21. Burnard P. A method of analysing interview transcripts in qualitative research. Nurse Educ Today. 1991 Dec;11(6):461–6.

22. Robbins NB, Heiberger RM. Plotting Likert and other rating scales. Proceedings of the 2011 Joint Statistical Meeting. 2011.

23. Bird AP, Walji MT. Our patients have access to their medical records. Br Med J Clin Res Ed. 1986;292(6520):595–6.

24. Cohen RN. Whose file is it anyway? Discussion paper. J R Soc Med. 1985;78(2):126–8.

25. Fisher B, Britten N. Patient access to records: expectations of hospital doctors and experiences of cancer patients. Br J Gen Pract J R Coll Gen Pract. 1993;43(367):52–6.

26. Jimison H, Sher P. Consumer Health Informatics: Health Information Technology for Consumers. J Am Soc Inf Sci. 1995;46(10):783–90.

27. Maly RC, Bourque LB, Engelhardt RF. A randomized controlled trial of facilitating information giving to patients with chronic medical conditions: effects on outcomes of care. J Fam Pract. 1999;48(5):356–63.

28. Elwyn G, Edwards A, Kinnersley P, Grol R. Shared decision making and the concept of equipoise: the competences of involving patients in healthcare choices. Br J Gen Pract. 2000;50(460):892.

29. Greenfield S, Kaplan SH, Ware JE, Yano EM, Frank HJL. Patients' participation in medical care. J Gen Intern Med. 1988 Sep;3(5):448–57.

30. Bailey E. The patient's checklist : what every hospital patient needs to know to stay organized, safe & sane. New York: Sterling Pub.; 2011. xxvi, 133 p. p.

31. Delbanco T, Walker J, Darer JD, Elmore JG, Feldman HJ, Leveille SG, et al. Open notes: doctors and patients signing on. Ann Intern Med. 2010;153(2):121.

32. Trossman S. OpenNotes initiative aims to improve patient-clinician communication, care. Am Nurse. 2013 Oct;45(5):10.

33. Delbanco T, Walker J, Bell SK, Darer JD, Elmore JG, Farag N, et al. Inviting Patients to Read Their Doctors' Notes: A Quasi-experimental Study and a Look Ahead. Ann Intern Med. 2012 Oct 2;157(7):461–70.

34. Palma JP, Keller H, Godin M, Wayman K, Cohen RS, Rhine WD, et al. Impact of an EMR-Based Daily Patient Update Letter on Communication and Parent Engagement in a Neonatal Intensive Care Unit. J Particip Med. 2012 Dec 31;4.

35. King BJ, Mills PD, Fore A, Mitchell C. The Daily Plan®: Including patients for safety's sake. Nurs Manag Springhouse. 2012 Mar;43(3):15–8.

36. Zeng QT, Tse T. Exploring and Developing Consumer Health Vocabularies. J Am Med Inform Assoc. 2006 Jan 1;13(1):24–9.

37. Medicine NL of. MedlinePlus Connect: linking electronic health records (EHRs) to MedlinePlus health information. 2011.

38. Garcia-Retamero R, Okan Y, Cokely ET. Using Visual Aids to Improve Communication of Risks about Health: A Review. Sci World J [Internet]. 2012 May 2 [cited 2014 Mar 11];2012. Available from: http://www.hindawi.com/journals/tswj/2012/562637/abs/

39. Ancker JS, Senathirajah Y, Kukafka R, Starren JB. Design Features of Graphs in Health Risk Communication: A Systematic Review. J Am Med Inform Assoc. 2006 Nov 1;13(6):608–18.

40. Gaissmaier W, Wegwarth O, Skopec D, Müller A-S, Broschinski S, Politi MC. Numbers can be worth a thousand pictures: Individual differences in understanding graphical and numerical representations of health-related information. Health Psychol. 2012;31(3):286–96.

41. Sarkar U, Bates DW. Care partners and online patient portals. JAMA. 2014 Jan 22;311(4):357–8.

42. Goetz T. The Decision Tree: How to Make Better Choices and Take Control of Your Health. Rodale; 2011. 339 p.

Improving Clinical Data Integrity by using Data Adjudication Techniques for Data Received through a Health Information Exchange (HIE)

Pallavi Ranade-Kharkar, MS[1,2], Susan E. Pollock, MS[1], Darren K. Mann, BS[1], Sidney N. Thornton, PhD[1,2]
[1]Intermountain Healthcare, Murray, UT;
[2]Department of Biomedical Informatics, University of Utah, Salt Lake City, UT

Abstract

Growing participation in Healthcare Information Exchange (HIE) has created opportunities for the seamless integration of external data into an organization's own EHR and clinical workflows. The process of integrating external data has the potential to detect data integrity issues. Lack of critiquing external data before its incorporation can lead to data unfit for use in the clinical setting. HIE data adjudication, by detecting inconsistencies, physiological and temporal incompatibilities, data completeness and timeliness issues in HIE data, facilitates corrective actions and improves clinical data integrity.

Introduction

Health Information Exchange (HIE) holds the promise of improving the quality and efficiency of health care[1, 2]. Hospital-based HIE grew 41% between 2008 and 2012 as six out of ten hospitals exchanged information with providers and hospitals external to their organization[3]. Another study found three-fold increase in HIE participation from 2010 to 2012 when 10% ambulatory practices participated in 119 operational HIEs[4]. As the capability to exchange health information among healthcare organizations is a mandatory requirement for "Meaningful Use" and certification of electronic health records (EHR)[5], we can expect increased adoption and utilization of HIE in the future.

Growing participation in HIE has created opportunities for integrating data from external sources into an organization's own EHR. The exchange of data for patients served by more than one organization can also lead to uncovering inconsistencies in data. This may open up opportunities for corrective actions on one's own data. However, more data does not always mean better care. Kuperman and McGowan note that a flood of data from external sources can have the unintended consequence of overwhelming clinicians[6]. HIE-related errors can also reduce patient safety[7]. Incoming external data need to be critiqued for duplication, contradiction and physiological compatibilities. Data incorporation done in conjunction with data adjudication has the potential to improve clinical data integrity in the receiving organization's EHR. This manuscript addresses the following types of data integrity issues: the consistency, physiological compatibility, completeness, and timeliness of data.

In a previous study, we analyzed data from Intermountain Healthcare's (IH) Enterprise Data Warehouse (EDW) to identify error patterns, extracted requirements for a data adjudication framework, and designed architecture for an inferencing solution[8]. In this manuscript, we present a study which describes how the data adjudication infrastructure addresses data integrity issues for the use case of a point-to-point HIE between IH and a government trading partner that exchanges death information. Our approach as a partner in the Healtheway collaborative[23] (formerly Nationwide Health Information Network Exchange) has been point-to-point HIE connections rather than to an HIE as a consolidator organization. We believe this approach allows for local decision autonomy for issues of clinical data integrity as opposed to HIE organizations which are limited by consensus decisions and potentially by the lowest quality source of data.

This study is novel for a number of reasons: 1. It addresses the issue of data integrity for HIE data inbound into an organization; 2. It goes beyond passively accepting HIE data and takes it a step further into incorporating that data into an organization's own EHR; 3. It takes into consideration 4 axes of data integrity: consistency, compatibility, completeness and timeliness; 4. It applies a decision support infrastructure to adjudicate HIE data.

Background

Data integrity: EHRs have highly variable levels of data accuracy ranging from 30% to 100% data correctness and completeness[9]. The integrity of data in healthcare information systems affects decision making processes, both manual and automated. Berner et al.[10] report that gaps in patient data can affect the accuracy of the output of the Clinical Decision Support (CDS) tools. Guidance and recommendations provided by CDS systems may not be

trustworthy unless they are based on data with high levels of integrity[11]. Clinician perception of the effectiveness of CDS is dependent on data quality across settings, organizations and types of EHRs[12]. Data integrity issues also affect secondary use of health care data for the purposes of quality improvement, public health, and research[13].

In this study, we address the following types of clinical data integrity: consistency, physiological compatibility, completeness and timeliness of data. Current literature on data integrity in the context of HIE focuses on the data integrity issues surrounding patient identification[14, 15]. In this manuscript, we focus on death information exchanged in an HIE, its consistency and physiological compatibility with other clinical data in the receiving organization's EHR, and its completeness and timeliness.

Data incorporation: While HIE is about moving data from one organization to the other, data incorporation is about integrating appropriate data into an organization's own EHR and workflows. Literature on data incorporation in the realm of HIE is sparse. Frisse et al.[16] describe data incorporation performed in the Memphis HIE which was done at the display level. In another system, a high-tech approach was applied to medication reconciliation where data from a pharmacy system was incorporated into an EHR[17]. Cerner, a commercial vendor of EHR and IH's strategic partner, uses a document-based data model for integration of HIE data. The incoming documents with patient data are stored in a repository and the data in it is transformed for display, transfer, etc., as needed[18]. Data incorporation becomes more challenging if the EHRs on both sides of the exchange are different.

Another way to accomplish data integration at the discrete level is through electronic messaging interfaces between organizations using standards such as HL7 v2.x. However, discrete data incorporation done using document-based exchange is more scalable without incurring the expense of traditional messaging interfaces. We did not find evidence of network exchange based "complete" data incorporation in literature where the incorporated data becomes fully integrated with the EHR and is stored just like other similar data natively captured and stored by the EHR.

Setting for Health Information Exchange: This study was conducted at IH in Salt Lake City, UT. IH is currently participating in multiple HIE projects including the Care Connectivity Consortium (CCC)[19] at the national level, the Clinical Health Information Exchange (cHIE)[20], the Utah Health Information Network (UHIN)[21] and a private exchange with a government trading partner at the state level. IH has developed an infrastructure that processes different types of standards-based documents and can be universally used for any of the various exchanges. The exchanges have had to face administrative, logistical, policy, and other delays. As a result, the exchanges are at different stages of maturity.

We chose the exchange of death-related information between our government trading partner and IH as a convenience sample for this study. This type of data is available for all patients regardless of whether they opted-in or opted-out of HIE. Moreover, death is an important clinical outcome. Having high integrity death-related data is important for secondary uses such as patient outcome, resource utilization and cost analyses.

Previous work: The authors (PRK, DKM and SNT) analyzed 2.2 million Hemoglobin A1C (HgbA1C) result records from IH's enterprise data warehouse (EDW) for potential error patterns[8]. The HgbA1C results records are generated at IH, at point-of-care and external non-Intermountain laboratories. The authors then extracted requirements and designed architecture for a data adjudication infrastructure that integrates into the overall system architecture to detect data integrity issues and facilitate corrective actions[8].

Figure 1 shows the architectural diagram of Intermountain's HIE data adjudication system along with the document processing system used for HIE. External HIE data in the form of a summary of care document in the HL7 Consolidated Clinical Document Architecture format (CCDA) or an HL7 clinical document for reporting death information, inbound into IH, is initially processed to establish positive patient identity and positive consent. Data that pass these criteria are then transferred to the Inbound Message Orchestrator. This component is responsible for managing the workflow marked 2 through 6 in Figure 1. First, incoming data are filtered for new data at the document level to eliminate duplicative processing. Data is then extracted into internal data models and mapped from standard terminologies such as SNOMED and LOINC codes to Intermountain's own terminology using Intermountain's proprietary data dictionary.

Data that has been extracted and mapped to local terminology is fed into the HIE Data Adjudicator where a series of JBoss® Drools rules are run on it. Context-independent rules, i.e. rules that can run solely on the external data are run first for maximizing efficiency. These include data completeness and temporal compatibility rules. Data completeness rules check whether the incoming data meets mandatory data requirements as determined by the analysts at IH. Temporal compatibility rules check whether data are current and relevant. If the data passes the

context-independent rules, the rules processing continues to the context-dependent rules, i.e. rules which infer on external as well as internal data. These include data redundancy and data compatibility rules. Data redundancy rules compare incoming data to the data that is stored in IH's longitudinal patient data repository called the "Clinical Data Repository" (CDR) and determine if the data are duplicative. The data compatibility rules are responsible for detecting inconsistencies and physiological incompatibilities between the external HIE data and data recorded at IH. The HIE Data Adjudicator result statuses include "accept", "accept_as_new" and "reject". The "accept_as_new" status results in new records in the database. The "accept" adjudication status indicates that the new data is an update to existing data. Data in either statuses of "accept_as_new" and "accept" are stored to Intermountain's CDR. Data "rejected" by the HIE Data Adjudicator goes through various data integrity workflows. This architecture is fully implemented and is undergoing verification and validation at the time this manuscript was written.

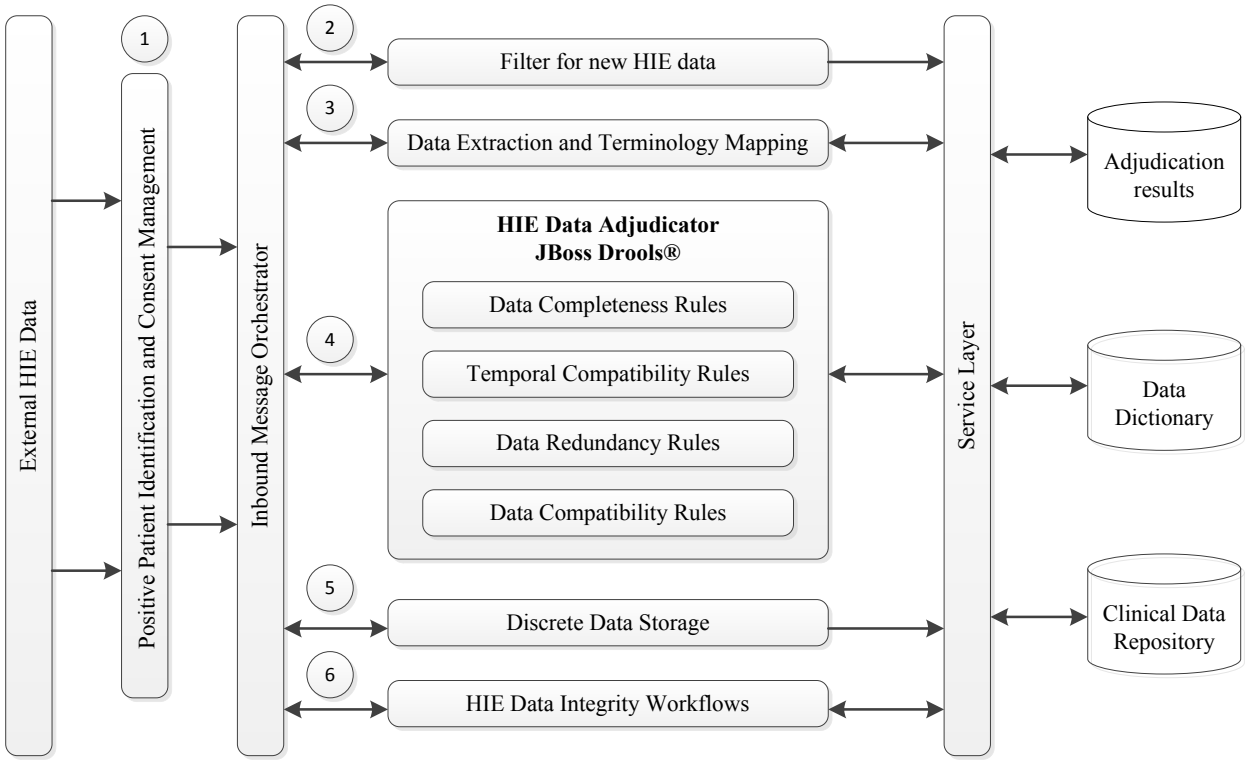


Figure 1. Architecture for HIE Document Processing and Data Adjudication

Methods

The data source (IH's government trading partner) receives death certificate data from multiple sources within the state of Utah. IH maintains the death records for patients who pass away at any of its own facilities. We performed a retrospective, descriptive analysis to compare death information recorded at IH facilities to that received from the data source between January 1, 1996 and August 31, 2013. The patient identities were matched prior to this analysis using research-driven technologies that included a combination of algorithmic and manual processes. Algorithms used for determining positive patient identity included Jaro-Winkler and Levenshtein distance algorithms. We measured the inconsistencies, physiological incompatibilities, completeness and timeliness of the death information. We enhanced the document processing system developed at IH for HIE by implementing the data adjudication architecture. We then ran a sample of the data through the data adjudication infrastructure to validate the errors we found in our analysis.

Data Inconsistencies/Physiological Incompatibility: We first compared the date/time of death in the data generated at IH to that in the data received from the data source. We then analyzed the data to determine if clinical data was recorded for deceased patients to understand the physiological incompatibility issues. We included billed encounters, vital signs, laboratory results, medication orders and problems in our analysis because these were most frequently used data types to record data for all patients at IH. We excluded data recorded for reasons that were

considered appropriate for deceased patients such as encounters for transplants, equipment, and organ and cadaver donors. We also included an allowance of 24 hours after the earliest recorded death date to accommodate for systems that have different ways to record time of death. We did not include data that was marked canceled or error. In the cases where we found multiple recorded death dates, we considered the earliest recorded death date as the death date for our analysis. Finally, we checked for temporal compatibility by looking at date/time of death and determining if it is in the future.

Data Completeness: The analysts at IH determined the criteria for completeness for death information. Death information is considered "complete" at IH if a valid patient identification and date/time of death is recorded.

Timeliness: We calculated the latency of recording of death information as the difference between the actual death date/time and the date/time the information was processed by IH workflows. The Utah State Code (Title 26 Chapter 2 Section 13)[22] stipulates that death certificates must be filed with the health department within five days of death. Our analysis adjusted for this allowance.

We had to simulate the exchange between IH and the government trading partner because of unforeseen administrative and logistical delays. The simulation system translated raw death data from the data source into an HL7 clinical document for the template id: 2.16.840.1.113883.10.20.24.1. This document was then input to and processed by the Inbound Message Orchestrator as shown in Figure 1. Process steps 2 through 6 were identical between the simulated and model systems.

We randomly selected 100 death records which were found to have inconsistencies ("error" records) and 100 death records for which we did not find data integrity issues ("good" records). The simulation system created HL7 clinical documents using this data. These records were run through the adjudication process for two iterations. The first iteration was performed to validate all the rules. We then added a dummy cause of death to half of the "good" records and performed a second iteration of the data through the system with the updated records as well as the records that were not changed. The second run was performed to validate the data redundancy rules (that detect duplicates) in addition to the other types of rules. We further created "empty" documents that did not have either date or time of death ("empty" records). We ran these through the system to validate the rejection logic in the data completeness rules. We also created documents with date/time of death in the future ("future" records) to validate the temporal compatibility rules.

Results

IH received a total of 198,453 death records from the data source between January 1, 1996 and August 31, 2013. A total of 47,175 patient deaths were recorded at IH facilities during the same time period.

Table 1 describes data inconsistencies and physiological incompatibilities with the clinical data recorded for deceased patients at least 24 hours after their death date/time. We found a total of 15,721 inconsistent records for different types of clinical observations. A total of 4483 unique patients were affected by these inconsistencies across all types of clinical observations we analyzed. We did not find any temporal compatibility issues with the data.

Table 1. Death data related inconsistencies and physiological incompatibilities

Type of Clinical Observation	Number of Unique Patients	Number of Inconsistent Records	Comments
Billed encounters	401	819	Data was recorded under 110 different types of billed encounters.
Vital Signs	23	2820	12 types of vital signs were recorded (10 from inpatient and 2 from the outpatient setting).
Laboratory Results	1271	2614	Data was recorded for 704 different test codes.
Medications Orders	2889	9467	Out of all the inconsistent records, 9359 were from the inpatient setting, 96 were from the outpatient setting and 12 were from other settings.
Problems	1	1	

We found that all data met the data completeness criteria decided upon by our analysts. However, we found that some records were missing cause of death. Table 2 details the results of the data completeness analysis.

Table 2. Data completeness profile for death related data

Description	Number of records	Contributes to data completeness
Missing patient id	0	Yes
Missing date/time of death	0	Yes
Missing cause of death	852	No
Missing location of death	0	No

We currently manually process death data from our government trading partner. This is typically done at unpredictable intervals. We had timeliness data for only a subset of the death records (35962 or 18.1% of death records). Table 3 shows the results of the latency calculations for the death data from the data source.

Table 3. Data timeliness profile for death related data

Latency description	Latency value (days)
Minimum	76
Maximum	1051
Mean	549.6304
Median	544

The data adjudication system validation results were as expected. All the data integrity issues were successfully detected and the related records were "rejected". The "good" records were "accepted" or "accepted_as_new" and stored to the test CDR. Table 4 gives the validation results.

Table 4. Validation of the data adjudicator framework for death related data

Iteration Number	Type of records	Number of records	% "Accepted _as_new"	% "Accepted"	% "Rejected"	Reason for Rejection
Iteration 1	"Good"	100	100	0	0	n/a
	"Error"	100	0	0	100	"Failed Data Compatibility Rules"
Iteration 2	"Good" (with dummy cause of death)	50	0	100	0	n/a
	"Good"	50	0	0	100	"Failed Data Redundancy Rules"
	"Error"	100	0	0	100	"Failed Data Compatibility Rules"
Iteration 3	"Empty"	2	0	0	100	"Failed Data Completeness Rules"
Iteration 4	"Future"	2	0	0	100	"Failed Temporal Compatibility Rules"

Discussion

We uncovered inconsistencies and physiological incompatibilities for 4483 patients marked as deceased (2.259% of the HIE inbound data) even after we accounted for the types of data that are appropriate to be recorded on deceased patients (such as encounters for transplants, equipment, and organ and cadaver donors). We augmented our original time criteria for search of 24 hours after death date/time to 15 and 30 days after the death date/time. Although the number of inconsistencies and physiological incompatibilities decreased as we increased our time criteria for search, there appeared to be ongoing data issues that were not picked up by our data integrity workflows. There could be a variety of reasons, both benign and non-benign, for these data integrity issues and may be attributed to either side of the exchange (IH or the data source). An example of a benign reason could be delayed billing by any of the various systems (excluding equipment rentals) involved in a patient's care. A non-benign reason could be the possibility of fraudulent activity, where the identity of a deceased patient is used to procure healthcare services. Further investigation is necessary uncover details about the reasons. Data adjudication applied to HIE data gives us the opportunity to detect such errors and brings us one step closer to facilitating corrective actions and thus improving the data integrity of one's own EHR.

Data completeness and temporal incompatibility were found to be a non-issue with the death information use case. This could be attributed to the mandatory requirements imposed by the Utah State Code at the source system. We think that the timeliness issues we found (mean latency of 549.63 days) are more of a process issue rather than a data issue. We believe the automation of the HIE exchange between IH and the government trading partner for death information will take us closer to solving the problem.

Handling redundant data at the document as well as discrete data level is important to reduce inefficiencies in processing of HIE data. Without the ability to detect duplicate documents, systems may waste resources by re-processing previously processed data that provides no new information. Likewise, the ability to compare discrete data and decide whether the incoming HIE data is new, duplicate or an update to already existing data can greatly improve process and database resource utilization.

The data adjudication framework implemented in this study is highly generalizable and can be extended to handle other clinical data types by adding appropriate data extraction, terminology mapping and data integrity rules specific to the data type. The framework can also be applied to demographic or administrative data. It can be further enhanced to detect more data integrity issues such as clinical data trend incompatibilities. The framework also suggests an approach to integrate HIE data focusing on episodic care details into patient's longitudinal data record. This approach is scalable and has the advantage of inter-organizational electronic messaging interfaces without the overhead associated with such interfaces.

The approach described in this manuscript can be applied bi-directionally. In other words, it can be applied to improve clinical data integrity of the EHR sending the data as well as the one receiving the data. We envision that the data integrity critique (a report of inconsistencies, physiological and temporal incompatibilities and incompleteness of data) can be fed back to the sending system.

We believe that the information gained and lessons learned from our study will inform our national and international collaborators working towards improved exchange profiles, standards-based data adjudication rules and HIE best practices.

Limitations: In order to validate the data adjudication infrastructure we had to simulate the data exchange between IH and the data source for death information because the effort to establish a working exchange faced administrative and logistical delays. Although our verification and validation test cases adequately covered variations in data, we understand that the real working exchange may add some data or system conditions that we did not anticipate. However, should any issues arise, we feel that we will be able to rectify them efficiently because the implementation of the data adjudication framework allows for easy and efficient updates. Another limitation of the study is related to the accuracy of patient identity matching algorithms and processes. Any errors with patient matching can adversely affect the analysis and data adjudication results. However, we feel that because this study has used the patient identity matching workflows that have been implemented and incrementally improved at IH for over a decade, the potential negative effects of this are minimal. Finally, we acknowledge that clinically more complex data models exist as compared to the death data model we have used in this study. We feel that because of our modular approach, our framework can be augmented to use resources such as ontologies to handle hierarchical and complex data models.

Future work: Our next step is to connect the data adjudication infrastructure demonstrated in this study to a working HIE between IH and the government trading partner. We believe that the experience will help us fine tune the data adjudication logic. We plan to expand the data adjudication to include problems, medications, allergies, and other data types, that are exchanged in the HIE domain. Additionally, we want to understand, in more detail, the reasons for data integrity issues we found. This may further inform our effort to enhance the data adjudication logic. Finally, we intend to research ways to translate the data adjudication rules into interoperable and shareable knowledge.

Conclusion

We demonstrated the capabilities of a data adjudication infrastructure integrated into Intermountain Healthcare's HIE architecture. We showed that it can detect data integrity issues related to inconsistencies, physiological and temporal incompatibilities, and completeness. It also addressed new data integrity issues related to redundancy by detecting duplicates external and internal to the organization's EHR. Data adjudication reduces the burden of manual review and resolution of data integrity issues by automatically connecting the data adjudication results to HIE data integrity workflows and facilitating corrective actions.

Acknowledgements

We would like to acknowledge Shannon Hood and Dallin Rogers for providing domain knowledge for the patient identification process at Intermountain Healthcare. We would also like to acknowledge Jeff Duncan for his expertise with death related information.

References

1. Walker J, Pan E, Johnston D, Adler-Milstein J, Bates DW, Middleton B. The value of health care information exchange and interoperability. Health Aff (Millwood). 2005;Suppl Web Exclusives:W5-10-W5-8. Epub 2005/01/22.
2. Fontaine P, Ross SE, Zink T, Schilling LM. Systematic review of health information exchange in primary care practices. Journal of the American Board of Family Medicine : JABFM. 2010;23(5):655-70. Epub 2010/09/09.
3. Furukawa MF, Patel V, Charles D, Swain M, Mostashari F. Hospital electronic health information exchange grew substantially in 2008-12. Health Aff (Millwood). 2013;32(8):1346-54. Epub 2013/08/07.
4. Adler-Milstein J, Bates DW, Jha AK. Operational health information exchanges show substantial growth, but long-term funding remains a concern. Health Aff (Millwood). 2013;32(8):1486-92. Epub 2013/07/11.
5. Health Information Exchange (HIE). [URL]; Available from: http://www.healthit.gov/HIE. (Accessed on March 13, 2014)
6. Kuperman GJ, McGowan JJ. Potential unintended consequences of health information exchange. Journal of general internal medicine. 2013;28(12):1663-6. Epub 2013/05/22.
7. Kaelber DC, Bates DW. Health information exchange and patient safety. Journal of biomedical informatics. 2007;40(6 Suppl):S40-5. Epub 2007/10/24.
8. Ranade-Kharkar P, Mann D, Thornton S. Data adjudication architecture for health information exchange (HIE): a case of adjudicating and storing hemoglobin a1c values. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2013.
9. Hogan WR, Wagner MM. Accuracy of data in computer-based patient records. Journal of the American Medical Informatics Association : JAMIA. 1997;4(5):342-55. Epub 1997/09/18.
10. Berner ES, Kasiraman RK, Yu F, Ray MN, Houston TK. Data quality in the outpatient setting: impact on clinical decision support systems. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2005:41-5. Epub 2006/06/17.
11. Hasan S, Padman R. Analyzing the effect of data quality on the accuracy of clinical decision support systems: a computer simulation approach. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2006:324-8. Epub 2007/01/24.
12. McCormack JL, Ash JS. Clinician perspectives on the quality of patient data used for clinical decision support: a qualitative study. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2012;2012:1302-9. Epub 2013/01/11.
13. Ancker JS, Shih S, Singh MP, Snyder A, Edwards A, Kaushal R. Root causes underlying challenges to secondary use of data. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2011;2011:57-62. Epub 2011/12/24.

14. AHIMA. Ensuring data integrity in health information exchange. Available from: http://library.ahima.org/xpedio/groups/public/documents/ahima/bok1_049675.pdf. (Accessed on March 13, 2014)
15. Thornton SN WJ, Russ GE, Westburg LJ, Mann DK, Rasumssen DN. Sharing qualitative matching parameters among master patient indices. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2013.
16. Frisse ME, Tang L, Belsito A, Overhage JM. Development and use of a medication history service associated with a health information exchange: architecture and preliminary findings. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2010;2010:242-5. Epub 2011/02/25.
17. High-tech approach to medication reconciliation saves time, bolsters safety at hospital in northern Virginia. ED management : the monthly update on emergency department management. 2011;23(10):117-9. Epub 2011/10/07.
18. Cerner Corporation. Secure exchange of medical information: clinical exchange platform. Available from: https://www.cerner.com/uploadedFiles/fl03_789_10_v1_CXP_hr[1].pdf. (Accessed on March 13, 2014)
19. Care connectivity consortium. Available from: http://www.careconnectivity.org/. (Accessed on March 13, 2014)
20. My clinical health information exchange. Available from: http://mychie.org/. (Accessed on March 13, 2014)
21. Utah Health Information Network. Available from: http://www.uhin.org/. (Accessed on March 13, 2014)
22. Utah State Code (Title 26 Chapter 2 Section 13). Available from: http://le.utah.gov/code/TITLE26/htm/26_02_001300.htm. (Accessed on March 13, 2014)
23. Healtheway. Available from: http://healthewayinc.org/. (Accessed on July 22, 2014)

Development and Evaluation of Reference Standards for Image-based Telemedicine Diagnosis and Clinical Research Studies in Ophthalmology

Michael C. Ryan, MS,[1] Susan Ostmo, MPH,[1] Karyn Jonas, RN,[3] Audina Berrocal, MD,[4] Kimberly Drenser, MD,[5] Jason Horowitz, MD,[6] Thomas C. Lee, MD,[7] Charles Simmons, MD,[8] Maria-Ana Martinez-Castellanos, MD,[9] R.V. Paul Chan, MD,[3] Michael F. Chiang, MD[1,2]

Departments of [1]Ophthalmology and [2]Medical Informatics & Clinical Epidemiology, Oregon Health & Science University, Portland, OR; [3]Department of Ophthalmology, Weill Cornell Medical College, New York, NY; [4]Department of Ophthalmology, University of Miami, Miami, FL; [5]Department of Ophthalmology, William Beaumont Hospital, Royal Oak, MI; [6]Department of Ophthalmology, Columbia University, New York,NY; [7]Department of Ophthalmology, Children's Hospital Los Angeles, Los Angeles, CA; [8]Department of Pediatrics, Cedars-Sinai Medical Center, Los Angeles, CA; [9]Department of Ophthalmology, APEC, Mexico City, Mexico

Abstract

Information systems managing image-based data for telemedicine or clinical research applications require a reference standard representing the correct diagnosis. Accurate reference standards are difficult to establish because of imperfect agreement among physicians, and discrepancies between clinical vs. image-based diagnosis. This study is designed to describe the development and evaluation of reference standards for image-based diagnosis, which combine diagnostic impressions of multiple image readers with the actual clinical diagnoses. We show that agreement between image reading and clinical examinations was imperfect (689 [32%] discrepancies in 2148 image readings), as was inter-reader agreement (kappa 0.490-0.652). This was improved by establishing an image-based reference standard defined as the majority diagnosis given by three readers (13% discrepancies with image readers). It was further improved by establishing an overall reference standard that incorporated the clinical diagnosis (10% discrepancies with image readers). These principles of establishing reference standards may be applied to improve robustness of real-world systems supporting image-based diagnosis.

Introduction

Medical diagnosis has traditionally required examination by a physician. However, advances in imaging technology have affected specialties such as ophthalmology, dermatology, radiology, and cardiology to the point where clinical decision-making in these specialties is based largely on review of these imaging studies. Meanwhile, there have been persistent concerns about the accessibility of health care, particularly in rural and medically underserved areas. As a result, store-and-forward telemedicine strategies have emerged as a potential strategy for improving the delivery and cost of health care by replacing some in-person physician examinations with remote image-based evaluations.[1-2] Real-world implementation of these strategies will require appropriate validation of diagnostic accuracy, as well as agreement among different image readers.

At the same time, institutional and regulatory pressures are placing increased emphasis on quality and adherence to evidence-based practice guidelines.[3-4] Clinical examination by a physician is generally considered the gold standard in medical practice. However, there are often significant variations in diagnosis and management among physicians, even when they are presented with the exact same clinical scenarios.[5-6] Similarly, although image-based diagnosis is made from the appearance of structural and morphological features, numerous studies in ophthalmology have demonstrated that there may be significant discrepancies in image reading, even among experts looking at the same images.[7-9]

This variability among experts creates challenges for the implementation of image-based clinical information systems. For telemedicine systems, an accurate reference standard must be defined for proper validation, yet different remote image readers may disagree with regard to diagnosis. Furthermore, it is often unclear whether the

actual clinical examination or remote interpretation is more correct. For clinical research systems, it is critical to define a reference standard with the highest accuracy, yet there may be discrepancies between the interpretations of clinical examination and imaging data. Understanding the factors contributing to accurate diagnosis, as well as having a clear definition of a reference standard defining the correct diagnosis, is essential for managing image-based data for applications such as telemedicine and clinical research.

The purpose of this paper is to describe the development and evaluation of reference standards for image-based diagnosis, which combines the diagnostic impressions of multiple image readers with the actual clinical diagnoses by expert physicians. Sources of discrepancy will be identified and analyzed. Retinopathy of prematurity (ROP), an ophthalmic disease affecting low birth-weight infants during the first several months of life, is used as the study domain. Results from clinical ophthalmoscopic exams by an expert on a study cohort of infants are compared to results from image interpretation by three readers. In this way, we evaluate the variability among multiple readers performing image-based diagnosis, the impact of integrating diagnoses by multiple readers into an image-based reference standard for telemedicine applications, and the impact of integrating clinical diagnosis into an overall reference standard for clinical research applications.

Study Domain: Retinopathy of Prematurity (ROP)

ROP is diagnosed from dilated fundoscopic examination by an ophthalmologist, and there are established guidelines for identifying high-risk premature infants who need serial screening examinations.[10] When ROP occurs, approximately 90% of cases improve spontaneously and require only close follow-up examinations every 1-2 weeks. However, approximately 10% are at high risk for complications leading to blindness and require treatment.[11-12]

ROP has several characteristics that make it an ideal topic for research in telemedicine, biomedical informatics, and clinical research: (1) Diagnosis is based solely on the appearance of disease in the retina. (2) There is a universally-accepted, evidence-based, diagnostic classification standard for ROP.[13] (3) Although it is treatable if detected early, ROP continues to be a leading cause of childhood blindness throughout the world because of inadequacies in screening.[14] (4) Current ROP exam methods are time-intensive and physiologically stressful to infants. (5) Clinical expertise is often limited to larger academic centers, and is therefore unavailable at the point of care. Therefore, following the establishment of an appropriate reference standard, there should be diagnostic reliability and reproducibility of telemedical diagnosis via the application of objective diagnostic criteria. Furthermore, the physiologic stress on the infants may be reduced[(15), (16)DL, MM] Finally, while the expertise required to diagnose and treat ROP tends to be found only at large academic medical centers, the skill needed to acquire images adequate for diagnosis can be taught and disseminated across myriad practice settings and communities of varying size, thereby increasing access to care.[(17, 18, 19, 20)PC, RW, YM, DW]

Methods

Ophthalmoscopic Examination and Image Capture

This is a multicenter study with eight participating academic medical centers: (1) Oregon Health & Science University (OHSU), (2) Weill Cornell Medical College, (3) University of Miami, (4) Beaumont Health System, (5) Columbia University Medical Center, (6) Children's Hospital Los Angeles, (7) Cedars-Sinai Medical Center, and (8) Asociación para Evitar la Ceguera en México (APEC). Each institution's IRB approved the study protocol. Subject enrollment began in July 2011. All infants admitted to a participating Neonatal Intensive Care Unit (NICU) were eligible for the study if they met published criteria for ROP screening examination, or if they were transferred to the study center for specialized ophthalmic care.[10]

Study infants underwent serial examinations in accordance with the most recent evidence-based ROP guidelines.[10] Dilated ophthalmoscopic examinations were performed by an expert ophthalmologist, and findings were documented according to the international classification standard[13]. Retinal images were captured by a trained photographer after each eye examination using a wide-angle camera (RetCam; Clarity Medical Systems, Pleasanton, CA) using a standard protocol following manufacturer guidelines. De-identified clinical and image data were uploaded to a secure database (ASP.net, C#; State33, Portland, OR).

Image-based Reading

Remote image-based readings were conducted by three study authors (MFC, RVPC, SO) using an SSL-encrypted web-based grading module. In some cases, the image readers were the same ophthalmologists who had performed the ophthalmic examination. To best simulate ophthalmoscopy, where both eyes are examined sequentially before a final diagnosis is made, images from both eyes were displayed side-by-side (**Figure 1**). Demographic information, such as gestational age, postmenstrual age, and birth weight were also visible during image reading. Image readings were graded on an ordinal scale based on criteria from NIH-funded clinical trials: (1) No ROP; (2) Mild ROP; (3) Type-2 (moderate) ROP; and (4) treatment-requiring (severe) ROP,[11-12]

Development and Rationale for Reference Standards

Two reference standards were developed for this study: (1) Image-based reference standard, which was defined as the diagnosis given by a majority of the three readers. The rationale for this definition is that pooling the expertise of multiple image readers may improve the overall diagnosis. (2) Overall reference standard, which integrated the image-based reference standard with the actual clinical diagnosis provided by the examining ophthalmologist. The rationale for this definition is that combining information from clinical and telemedicine data may provide the most accurate diagnosis possible, and that this may be applicable in setting such as rigorous clinical research. In instances when there discrepancies between the image-based reference standard and the clinical diagnosis, all medical records reviewed by the three image readers and a moderator (KJ) to reach a consensus for the overall reference standard.

Data Analysis

Data were analyzed using spreadsheet software (Excel 2011; Microsoft, Redmond, WA). All records that had three image readers as of December 3, 2013 were analyzed. Additional analysis was conducted on the subset of these records that also had a submitted overall reference standard.

Inter-reader agreement in diagnostic classification was calculated for each pair of image readers using absolute agreement, kappa (κ) statistic for chance-corrected agreement, and weighted κ. Agreement was also investigated for the following comparisons: (1) individual readers vs. clinical diagnosis, (2) individual readers vs. image-based

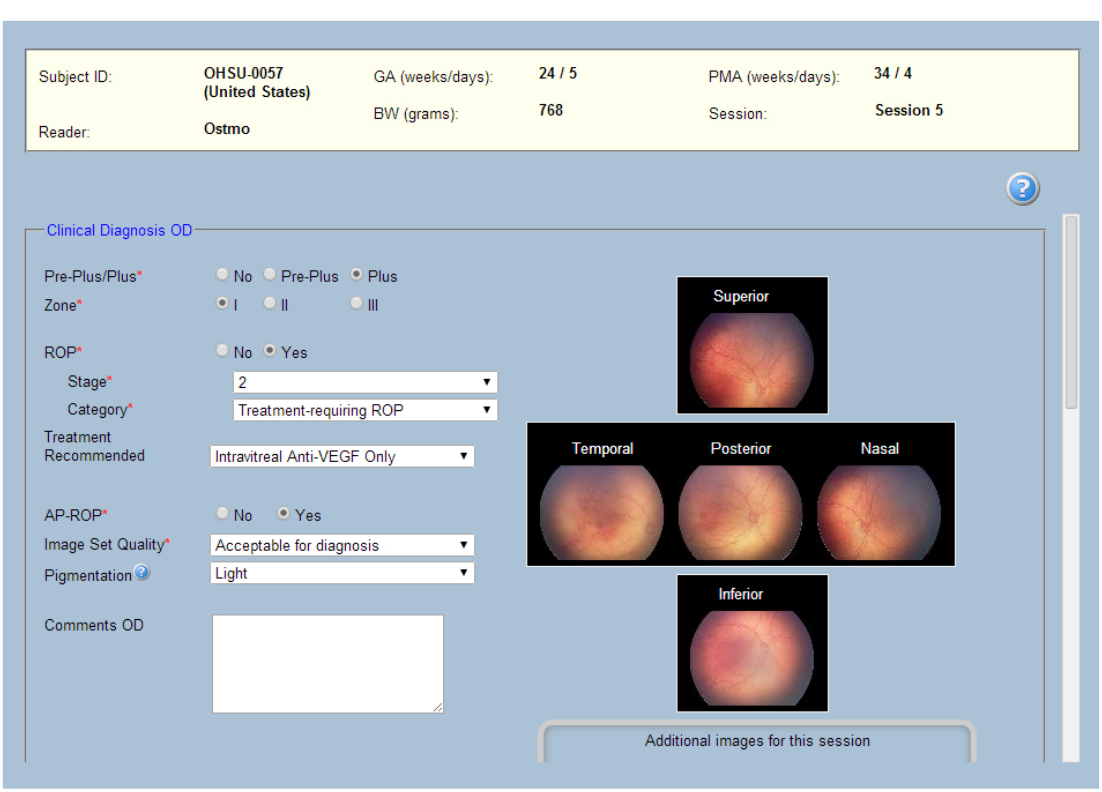


Figure 1. Example of web-based interface used for image evaluation and imaged-based diagnosis of ROP.

Table 1. Inter-reader agreement for ordinal ROP classification expressed as κ, weighted κ, and absolute agreement.

Reader Pair	κ (SE)	Weighted κ (SE)	Absolute Agreement (%)
1 vs 2	0.490 (0.027)	0.586 (0.024)	67%
1 vs 3	0.591 (0.026)	0.668 (0.023)	74%
2 vs 3	0.652 (0.025)	0.723 (0.021)	67%

reference standard, (3) individual readers vs. overall reference standard, (4) image-based reference standard vs. clinical diagnosis, (5) image-based reference standard vs. overall reference standard, and (6) clinical diagnosis vs. overall reference standard.

In all instances where there were diagnostic discrepancies with reference standards, all study data were reviewed by the authors (MCR, SO) and the reason for discrepancy was classified as either: (1) no ROP identified by ophthalmoscopic exam, (2) no ROP identified by image-based exam, (3) disagreement in classification of ROP severity ("stage"), (4) disagreement in classification of ROP location ("zone"), and/or (5) disagreement in classification of blood vessel appearance ("dilation" and "tortuosity").

Results

Characteristics of Study Population

A total of 150 infants who underwent 358 clinical exams met eligibility criteria for analysis. Both eyes of each infant underwent ophthalmoscopic examination at each visit, for a total of 716 study eyes. Based on clinical exam, 335 (47%) had no ROP, 283 (40%) had mild ROP, 67 (9%) had Type-2 ROP, and 31 (4%) had treatment-requiring ROP.

Inter-Reader Reliability and Reader Agreement with Clinical Exam

Table 1 summarizes agreement for each pair of readers based on ordinal ROP classification. Readers 1 & 2 had moderate agreement, while readers 1 & 3 and readers 2 & 3 had substantial agreement. Overall absolute agreement among the three readers was 73%. **Table 2** summarizes agreement between individual readers and the actual clinical diagnosis. Readers 1 and 2 demonstrated moderate agreement with the clinical diagnosis, while reader 3 had substantial agreement with the clinical diagnosis.

Agreement of Image-based Reference Standard with Individual Readers and Clinical Exam

Table 3 displays agreement for individual readers with the image-based reference standard. Reader 1 had moderate agreement with the image-based reference standard, while readers 2 and 3 had near-perfect agreement. Overall absolute agreement between the image-based reference standard and clinical diagnosis was 72%, while the κ (SE) and weighted κ (SE) were 0.551 (0.027) and 0.621 (0.024) respectively, indicating substantial agreement.

Agreement of Overall Reference Standard

Table 4 displays agreement for individual readers with the overall reference standard. There was near-perfect agreement for all three readers. Absolute agreement between the overall reference standard and clinical diagnosis was 81%. The κ (SE) and weighted κ (SE) were 0.679 (0.032) and 0.751 (0.027) respectively, indicating substantial agreement.

Table 2. Agreement between individual readers and clinical diagnosis for ordinal ROP classification expressed as κ and weighted κ statistics.

Reader	κ (SE)	Weighted κ (SE)	Absolute Agreement (%)
1	0.460 (0.028)	0.542 (0.026)	66%
2	0.465 (0.028)	0.556 (0.024)	66%
3	0.553 (0.027)	0.624 (0.024)	72%

Table 3. Agreement between individual readers and the image-based reference standard for ordinal ROP classification expressed as κ, weighted κ, and absolute agreement.

Reader	κ (SE)	Weighted κ (SE)	Absolute Agreement (%)
1	0.713 (0.023)	0.769 (0.020)	80%
2	0.773 (0.021)	0.824 (0.018)	85%
3	0.900 (0.015)	0.920 (0.012)	92%

Comparing the imaged-based reference standard to the overall reference standard, the κ (SE) was 0.911 (0.019) and the weighted κ (SE) was 0.931 (0.015). This signifies near-perfect agreement between the two standards.

Classification of Discrepancies with Reference Standards

Among 434 eye exams in this study with an overall reference standard, there were 14 (3%) discrepancies with the image-based reference standard derived from majority vote among the three image graders. When these medical records were reviewed, it was discovered that 6/14 discrepancies were due to disagreements over disease location, 5/14 were due to disagreements over blood vessel morphology, in 2/14 cases the image-based reference standard did not identify any signs of ROP, and 1/14 was due to a disagreement in disease severity. **Table 5** summarizes the absolute number of discrepancies and discrepancy rates for each study comparison.

Discussion

Summary of Key Findings

To our knowledge, this is the first study to develop and evaluate the accuracy of reference standards in ophthalmology that integrate clinical diagnosis with a consensus image-based diagnosis from multiple image readers. The performance of the reference standard was evaluated against the telemedical ROP diagnosis of three image readers, the clinical exam, and an image-based reference standard. The key findings from this study are: (1) There is imperfect agreement in image-based diagnosis among experts, and between image-based and clinical diagnoses. (2) Using multiple image readers is a potentially useful approach to increase the accuracy and reliability in telemedical diagnosis of ROP. (3) Use of an overall reference standard that integrates information from the clinical exam with image-based diagnoses may be of value in image-based clinical research.

Telemedicine holds great promise for improving the accessibility and quality of health care.[2,21] However, variation in telemedical image interpretation and the lack of definitive reference standards have presented challenges to the implementation of telemedicine systems for the diagnosis and management of retinopathy of prematurity (ROP).[7]

Inter-reader Agreement

This study confirms findings from previously published research showing that inter-reader agreement in image-based ROP diagnosis is imperfect (**Table 1**). For example, a previous study examining interphysican agreement in the telemedical diagnosis of ROP found a weighted κ statistic that ranged from 0.38 (fair agreement) to 0.81 (near-perfect agreement).[22] This prior study's higher maximum value for κ may be a consequence of only using ophthalmologists for image interpretation, whereas one of the image readers in the current study was not a physician.[23] However, inter-reader agreement for the non-ophthalmologist reader (reader 3) was comparable to what

Table 4. Agreement between individual readers and the overall reference standard for ordinal ROP classification expressed as κ and weighted κ statistics.

Reader	κ (SE)	Weighted κ (SE)	Absolute Agreement (%)
1	0.802 (0.026)	0.850 (0.021)	88%
2	0.829 (0.025)	0.877 (0.019)	90%
3	0.839 (0.024)	0.877 (0.020)	90%

Table 5. Summary of absolute number of discrepancies and the discrepancy rate for each comparison in the study.

Comparison	No. of Eye Exams	No. of Disagreements	Disagreement Rate
Individual Readers vs Clinical Exam	2148	689	32%
Individual Readers vs Image-based Reference Standard	2148	269	13%
Individual Readers vs Overall Reference Standard	1302	133	10%
Clinical Exam vs Image-based Reference Standard	716	198	30%
Clinical Exam vs. Overall Reference Standard	434	82	19%
Overall Reference Standard vs. Image-based Reference Standard	434	14	3%

was observed for the two ophthalmologist readers. Interestingly, agreement between the non-ophthalmologist reader and the clinical diagnosis was actually higher than what was observed for the ophthalmologist readers.

The degree of inter-reader agreement noted in the current study is also consistent with what has been observed in non-ROP related ophthalmologic studies investigating image-based diagnosis. In a major multicenter study involving diabetic retinopathy, the weighted κ for intergrader reliability was 0.41 (moderate agreement) to 0.80 (substantial agreement), depending on the type of retinal lesion being observed.[24] In a study examining interobserver agreement in the diagnosis of age-related macular degeneration based on fluorescein angiography imaging methods, the κ was 0.37 to 0.40 (fair agreement).[25]

Additional published literature suggests that this finding is generalizable across other medical domains. In one dermatologic study examining agreement in the evaluation and diagnosis of skin tumors the κ statistic was 0.32 (fair agreement).[26] In a radiographic study comparing expert radiologists and pulmonologists in their diagnoses of upper lobe-predominant emphysema the κ was 0.20 (slight agreement) to 0.60 (moderate agreement).[27]

When considered in conjunction with the existing evidence base, findings from this study suggest that the reliability in the telemedical diagnosis of ROP, while imperfect, is comparable or better than interobserver agreement for other ophthalmic or medical diagnoses. This supports the validity of telemedicine programs for ROP diagnosis.

Image-Based Reference Standard Combining Multiple Readers

While the telemedicine literature examining inter-reader agreement and reader agreement against a gold standard is relatively robust, data investigating the utility of pooling telemedical diagnoses is sparse. A 2010 study investigating the accuracy of "non-expert" graders in diagnosing ROP noted slight improvements in sensitivity and specificity when using the majority diagnosis of the three best non-expert readers, but the results were not statistically significant.[28] They found that sensitivity in diagnosing treatment-requiring ROP increased from 0.82 to 1.00, while specificity increased from 0.92 to 0.94.[28] Similarly, a 2013 teledermatology study investigating the diagnostic accuracy of remote reflectance confocal microscopy found that sensitivity was improved by more than eight percentage points by combining reader diagnoses.[29]

This study confirms and expands upon these earlier findings (**Tables 1 and 3**). By merging the image-based diagnoses of the individual readers to form the image-based reference standard, the reliability of telemedical diagnosis of ROP was significantly improved. The range of individual reader's weighted κ increased from 0.586 – 0.723 for inter-reader agreement to 0.769 – 0.920 for agreement with the image-based reference standard. Importantly, there was little-to-no improvement in the agreement of the clinical diagnosis with the image-based reference standard, with only a 2-percentage point increase in absolute agreement (69% vs. 71%).

Overall Reference Standard Combining Clinical and Image-Based Diagnoses

While the establishment of an image-based reference standard significantly improved the reliability of telemedical diagnosis, the relative lack of improvement in agreement between clinical and the image-based diagnoses suggests the need for further refinement of potential reference standards. A reference standard that allows for direct comparison of image-based and ophthalmoscopic diagnoses would be particularly useful in clinical research, where both diagnostic approaches need to be evaluated simultaneously. We created such a reference standard by

incorporating the clinical diagnosis with the image-based reference standard. The utility of this approach is demonstrated by the increased levels of agreement (**Table 4)** and the decreased levels of discrepancies (**Table 5**). Each time more information was aggregated together to form a new reference standard there were commensurate increases in reliability and accuracy of the telemedical diagnosis.

Limitations

There are several limitations of this study. (1) While the data in this study were analyzed by eye, the classification and diagnosis of ROP in the right and left eyes of the same patient is not independent. At the time of image interpretation, images of both eyes were presented to the study readers simultaneously so as to simulate ophthalmoscopy, where both eyes are examined together. This approach minimizes bias that might favor either examination and allows for the analysis of both eyes of the infant. (2) There was no standardization of image reading conditions, such as resolution, contrast, and luminance. Previous radiographic studies have demonstrated the ability of these parameters to affect diagnostic accuracy.[30] (3) While two of the readers involved in this study have extensive clinical experience with ROP (RVPC, MFC), the third was a research coordinator who was trained to interpret images (SO). While the use of "non-expert" readers is common in the literature, it is unclear what affect, if any, this may have had on the study findings. Future studies examining the influence of reader background on inter-reader reliability and diagnostic accuracy may be revealing. (4) The sample size of three image readers, while consistent with the published literature, is still relatively small. As such, it is difficult to make generalizations to the broader community of ophthalmologists. (5) As some of the ophthalmic examinations included in this study were performed by the image readers, there is the potential for recall bias. However: (a) Study images were collected at eight sites, and readers only worked at two of those sites. (b) Study images were reviewed by readers several months after clinical exams. (c) No clinical data beyond the retinal images and basic demographic information (e.g. birth weight & gestational age) were shown to readers about subjects. For these reasons, we suspect that the impact of recall bias is minimal.

Conclusions

Telemedicine has the potential to improve the quality, cost, and accessibility of medical care, particularly in image-oriented specialties. Previous research investigating the accuracy and reliability of telemedicine systems in ophthalmology has typically compared telemedical diagnosis to the current gold standard of a dilated fundoscopic exam by an ophthalmologist. However, defining absolute reference standards for the diagnosis of ROP has been difficult and has hindered the implementation of image-based clinical information systems. This study demonstrates that the reliability and accuracy of ROP diagnosis can be improved through the implementation an overall reference standard that integrates diagnostic information from the clinical examination and an image-based telemedicine system. Findings from this study have important implications for groups designing and implementing image-based telemedicine systems and for future research aimed at improving the accuracy of ROP diagnosis.

References

1. Grigsby J, Sanders JH. Telemedicine: where it is and where it's going. Ann Int Med 1998; 129: 123-7.
2. Bashshur RL, Reardon TF, Shannon GW. Telemedicine: a new health care delivery system. Annu Rev Public Health 2000; 21: 613-37.
3. Sittig DF, Singh H. Electronic health records and national patient safety goals. New Engl J Med 2012; 367: 1854-60.
4. Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. New Engl J Med 2010; 363: 501-4.
5. Peabody JW, Luck J, Glassman P. Comparison of vignettes, standardized patients, and chart abstraction: a prospective study of 3 methods for measuring quality. JAMA 2000; 283: 1715-22.
6. Veloski J, Tai S, Evans AS, Nash DB. Clinical vignette-based surveys: a tool for assessing physician practice variation. Am J Med Qual 2005; 20: 151-7.
7. Chiang MF, Jiang L, Gelman R, et al. Inter-expert agreement of plus disease diagnosis in retinopathy of prematurity. Arch Ophthalmol 2007; 125:875-80.

8. Moss SE, Klein R, Kessler SD, Richie KA. Comparison between ophthalmoscopy and fundus photography in determining severity of diabetic retinopathy. Ophthalmology 1985; 92: 62-7.
9. Kinyoun JL, Martin DC, Fujimoto WY, Leonetti DL. Ophthalmoscopy versus fundus photographs for detecting and grading diabetic retinopathy. Invest Ophthalmol Vis Sci 1992; 33: 1888-93.
10. Fierson WM, American Academy of Pediatric, American Academy of Ophthalmology, et al. Screening examination of premature infants for retinopathy of prematurity. Pediatrics 2013; 131: 189-95.
11. Cryotherapy for ROP Cooperative Group. Multicenter trial of cryotherapy for retinopathy of prematurity: preliminary results. Arch Ophthalmol 1988; 106: 471-9.
12. Early Treatment for ROP Cooperative Group. Revised indications for the treatment of ROP. Arch Ophthalmol 2003; 121: 1684-94.
13. Committee for the classification of retinopathy of prematurity. The international classification of ROP revisited. Arch Ophthalmol 2005; 123: 991-9.
14. Steinkuller PG, Du L, Gilbert C, et al. Childhood blindness. J AAPOS 1999; 3: 26–32.
15. Laws DE, Morton C, Weindling M, Clark D. Systemic effects of screening for retinopathy of prematurity. Br J Ophthalmol. 1996;80(5):425-8.
16. Moral-pumarega MT, Caserío-carbonero S, De-la-cruz-bértolo J, Tejada-palacios P, Lora-pablos D, Pallás-alonso CR. Pain and stress assessment after retinopathy of prematurity screening examination: indirect ophthalmoscopy versus digital retinal imaging. BMC Pediatr. 2012;12:132.
17. Paul Chan RV, Williams SL, Yonekawa Y, Weissgold DJ, Lee TC, Chiang MF. Accuracy of retinopathy of prematurity diagnosis by retinal fellows. Retina (Philadelphia, Pa). 2010;30(6):958-65.
18. Wong RK, Ventura CV, Espiritu MJ, et al. Training fellows for retinopathy of prematurity care: a Web-based survey. J AAPOS. 2012;16(2):177-81.
19. Murakami Y, Jain A, Silva RA, Lad EM, Gandhi J, Moshfeghi DM. Stanford University Network for Diagnosis of Retinopathy of Prematurity (SUNDROP): 12-month experience with telemedicine screening. Br J Ophthalmol. 2008;92(11):1456-60.
20. Weaver DT. Telemedicine for retinopathy of prematurity. Curr Opin Ophthalmol. 2013;24(5):425-31.
21. Ekeland AG, Bowes A, Flottorp S. Effectiveness of telemedicine: a systematic review of reviews. Int J Med Inform. 2010;79(11):736-71.
22. Scott KE, Kim DY, Wang L, et al. Telemedical diagnosis of retinopathy of prematurity intraphysician agreement between ophthalmoscopic examination and image-based interpretation. Ophthalmology. 2008;115(7):1222-1228.e3.
23. Chiang MF, Keenan JD, Starren J, et al. Accuracy and reliability of remote retinopathy of prematurity diagnosis. Arch Ophthalmol. 2006;124(3):322-7.
24. Early Treatment Diabetic Retinopathy Study Research Group. Grading diabetic retinopathy from stereoscopic color fundus photographs—an extension of the modified Airlie House classification. ETDRS report number 10. Ophthalmology 1991; 98(suppl):786 – 806.
25. Holz FG, Jorzik J, Schutt F, et al. Agreement among ophthalmologists in evaluating fluorescein angiograms in patients with neovascular age-related macular degeneration for photo- dynamic therapy eligibility (FLAP-Study). Ophthalmology 2003;110:400 –5
26. Phillips CM, Burke WA, Allen MH, Stone D, Wilson JL. Reliability of telemedicine in evaluating skin tumors. Telemed J. 1998;4(1):5-9.
27. Hersh CP, Washko GR, Jacobson FL, et al. Interobserver variability in the determination of upper lobe-predominant emphysema. Chest. 2007;131(2):424-31.
28. Williams SL, Wang L, Kane SA, et al. Telemedical diagnosis of retinopathy of prematurity: accuracy of expert versus non-expert graders. Br J Ophthalmol. 2010;94(3):351-6.
29. Rao BK, Mateus R, Wassef C, Pellacani G. In vivo confocal microscopy in clinical practice: comparison of bedside diagnostic accuracy of a trained physician and distant diagnosis of an expert reader. J Am Acad Dermatol. 2013;69(6):e295-300.
30. Herron JM, Bender TM, Campbell WL, Sumkin JH, Rockette HE, Gur D. Effects of luminance and resolution on observer performance with chest radiographs. Radiology. 2000;215(1):169-74.

Acknowledgements
Supported by grant EY19474 from the National Institutes of Health, Bethesda, MD (MFC, RVPC), and by unrestricted departmental funding from Research to Prevent Blindness, New York, NY (MCR, SO, KJ, RVPC, MFC). The St. Giles Foundation (RVPC), Departmental Grant from Research to Prevent Blindness (RVPC, KEJ), The iNsight Foundation (RVPC, KEJ)

Address for Correspondence
Michael F. Chiang, MD
Departments of Ophthalmology & Medical Informatics and Clinical Epidemiology
Oregon Health & Science University
3375 SW Terwilliger Boulevard
Portland, OR 97239
Tel: 503-418-3087 | Fax: 503-494-5347 | Email: chiangm@ohsu.edu

Developing a Formal Representation for Medication Appropriateness Criteria

Hojjat Salmasian MD MPH[1], Tran H Tran Pharm. D[2,3], Carol Friedman PhD[1]

1. Columbia University, New York, NY 2. NewYork-Presbyterian Hospital, New York, NY 3. St. John's University, Queens, NY

Abstract

Inappropriate medications use (IMU) is a serious issue of global concern that leads to a waste of resources and potentially harms the patients. IMU can usually be identified by extracting information about the patient's conditions and treatments, and comparing them with "medication appropriateness criteria". To enable automation of these criteria, we developed a formal representation for them, which we called Objective Medication Appropriateness Criteria (OMAC). OMAC represents four aspects of the criteria: trigger, rules, action and meta-data. Our evaluation showed that OMAC can completely represent explicitly defined medication appropriateness criteria using links to external knowledge sources. OMAC is the first formal representation for medication appropriateness criteria, and will enable development of structured rules for appropriate use of medications that can be implemented using standards for clinical decision support.

Introduction

Inappropriate medication use (IMU) is a serious issue of global concern that not only leads to a waste of healthcare resources, but also potentially harms patients due to inadvertent side effects [1–3]. Previous studies have shown that several groups of medications are commonly subject to inappropriate use, including antibiotics, antidepressants, antipsychotics, bronchodilators, non-steroidal anti-inflammatory drugs (NSAIDs), proton pump inhibitors (PPIs), and statins [4–8]. Numerous studies developed methods for identifying and reducing IMU, with the focus of their intervention spanning from healthcare professionals and patients, to financial, organizational, and regulatory approaches [9,10]. Typically, these approaches rely on manual identification of IMU. Developing automated methods to reduce IMU is challenging. Inappropriate use of medications is still considered an understudied problem in general [11,12].

Manual identification of IMU requires extracting information about the patient, the medication, and other treatments, and comparing them with "medication appropriateness criteria", which are standards that define appropriate use of medications [13]. Consequently, developing automated solutions to reduce IMU entails two requirements: a framework to represent the medication appropriateness criteria formally, and methods to extract the information needed to compute these criteria. This article focuses on developing a framework for formal representation of medication appropriateness criteria.

Previous researchers have identified several medication appropriateness criteria and metrics through systematic review of the literature [12,14]. These criteria can be categorized into three groups: the first group enumerates the conditions in which the use of medication is appropriate (e.g. see Choudhrey *et al.*'s criteria for appropriate use of proton pump inhibitors (PPIs)[15]), the second group lists conditions in which the use of medication is deemed inappropriate (e.g. see the Beers' criteria[16]), and the third group provides a combination of both (e.g. see Oborne *et al.*'s criteria on appropriate use of neuroleptics[17]). Since older adults are frequently subject to polypharmacy and therefore more likely to experience the negative impacts of IMU (e.g. drug-drug interactions, adverse drug reactions and increased risk of hospitalization) [18–20] larger collections of medication appropriateness criteria exist for the geriatric population. Examples include the Beers' criteria[16] and the Screening Tool of Older Persons' potentially inappropriate Prescriptions (STOPP)[21] which aim to reduce inappropriate use of medications (overuse), and the Screening Tool to Alert doctors to Right, i.e. appropriate indicated Treatment (START)[22] which promotes appropriate use of medications that are omitted (underuse).

Medication appropriateness criteria can be described as a special form of clinical guidelines, although they have distinct features that separate them from the majority of clinical guidelines. Clinical guidelines provide best

practices for diagnosis and therapy of diseases, but medication appropriateness criteria are focused on proper utilization of a resource (namely, medications). Clinical guidelines are primarily developed by major medical associations, are organized in a common format and are hosted on repositories such as the National Guideline Clearinghouse [23]; in contrast, medication appropriateness criteria are mostly developed by independent groups of researchers and distributed without using a common format or central repository.

Medication appropriateness criteria are currently only available in narrative form, and transforming them into a computable format is challenging because a formal representation for the components of medication appropriateness criteria does not exist. Different criteria have varying levels of granularity and specificity in defining the medications, diagnoses, and symptoms; in addition, some but not all of the criteria are accompanied by information regarding the level of evidence, target population, or extent of clinical relevance. A framework in which these criteria can be explicitly and comprehensively represented is needed. We developed such a representation framework, which we call the Objective Medication Appropriateness Criteria (OMAC).

Methods

In order to study existing medication appropriateness criteria, we started by identifying these criteria by searching PubMed using the following keywords and their variations to identify published medication appropriateness criteria: inappropriate prescribing, overuse, overtreatment, overutilization, and utilization review. We grouped the relevant studies based on the actual criteria they used to identify IMU. We then curated a collection of published medication appropriateness criteria and used a random subset of those to develop OMAC (for examples, see Table 1). To ensure that we didn't have any biases in our component selection and semantic aggregation of the concepts, we used another independent set of criteria for evaluation of OMAC.

Developing OMAC

We manually analyzed a randomly selected subset of medication appropriateness criteria to identify their components. Each criterion can be described as one or more rules, and the purpose of OMAC was to provide a formal representation for these rules as well as any other aspects of the criteria. We semantically grouped the components we found in the sample criteria to define concepts that comprise the criteria, including high-level elements (such as the general sections of a criterion) and low-level elements (such as modifiers, identifiers, names, etc.) and we also identified the relationships between these concepts and represented them in OMAC.

Different medical concepts are frequently mentioned in the medication appropriateness criteria, such as medications and diseases. Since the purpose of OMAC was to provide a representation for the criteria, and not to enumerate each individual medication or disease, we ensured that OMAC takes advantage of previously developed ontologies and terminologies, by linking to external ontologies and terminologies to the extent possible. We saved OMAC using frames and properties in Protégé version 3.5.

Evaluating OMAC

After the initial design of OMAC was completed, we presented a separate set of 10 medication appropriateness criteria to a group of domain experts (physicians and pharmacists) in form of a questionnaire, asking them to identify and categorize the components of these criteria independently. Each item in the questionnaire consisted of one medication appropriateness criterion statement in its original narrative form, and requested that the participant breaks the statement into basic elements (such as medication names, medication class names, disease names, logical statements, or temporal modifiers). Disagreements in the experts' responses were identified through qualitative analysis of the responses. Three types of disagreements were considered: differences in the classification of the same word or phrase (e.g. classifying "hypertension" as a disease versus a problem), differences in specification of the elements in the statements (e.g. considering "severe hypertension" as two separate concepts versus one), and classification of terms into concepts that are not explicit (e.g. classifying "long-term use of drug X" as one concept of type "overuse"). In a subsequent questionnaire, we presented the experts with the same narrative criteria but clearly marked these areas of disagreement and asked the experts to translate those terms and phrases into more detailed, explicitly defined concepts. Note that the purpose of this process was not to reach perfect agreement, but rather to identify what "elements" constitute the criteria and also to describe the elements so that they are well-defined, so that we can evaluate OMAC's coverage for those elements.

	Source	Narrative criterion
1	2002 Beers' criteria	Disease: Seizures or epilepsy Drug: Clozapine, chlorpromazine, thioridazine, and thiothixene Concern: May lower seizure thresholds Severity Rating: High
2	STOPP, section A, item 2	Loop diuretic for dependent ankle edema only i.e. no clinical signs of heart failure
2	STOPP, section B, item 3	TCA's with cardiac conductive abnormalities
3	STOPP, section A, item 6	Beta-blocker in combination with verapamil
4	STOPP, section A, item 7	Use of diltiazem or verapamil with NYHA Class III or IV heart failure
5	STOPP, section B, item 9	Use of aspirin and warfarin in combination without histamine H2 receptor antagonist (except cimetidine because of interaction with warfarin) or proton pump inhibitor

Table 1 – Examples of medication appropriateness criteria previously published in the literature. These examples are all adopted from Beers' criteria[16] and STOPP[24] and each criteria expresses what medication should be *avoided* in the provided context. Other examples from other sources were also used in the development and evaluation of OMAC (not shown in the table). The narrative criteria are shown as they appear in the original source. *TCA: tricyclic antidepressant; NYHA: New York Health Association; H2 receptor: histamine 2 receptor.*

Subsequently, we evaluated whether OMAC could represent all of the explicitly defined concepts provided by the experts. We froze the development of OMAC before we started sending out the questionnaires, to ensure that our knowledge of the results of the previous step would not affect our evaluation of OMAC's completeness. We planned to correct OMAC for any areas of deficiency that would be found throughout this evaluation, only after the evaluation was completed.

Results

We identified 110 medication appropriateness criteria through literature review, and used a random subset of 40 to develop OMAC. We designed OMAC such that each criterion in this subset could be represented using four types of information: '*trigger'*, '*rules'*, '*action'*, and '*meta-data'*. The *trigger* may consist of one or more medications that are the primary focus of the criterion (when prescribing these medications the criterion would be triggered) or one or more clinical conditions in which the use of a certain medication is desirable (in this case the criterion would focus on underuse). *Rules* specify the conditions that a patient must meet to be eligible for the criterion (such as age limit, past medical history, medications prescribed, symptoms, or paraclinical findings). *Action* specifies the recommendation that the criterion makes once the patient meets all the rules; generally, actions are in two forms, either to *avoid* prescribing a medication or to *consider* prescribing a medication. *Meta-data* includes all the additional information that is used to describe the criterion (examples include a name or unique identifier, references to citations, or a justification or concern). As an example, Beers' criteria not only lists medications or combinations of drugs that should be avoided in the elderly, but also specifies what "concern" exists around using these medications, and also provides a "severity rating" for this concern (low vs. high) to help the clinicians determine the importance of each item in this criteria (Table 1) [16]. We represented the trigger, action and meta-data components using properties for the "criterion" class (Figure 1, right). We used a more complex classification as described below to represent the rules.

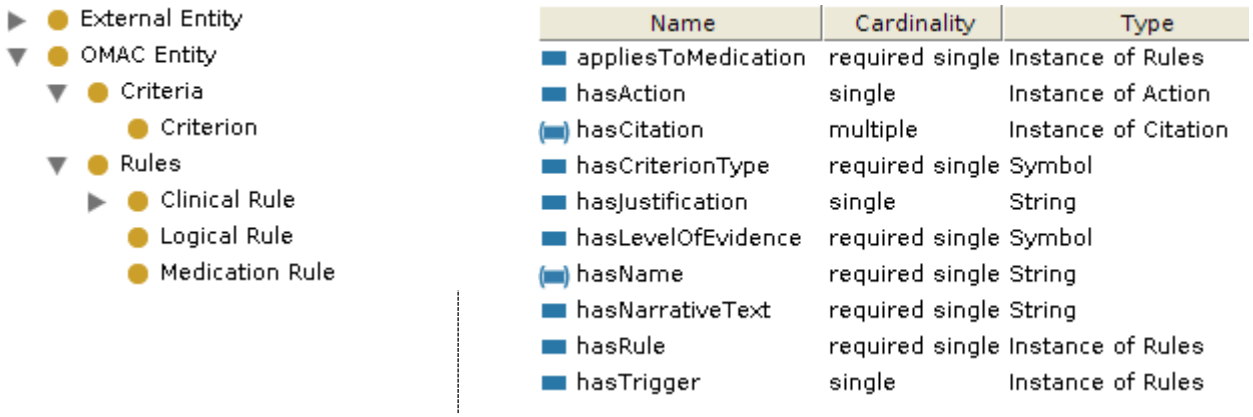


Figure 1 – Main concepts in the OMAC (left) and the properties of the *criterion* class (right).

Each criterion can contain one or more rules, and there are various types of rules in different criteria. These include *'medication rules'* which specify the medication that is the subject of the criterion as well as co-prescribed medications that need to be considered, and '*clinical rules'* which specify the diseases, symptoms, laboratory tests results, and demographics that have to be present or that should be absent for the patient to meet the criterion. This can be clarified using the third example shown in Table 1: "TCA's with cardiac conductive abnormalities"; this item from STOPP criteria states that in elderly patients who have cardiac conductive abnormalities, tricyclic antidepressants (TCAs) should be avoided because of their pro-arrhythmic effects [24]. To apply this criterion to a patient, three rules must be satisfied: (i) the patient must belong to the 'elderly' demographic group (formally defined as age ≥ 65 years), (ii) the patient must have been diagnosed with a cardiac conductive disorder (including, but not limited to Type I heart block, Type II heart block, or right bundle branch block), and (iii) the patient must have been prescribed a medication that belongs to the TCA class. The first two rules in this example are clinical rules, and the latter is a medication rule.

Concept Type	**External ontology or terminology**
Medication	RxNorm, ATC, NDC
Disease	ICD, SNOMED CT
Symptom	SNOMED CT, Symptom Ontology
Procedure	CPT, ICD, SNOMED CT

Table 2 – Examples of external ontologies and terminologies that can be used to define concepts that are contained in a clinical or medication statement, as part of a medication appropriateness criterion. *ATC: Anatomical Therapeutic Classification; NDC: National Drug Code; ICD: International Classification of Diseases; SNOMED CT: Systematized Nomenclature of Medicine, Clinical Terms; CPT: Current Procedural Terminology.*

Clinical and medication rules have different properties: clinical rules may focus on the existence, temporality and duration of a clinical finding or condition, or the value of a measurement, but medication rules may specify the dose, route, frequency and form of a medication. Both clinical and medication rules may include concepts that are externally defined in other ontologies or terminologies (Table 2). In the example provided above, "cardiac conductive abnormalities" can be represented as a clinical rule, which can refer to a pertinent concept in International Classification of Diseases, version 10 (ICD-10) or Systematized Nomenclature of Medicine, Clinical Terms (SNOMED CT), and thus, it is possible to link the concept to a standardized external knowledge source. A link to an external concept consists of four parts: the local name of the concept (e.g. 'cardiac conductive

abnormalities'), the name of the external ontology or terminology, including version number (e.g. ICD-10), URL of the external ontology or terminology (e.g., http://purl.bioontology.org/ontology/ICD10) and the unique identifier of the corresponding concept in that external ontology or terminology (in this case 'I44').

Clinical and medication rules can be combined with each other using *'logical rules'*. Each logical rule has a mandatory field which specifies the Boolean operator it is representing ('AND', 'OR', or 'NOT'). In the example above, the clinical and medication rules are combined using a logical rule with 'AND' logic (i.e. the patient must be among the elderly AND have a cardiac conductive disorder to be eligible for this criterion).

We grouped all the three aforementioned types of rules under a parent class called *'rules'* (Figure 1, left). To represent complex statements, these rules can be nested to create 'rule trees'. Clinical and medication rules can only appear as the leaves of the rule tree. Logical rules appear as branches of the tree, and each logical rule references one or more rules of any type. The latter enables nested rules which allow representation of complex logical statements. The last example in Figure 1 (STOPP, item B9) demonstrates a criterion with a complex logic. This complex statement can be encoded through nesting different types of rules, as shown in Figure 2.

Evaluating OMAC

Eight domain experts collaborated in the first questionnaire. There was no disagreement among experts in their responses for simple and well-defined criteria; for instance, all collaborators described STOPP criteria item A6 (shown in Table 1) using similar components. We observed disagreements in the way more complex criteria were broken down; for example, there was lower agreement on how the terms 'dependent ankle edema' and 'no clinical sign of heart failure' were categorized by different experts. When experts clarified the areas of vagueness using detailed explicit concepts, we noticed that although they clarified these vague terms using different sets of explicit concepts, they used similar 'types' of concepts to describe them. For example, each expert used a different set of 'signs', 'symptoms' and 'paraclinical findings' to describe the phrase 'clinical signs of heart failure', but all experts used exactly those three types of information.

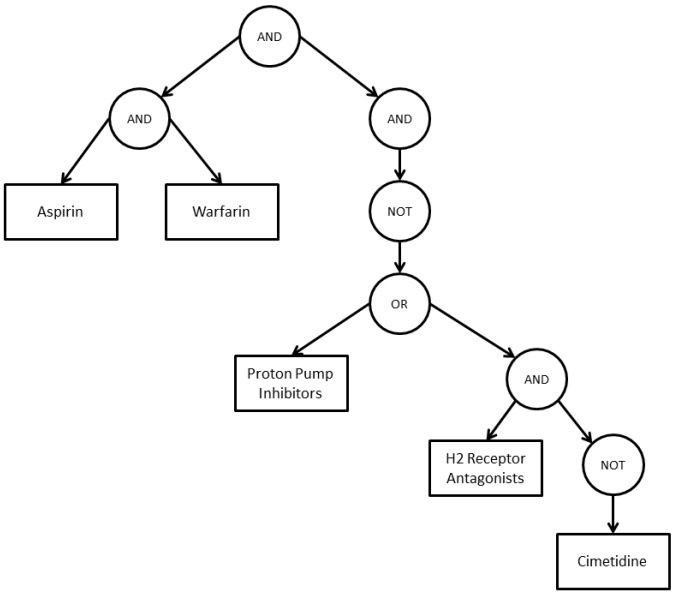


Figure 2 – A schematic diagram showing nesting of logical statements representing a complex criterion. For simplicity, logical statements are shown as circles that are labeled based on the Boolean operator associated with them, and medication statements are shown as boxes corresponding to the pertinent medication or medication class. Each of the medication statements may refer to an external ontology or terminology which formally defines the specific medication or drug class; those links are not shown in this figure. The narrative form of this criterion is available in Table 1.

All of the types of information that experts used to transform the vague phrases into explicit forms corroborated with the types of information that we had already incorporated into OMAC's clinical or medication rules. In other words, OMAC had completed coverage for all of the criteria that were coded by the experts, and as a result we did not modify OMAC after this evaluation.

Discussion

Developing a representation format for medication appropriateness criteria is the first step towards developing computable, interchangeable and reusable solutions to prevent inappropriate medications use. OMAC formally defines the structure of explicitly defined medication appropriateness criteria, and allows referencing to external ontologies and terminologies when applicable.

The results of our questionnaire study indicate that at least some of the medication appropriateness criteria are defined using vague terms that were interpreted differently by the experts. These criteria only provide guidelines for appropriate use of medications, and variability in the application of guidelines is a well-established phenomenon in health care practice; however, ideally the guideline itself should be interpreted identically by all of its users so that the variability should be only due to the specific characteristics of the patient or the settings in which the guideline is used, and not due to different interpretations of the appropriate care[25,26]. Although our questionnaire study has a small sample size, it signifies the need for well-defined medication appropriateness criteria. OMAC can facilitate this process, as encoding the criteria into OMAC requires translating all terms into explicitly defined medication, clinical or logical rules.

OMAC is designed to be flexible, and allow for multiple ways of defining concepts and their relationships. Through the use of logical rules, it is possible to model the steps that are used to implement medication appropriateness criteria in clinical practice and encode these steps in a computable way. When a clinical or medication concept is in fact referring to a class of diseases or medications, logical steps can be used to internally define these sets instead of referencing external knowledge sources, which is important when defining a concept that does not exist in any external knowledge source. Therefore, the user has the choice of either specifying a medication class by referencing an external entity, or by defining external references to each member of that class and then combining them using an 'OR' logic (Figure 3). Each approach has its own advantages: using an external reference for each of the elements in that class makes the local definition of the criteria more explicit, while using an external reference for the class itself reduces the amount of effort needed to encode the criteria in OMAC.

Using an external ontology or terminology to define the concepts in OMAC also has the advantage of reusing knowledge that has been vetted by a group of experts, but a suitable external knowledge source may not be available in all cases, or it may not be as accurate or complete. In addition, not all of the concepts that are found in medication appropriateness criteria can be identically found in external knowledge sources. For instance, one medication appropriateness criteria may specify the severity levels for heart failure using the classification provided by the New York Health Association (Table 1), but this classification may not be already defined in any existing disease ontologies and terminologies. OMAC flexibly supports defining these complex concepts either by external links (when possible) or locally, and the users can choose their preferred method based on the task at hand.

OMAC is different from a guideline representation language. While medication appropriateness criteria can be described as a special form of guidelines, guideline representation languages (e.g. GEM[27], GLIF[28], EON[29], PRO*forma*[30], and SAGE[31]) do not enforce the mandate level of detail in their formalism that is needed for representing medication appropriateness criteria. Guideline representation languages provide a structured way to encode the "flow" of decisions in a guideline. However, to ensure that they can support different types of decision and various forms of guidelines, they provide a significant amount of flexibility as to how each decision step is defined. Previous research has shown that these guideline representation models have limitations when applied to medication related guidelines used for chemotherapy, and that representing medication related guidelines as rules can address this limitation.[32] OMAC combines this rule-based approach with specific features of guideline representation language (such as the inclusion of meta-data about provenance of the guidelines), to provide a more strict structure to represent the medication appropriateness criteria than guideline representation models, thereby providing a common framework for encoding all such criteria in a similar, interchangeable way. In that sense, OMAC complements the guideline representation languages by providing the formalism that is necessary for a certain type of decisions, namely the decision about appropriateness of medications.

One potential challenge in interchanging OMAC-encoded criteria is that a criterion may be encoded using an external ontology or terminology which may be different from what is desirable for a second user of the criterion. This challenge can be addressed by creating cross-walks between these external knowledge sources; in many cases, this can be easily possible using the Unified Medical Language System (UMLS). Finally, our study is also limited in that we did not conduct a large scale evaluation of the completeness of OMAC. We intend to address this limitation in future research. We also intend to use OMAC to develop structured representations of well-established medication appropriateness criteria and then export them into a format supported by HL7 Clinical Decision Support (CDS) standards. Namely, we intend to use the virtual medical record (vMR) format[33] to represent the patient data, and use the OpenCDS platform[34] to integrate the computable medication appropriateness criteria with the medical records and evaluate the accuracy and impact of using this approach to provide decision support regarding appropriateness of medications**.**

Conclusions

OMAC provides the necessary flexibility for defining concepts using external ontologies or terminologies whenever applicable, and through the use of rules, it enforces the necessary formalism to ensure that all essential concepts of the medication appropriateness criteria are represented using a common structure. OMAC is the first framework that specifies encoding the medication appropriateness criteria into a formal, structured form, which is necessary to incorporate a decision support component aimed at reducing IMU.

Acknowledgements

This work has been supported by the National Library of Medicine grants R01 LM010016, R01 LM010016-0S1, R01 LM010016-0S2, R01 LM008635, and 5 T15 LM007079. All authors reviewed and approved the final draft. HS developed the framework, designed the evaluation, collected and analyzed data, and drafted the manuscript. TT and CF collaborated in study design and interpretation of the results. Authors would like to thank Rimma Pivovarov, Nicole Weiskopf and Janet Woolen for their insightful comments on the manuscript.

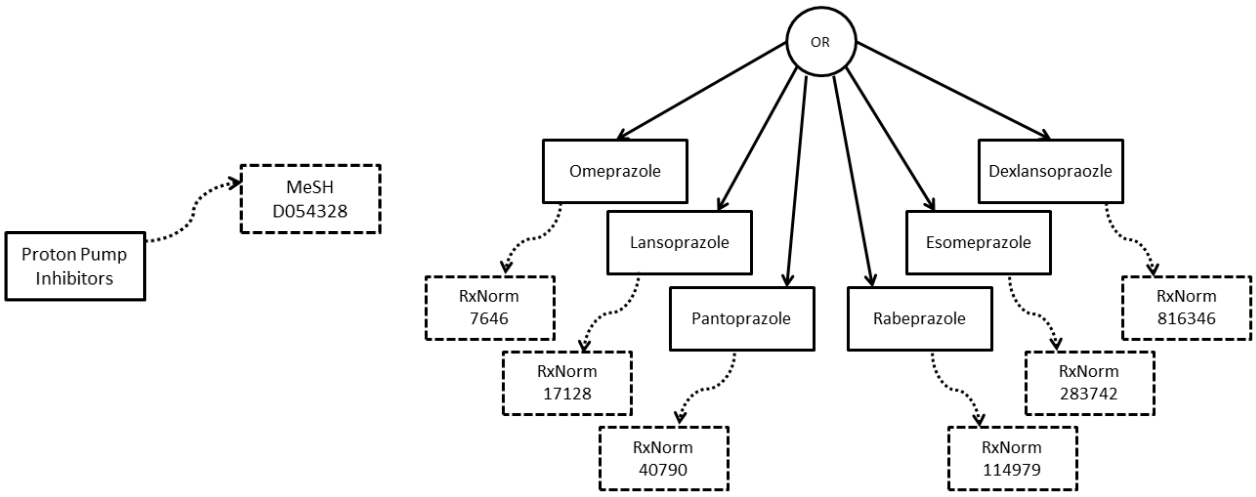


Figure 3 – Classes of medications or diseases can be defined either by creating an external link to the 'class' itself (left), or by creating external links to each of the elements of the class and then combining them using 'OR' logic locally (right). Boxes with solid borders indicate concepts specified in OMAC, while dotted arrows indicate a link to an external knowledge source, and boxes with dashed borders specify the external knowledge source and the unique identifier of the respective concept in that knowledge source. The links will also include the version of the external ontology as well as a URL linking to that external ontology (not shown in this schematic).

References

1. Thomson F, Masters IB, Chang a B. Persistent cough in children and the overuse of medications. *J Paediatr Child Health.* 2002 Dec;38(6):578–81.

2. Kapur PA. Pharmacy Acquisition Costs: Responsible Choices Versus Overutilization. 1994;:617–618.

3. Lindley CM, Tully MP, Paramsothy V, Tallis RC. Inappropriate medication is a major cause of adverse drug reactions in elderly patients. *Age Ageing.* 1992 Jul;21(4):294–300.

4. Naunton M, Peterson GM, Bleasel MD. Overuse of proton pump inhibitors. *J Clin Pharm Ther.* 2000 Oct;25(5):333–40.

5. Conti R, Busch AB, Cutler DM. Overuse of antidepressants in a nationally representative adult patient population in 2005. *Psychiatr Serv.* 2011 Jul;62(7):720–6.

6. Autier P, Creplet J, Vansant G, Brohet C, Paquot N, Muls E, et al. The impact of reimbursement criteria on the appropriateness of "statin" prescribing. *Eur J Cardiovasc Prev Rehabil.* 2003 Dec;10(6):456–62.

7. Mainous AG, Hueston WJ, Love MM, Evans ME, Finger R. An evaluation of statewide strategies to reduce antibiotic overuse. *Fam Med.* 2000 Jan;32(1):22–9.

8. Kester L, Stoller JK. Ordering respiratory care services for hospitalized patients: practices of overuse and underuse. *Cleve Clin J Med.* 1992;59(6):581–5.

9. Tjia J, Velten SJ, Parsons C, Valluri S, Briesacher B a. Studies to reduce unnecessary medication use in frail older adults: a systematic review. *Drugs Aging.* 2013 May;30(5):285–307.

10. Sketris I, Ingram EL, Lummis H. Strategic opportunities for effective optimal prescribing and medication management. *Can J Clin Pharmacol.* 2009 Jan;16(1):e103–25.

11. Korenstein D, Falk R, Howell EA, Bishop T, Keyhani S. Overuse of health care services in the United States: an understudied problem. *Arch Intern Med.* 2012 Jan 23;172(2):171–8.

12. Spinewine A, Schmader KE, Barber N, Hughes C, Lapane KL, Swine C, et al. Appropriate prescribing in elderly people: how well can it be measured and optimised? *Lancet.* 2007 Jul 14;370(9582):173–84.

13. Chan KS, Chang E, Nassery N, Chang H-Y, Segal JB. The State of Overuse Measurement: A Critical Review. *Med Care Res Rev.* 2013 Jun 26;

14. Kaufmann CP, Tremp R, Hersberger KE, Lampert ML. Inappropriate prescribing: a systematic overview of published assessment tools. *Eur J Clin Pharmacol.* 2013 Sep 10;

15. Choudhry MN, Soran H, Ziglam HM. Overuse and inappropriate prescribing of proton pump inhibitors in patients with Clostridium difficile-associated disease. *QJM.* 2008 Jun;101(6):445–8.

16. Fick DM, Cooper JW, Wade WE, Waller JL, Maclean JR, Beers MH. Updating the Beers criteria for potentially inappropriate medication use in older adults: results of a US consensus panel of experts. *Arch Intern Med.* 2003;163(22):2716–24.

17. Oborne CA, Hooper R, Li KC, Swift CG, Jackson SHD. An indicator of appropriate neuroleptic prescribing in nursing homes. *Age Ageing.* 2002 Nov;31(6):435–9.

18. Hanlon JT, Schmader KE, Ruby CM, Weinberger M. Suboptimal prescribing in older inpatients and outpatients. *J Am Geriatr Soc.* 2001 Feb;49(2):200–9.

19. Rancourt C, Moisan J, Baillargeon L, Verreault R, Laurin D, Grégoire J-P. Potentially inappropriate prescriptions for older patients in long-term care. *BMC Geriatr.* 2004 Oct 15;4:9.

20. Rollason V, Vogt N. Reduction of polypharmacy in the elderly: a systematic review of the role of the pharmacist. *Drugs Aging.* 2003 Jan;20(11):817–32.

21. Gallagher P, O'Mahony D. STOPP (Screening Tool of Older Persons' potentially inappropriate Prescriptions): application to acutely ill elderly patients and comparison with Beers' criteria. *Age Ageing.* 2008 Nov;37(6):673–9.

22. Barry PJ, Gallagher P, Ryan C, O'mahony D. START (screening tool to alert doctors to the right treatment)--an evidence-based screening tool to detect prescribing omissions in elderly patients. *Age Ageing.* 2007 Nov;36(6):632–8.

23. National Guideline Clearinghouse [Internet]. Available from: http://www.guideline.gov/

24. Gallagher P, Ryan C, Byrne S. STOPP (screening tool of older person's prescriptions) and START (screening tool to alert doctors to right treatment). Consensus validation. *Int J Clin Pharmacol Ther.* 2008 Feb;46(2):72–83.

25. Woolf SH. Practice guidelines: a new reality in medicine. I. Recent developments. *Arch Intern Med.* 1990 Sep;150(9):1811–8.

26. Kahol K, Vankipuram M, Patel VL, Smith ML. Deviations from protocol in a complex trauma environment: errors or innovations? *J Biomed Inform.* 2011 Jun;44(3):425–31.

27. Shiffman RN, Karras BT, Agrawal a, Chen R, Marenco L, Nath S. GEM: a proposal for a more comprehensive guideline document model using XML. *J Am Med Inform Assoc.* 7(5):488–98.

28. Wang D, Peleg M, Tu SW, Boxwala A a, Ogunyemi O, Zeng Q, et al. Design and implementation of the GLIF3 guideline execution engine. *J Biomed Inform.* 2004 Oct;37(5):305–18.

29. Tu SW, Musen M a. A flexible approach to guideline modeling. *Proc AMIA Symp.* 1999 Jan;:420–4.

30. Sutton DR, Fox J. The syntax and semantics of the PROforma guideline modeling language. *J Am Med Inform Assoc.* 2003;10(5):433–43.

31. Tu S, Campbell J, Glasgow J. The SAGE Guideline Model: achievements and overview. *J Am Med Inform Assoc.* 2007;14(5):589–598.

32. Chen R, Georgii-Hemming P, Ahlfeldt H. Representing a chemotherapy guideline using openEHR and rules. *Stud Health Technol Inform.* 2009 Jan;150:653–7.

33. Johnson PD, Tu SW, Musen MA, Purves I. A virtual medical record for guideline-based decision support. *Proc AMIA Symp.* 2001 Jan;:294–8.

34. OpenCDS: Open Clinical Decisoin Support Tools and Resources. [Internet]. Available from: http://www.opencds.org/

Design Considerations for Post-Acute Care mHealth: Patient Perspectives

Patrick Sanger, BA, Andrea Hartzler, PhD, William B. Lober, MD MS,
Heather L. Evans, MD MS, Wanda Pratt, PhD
University of Washington, Seattle, WA

Abstract

Many current mobile health applications ("apps") and most previous research have been directed at management of chronic illnesses. However, little is known about patient preferences and design considerations for apps intended to help in a post-acute setting. Our team is developing an mHealth platform to engage patients in wound tracking to identify and manage surgical site infections (SSI) after hospital discharge. Post-discharge SSIs are a major source of morbidity and expense, and occur at a critical care transition when patients are physically and emotionally stressed. Through interviews with surgical patients who experienced SSI, we derived design considerations for such a post-acute care app. Key design qualities include: meeting basic accessibility, usability and security needs; encouraging patient-centeredness; facilitating better, more predictable communication; and supporting personalized management by providers. We illustrate our application of these guiding design considerations and propose a new framework for mHealth design based on illness duration and intensity.

Introduction

Although much mHealth research supports managing chronic illness, relatively little is known about how mHealth could apply to acute conditions. New incentives (e.g., Accountable Care Organizations, bundled payments) will lead hospitals and providers to optimize care across the whole care spectrum, including areas such as post-acute care (i.e., happening after acute hospitalization). Post-acute care mHealth could facilitate care coordination by filling in gaps that occur during this significant transition of care. This coordination could prevent costly readmissions as well as improve patients' experiences during this often stressful time.[1]

Improving care coordination among surgical patients after discharge is of critical importance. These patients are at high risk for surgical site infections (SSI), most of which occur after discharge.[2,3] SSIs are the leading cause of readmission among surgical patients, which occurs in up to half of patients who experience SSI.[2] In our previous work we found that patients were ill equipped to recognize and manage wound complications, and faced many barriers to communicating with their providers after developing a concern. Patients found the concept of an mHealth wound tracking and communication tool highly acceptable and believed it could help address many gaps in the current system.[4]

In this paper, we extend our previous work by describing design considerations for post-acute care mHealth apps, derived from interviews with surgical patients who experienced wound complications while at home. Many of these patients had extreme experiences, both in terms of their physical and emotional state (e.g., anxiety, disorientation) and their interactions with the health care system (e.g., late night calls to triage nurses, emergency department visits, readmissions). These experiences allowed us to explore a breadth of the post-acute mHealth design space, although not all issues we identified will likely apply to every post-acute care mHealth app. In our discussion, we describe several core themes that emerged related to communication and management that could be applicable to a wider range of acute care apps. We then illustrate how we applied these design considerations to an mHealth app we are developing to engage surgical patients in post-discharge wound tracking. Finally, we introduce a new framework for mHealth design based on illness duration and intensity.

Background

The design space around mHealth for management of chronic conditions is relatively mature, especially for conditions such as cancer[5,6] or diabetes[7,8]. It is unclear whether design considerations for chronic mHealth apps apply as well in a post-acute setting. Apps for management of chronic conditions are characterized by achieving symptom control and long-term behavioral change. In contrast, the purpose of a post-acute care app might be to help avoid escalation around a single, limited duration episode of treatment while a patient is returning to a usual health state.

Some design considerations likely apply across a wide range of mHealth settings. For example, Klasnja et al[9] found that the ability for cancer patients to capture and access a variety of care-related information while on the go, in a

single application, helped them manage their care and feel more in control of their information and health. Since managing information is a key task for patients in any setting, enabling the organized storage of health information is likely to be a universal mHealth theme. Arsand et al[10] found that diabetes patients preferred to have some reward (e.g. education or feedback) at the time of data entry to provide a built-in motivation for use. This finding relates to a broader theme that, in order to continue using apps, patients must find them to have utility—not just in a theoretical sense, but also in an immediate, concrete sense. Liu et al[11] found that parents wanted to communicate with providers in different ways to suit their needs, both synchronously (e.g. telephone) or asynchronously (e.g. email). Communication with providers can be critical across a range of chronic and acute conditions, and supportive mHealth applications should facilitate communication using means that are both efficient and acceptable to patients. Kientz et al[12] suggest that the act of tracking health measures (e.g. infant development) has the potential to increase anxiety over trends that appear abnormal. Apps that capture patient data will have to carefully consider how to reflect that data back to patients, including whether and how to provide interpretation of that data.

Other design considerations for chronic mHealth apps might be less applicable or introduce new challenges in an acute context, ranging from privacy and self-reflection to automated feedback and engaging social networks. For example, Patel et al[5] identified giving breast cancer patients ownership over data (e.g. controlling what data is shared, capturing custom fields) as important to promote engagement in care and capture of sensitive data. Although patients should always have control over what information is shared with others, acute care providers may be concerned about patients sharing *too much* data that cannot be efficiently reviewed. Mamykina et al[7] identified promoting self-reflection as a primary design goal to aid in self-management for people with diabetes. Self-reflection is a critical element in the management of many chronic illnesses that rely on patients to make lasting behavioral changes. However, the importance of self-reflection is unclear given the short time horizons, cognitive impairments (e.g., due to pain medication), and limited control that patients often have over their care outcomes that might be common in the post-acute setting. In addition to self-reflection, Harris et al[8] identified automated, programmed responses as a key design requirement for diabetes self-management. Automated responses may support self-reflection and have the benefit of giving immediate feedback and gratification to patients without burdening a clinician, however more urgent or complex assessments associated with acute concerns might not be reliably made without human involvement. Finally, much work has been done using mHealth to help patients engage social networks and online communities for support in their care (e.g., Liu et al[11] in the parenting of high-risk infants). However, the utility of online communities is unclear over short durations and highly individualized recovery periods following hospital discharge. In addition, unlike with chronic conditions, acute conditions tend not to have dedicated online communities.

Though the examples above are not exhaustive, and many design considerations for chronic mHealth apps likely apply to acute apps as well, we suggest that the requirements and user experience in acute and chronic settings are sufficiently different to warrant further research. In this paper, we report on our work to explore a design space for apps which improve communication and decision support in the post-acute care setting.

Methods

We conducted semi-structured interviews with patients who experienced surgical wound complications after hospital discharge. The study was approved by the University of Washington Institutional Review Board.

Participants and setting

We interviewed patients who had post-discharge wound complications after undergoing abdominal surgery at one of two Seattle hospitals: an academic medical center or a county hospital/ regional trauma center. We identified English-speaking, adult patients using two different approaches: through clinic nurses at follow-up visits or through flyers placed in surgery clinics.

Data collection

We conducted one-on-one interviews lasting 60-90 minutes in a private setting near the clinics. We began by using the critical incident technique to guide participants in recounting their complication experience.[13] Then, grounded in their experience, we used scenarios to provide context to allow participants to walk through paper wireframe mockups (Figure 1) of an mHealth wound tracking application (e.g. “Imagine you are very concerned about your wound.

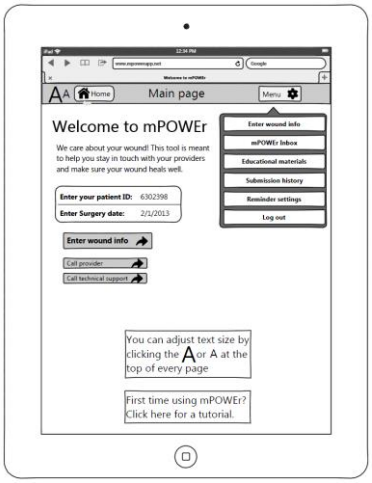


Figure 1: Paper mockups used during participant interviews to stimulate discussion.

You just clicked 'submit' to send your symptom data. What should happen now?"). We showed mockups of multiple versions of potential features such as symptom tracking, wound photography, secure communication, and informational content. Prior to showing mockups of each feature, the interviewer paused to ask the participant to describe how a particular feature might work; only then did the interviewer use the paper mockups to stimulate further discussion. Interviews were audio-recorded and transcribed. We collected data until thematic saturation was achieved.[14] We used written surveys to capture demographics and technology experience.

Data analysis

We collectively developed a codebook with two team members coding all the interviews using Atlas.ti (Atlas.ti v7, ATLAS.ti GmbH) while other team members spot-coded interviews to inform the codebook and check reliability. The whole team met periodically to resolve coding discrepancies. Cohen's Kappa between the two primary coders during early and late coding was 0.51 and 0.71, respectively, reflecting moderate to substantial inter-coder reliability.

Results

We interviewed 13 patients ranging from age 21 to 71 (mean 45), of whom 9 were female and 9 were white. Five were college graduates, 6 had some college, 1 graduated from high school and 1 had less than high school education. They self-rated their experience with computers as "some" (n=3), "intermediate" (n=4), "very" (n=4), or "expert" (n=2). Twelve used the internet at least occasionally and 8 owned a smartphone. Patients underwent major abdominal surgery, generally colorectal or ventral hernia repair, and struggled with complications for weeks or months after discharge. Five had one or more emergency department visits or hospital readmissions related to SSI.

From patient interviews, we identified 11 themes that we organized into 4 categories that describe qualities of a post-acute care mHealth application (Figure 2):

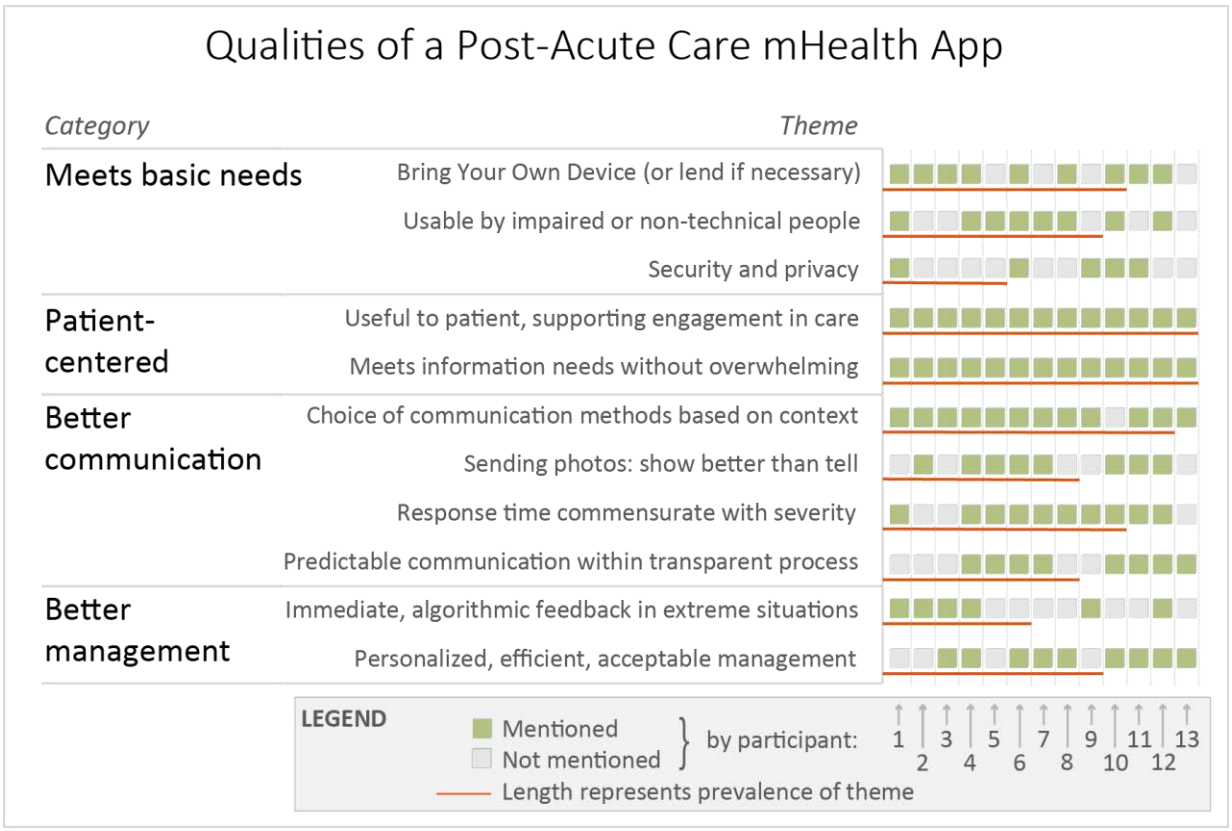


Figure 2. Each green square represents a participant who mentioned the theme during interviews. Length of red lines represents theme prevalence. Themes are organized into 4 major categories, visible on the left side of the figure.

Meets basic needs

One category of design qualities for a supportive mHealth tool expressed by participants revolved around meeting basic needs through accessibility of your own device, usability by impaired or non-technical people, and security to preserve privacy. P#'s following quotes are attributions to that particular participant.

Bring Your Own Device (or lend if necessary). Participants were concerned about an app being inaccessible on their preferred device or had concern for others who lacked access to a smartphone or computer. Several participants suggested that hospitals could loan patients devices.

> *What about people that don't have smart phones? Will you have this on a regular online website with it too? Because my phone is like ten years old.* ***P12***

> *I think it's a good idea if you have someone who's technology proficient in something like that. Some of the patients may not even have computer or computer access. But I think for me personally, I mean, because I'm used to computers, I think it's a great idea.* ***P4***

Usability by impaired or non-technical people. Participants had concern for technologically inexperienced users, and did not want to be overwhelmed with information or too many pages/functions. Participants mentioned the challenge of using an app while on pain medication, and wanted simple wording, obvious alerts and clear navigation.

> *But if you don't have [a smartphone] you might not know how to work it, so then you're going to have to get into who's smarter, the phone or you? You know, somebody's going to have to show you how to even operate the thing.* ***P10***

> *But it would make a complex website or doing something complex, it would require you to remember several steps. I think [navigating a complex website under influence of drugs] would make it very difficult for a lot of people.* ***P12***

Security and Privacy. While this was among the least prevalent themes, participants were most concerned about collection and transmission of particularly sensitive information such as photos of the groin area. Participants expressed concern that transmissions should be secure and go to the right recipients.

> *Some people might not want to [send pictures of private areas]. Your older generation.* ***P6***

> *As long as [the submissions] made it from point A to point B, it would be good. They didn't get lost in sending... to a bakery or something.* ***P11***

Patient-centered

A second major category of design qualities pointed to the importance of patient-centeredness: being genuinely useful to the patient to support engagement in their own care and meeting their individualized information needs without overwhelming them.

Useful to patient, supporting engagement in care. All participants voiced that the app should be genuinely useful to the patient—having an obvious benefit and not feeling like a burden on them. They felt that the app should allow patients to be engaged in and make decisions about their care, especially about how often and by what means they discuss their concerns with providers. Several expressed that the app should have a "personal feel" that gives the feeling that "we want to take care of you" (P1). Participants also felt that the app should connect them with a provider familiar with their history and with whom they already have a relationship/rapport.

> *The biggest thing is for me to feel like this is useful, because it's being sent to my doctor. This is a way of communication to my doctor, not just a survey I'm taking, you know what I mean?* ***P5***

> *[The ability to view past photos/history would be useful] because then you could see – oh, this is what this looked like 3 days ago and this is what it looks like now. This looks really different.* ***P7***

Meets information needs without overwhelming. Participants generally did not feel that their information needs were met well during their hospital discharge experience. Every participant saw an opportunity for the app to make up for this deficit by providing a personalized, succinct recap of their discharge instructions. They emphasized that they did not want to be overwhelmed, asking for just the highlights with links to more resources if needed. Most participants wanted to have information on procedures they themselves would be expected to perform after

discharge (e.g. how to clean and pack wound) and how to identify problems (e.g. infection). Participants preferred concrete examples through a variety of media over reams of pages (e.g. photos of infected vs normal wound, step-by-step instructions/video of wound care procedure). Several participants were also interested in receiving information about how to optimize their healing (e.g. dietary advice). Participants wanted the app itself to be well-documented with help/tutorials.

> *Like if you forget how to clean and pack your wound or whatever, or if your wound looks like this, then [it's infected] - or if your wound looks like this, then [it's not infected]. That might be helpful… Mainly just in terms of if this happens, don't freak out. If this happens, do freak out.* ***P7***

> *So if you're going to do a presentation for somebody coming out of the hospital, you should only have the highlights… [have a] mouseover if they [want] a big explanation.* ***P1***

Better communication

A third major category of design qualities pointed to the potential for enhanced communication, whether through more choice of communication methods appropriate to context (e.g., secure text for non-urgent matters), the ability to send photographs, rapid provider response when necessary, and patient control over and transparency about timing and method of provider contact.

Choice of communication methods based on context. Participants wanted to be able to choose the means of communication with their provider. Context was important—if they were very concerned about an issue, participants preferred a telephone call. When participants were not very concerned (e.g. routine check-ins, non-time sensitive care questions), many preferred text or email as it was more convenient for them and less interruptive to their providers than phone calls. Two participants suggested that real-time video conferencing should be incorporated into the app.

> *[The app should have] an option of how would you best like to be communicated with… Would you like it email, text message, phone call and they can select that, and it can go right in with the message. ... Because [grandmother] would pick a phone call, [mother] and I would pick a text message.* ***P6***

> *It depends on the situation, but I don't know, for me personally, I like to do stuff through email ... Unless it's like super urgent, so obviously phone call is the best way to get [urgent] communication. But I think for this type of stuff, I wouldn't mind email, as long as I knew that the doctor's looking at it throughout the day.* ***P4***

Sending photos: show is better than tell. Participants were very interested in sending photos to providers. They wanted their provider to really see that status of their wound rather than try to explain solely over the phone. They thought that this additional visual information would facilitate triage and management, be less subjective than patient-assessed symptoms (e.g. amount of redness), and help show trends across time through serial photography. Participants recognized that photos were necessary but not sufficient–some patient-assessed symptoms would be valuable to report (e.g. heat, pain).

> *I have a smart phone so I used that to take the picture. I thought that was very good to be able to send them an actual picture of what was happening so that way, you know, a little more hands on than "okay - this is… " - trying to describe it over the phone... The nurse commented about how good that was too to have a picture to look at.* ***P6***

> *So if you had it where you could take a picture of it… [the provider] might have said "oh boy, you need to go into the hospital" [or] they could say "hey – no, it's doing what it's supposed to do, just let it be."* ***P10***

Response time commensurate with severity. Participants wanted faster response times based on their level of concern and/or the apparent severity of their wound problem—in other words, the app should facilitate triage to enable provider feedback faster based on urgency. Many participants made comparisons to the main alternative to using the app—a phone call—saying that response times should be comparable (e.g. call back within 30 minutes). Participants voiced worries that waiting for even an hour might be too long for an acute concern and that if responses are too delayed, their condition could deteriorate. Of those who specified how quickly they would expect a response, 2 said under 30 minutes, 1 said within an hour, 2 said within 4 hours, and 4 said within 24 hours. Participants were willing to wait longer for a response if they had confidence in the system–that their responses were being monitored regularly and not falling into a "black hole" (P4).

I think that if it would have been really hurting, I would want a quicker response time for it. So I think based on the level of pain that somebody was having as to what the response - or felt they were having, the response time back would be quicker. ***P6***

When you pick up the phone you're getting a response. If you're using a tool and you're not getting anything back, then there's no reason to use it because the whole reason is to get communications. ***P1***

Predictable communication within a transparent process. Participants wanted a definite timeframe for a provider response, i.e. a shared expectation between patients and providers. Generally they expected the provider to set this parameter but several wanted to select a time and/or be able to "escalate" to request a faster response. Participants wanted the process to be transparent – to know when their data was received, viewed, and acted upon. They also wanted to be able to set the contact method so they would know what to expect (e.g. wouldn't have to wait around at their computer in case of email response).

Some type of timeframe. So it's not just kind of like sitting out there and you just submit it to a black hole, you know, when someone's going to get back to you. ***P4***

[After clicking submit, the app should say] 'please watch your email during the next three hours or something for a response'. Or whatever you guys decide the response time should be. And/or choose a phone call back. In other words, to know on here before I log off what I can expect next... ***P12***

Better management

A final major category of design qualities pointed to the potential for better, faster, more personalized and more acceptable (to the patient) management. Participants saw many potential benefits including earlier identification and treatment of problems, and the possibility of more efficient care through reduction in unnecessary visits.

Immediate, algorithmic feedback in extreme situations. Participants found the idea of algorithmic (i.e. immediate, app-generated) feedback most acceptable at the extremes—i.e. their situation appears very good or very bad. In less clear-cut cases, most favored the judgment of their health care provider. Participants thought algorithmic feedback was good if it was based on existing practices (e.g. algorithms used by triage nurses). Participants noted that algorithms could benefit both the patient (e.g. advise to go to emergency department immediately if reporting chest pain) and the provider (e.g. flag most concerning patients to review quickly). In general, most participants did not fully trust the computer to make unsupervised management decisions, noting that the quality of the patient input is critical; misjudgments about symptoms could lead patients to unnecessary emergency room visits.

I think yeah, that's all right for the extremes. But I still would feel more comfortable with the doctor responding. ***P4***

It's the same judgment that you would get if you called a nurse, well, it's probably the same thing you were told outright – if you see this, call the nurse or come into the emergency room. So if the app is just reinforcing that, it seems perfectly natural. ***P3***

Personalized, efficient, acceptable (to the patient) management. Almost universally, patients saw the potential for an app to facilitate better triage. For example, the app could help the patient answer the anxiety-ridden question, "What do I do? Come in or stay home?" (P8). Through better triaging, patients expected a variety of potential benefits. If they were healing normally, for example, the app could save time and unnecessary clinic/emergency room visits, which is especially important for distant patients, as well as alleviate stress and provide reassurance. If their wound was not healing normally, the app could facilitate earlier problem identification and quicker/easier re-admission than the current management process patients experienced. Patients liked the idea of giving their provider more data (especially photos) to track their progress, and would be more willing to accept providers' management decisions based on that more complete, standardized, and personalized data. For example, they would be more willing to go to ER if advised to do so.

That's pretty much what the triage nurse tells you anyway. You have to come in [to the emergency room]. *But if you have a picture of it, and it's nothing, then that would make it so that you wouldn't have to go in necessarily… It would be more advantageous and you wouldn't have to sit there for five hours (laughs) in the ER.* ***P10***

But it would have been really helpful, especially the first time that it started getting infected, I could have sent them a picture or whatever and then if a day later - because it did, it got a lot worse. It was itching, it

was bleeding and stuff - then I could have sent another picture and said it's a lot worse and they could have seen right then you need to come in now. Instead of waiting until it got really bad. ***P7***

Discussion

Our findings illustrate the large potential benefits that patients see in post-acute care mHealth apps. Indeed, such apps are probably inevitable, but the key question is: will they be embraced by patients? Due to the hectic and stressful time during care transitions after acute illness, it is critical that apps be immediately usable and obviously useful to patients or they will not be used. Both patients and clinicians will lose out if patients reject this powerful method to facilitate data gathering and communication in favor of the highly usable yet limiting alternative—the telephone. However, it is challenging to design for short-term post-acute episodes for a number of reasons. First, there are a large number of possible use cases, and acute problems do not necessarily follow a predictable disease course. Second, related to user-centered design, it is challenging to engage patients in the moment, while they are actually sick, and due to the short-term nature of acute conditions, patients may lack the expertise about managing their condition that patients with chronic illness may have. Finally, it is unclear whether prior work on mHealth for chronic illness management is applicable in a post-acute context.

Although many themes voiced by our participants were common across mHealth, several appeared novel, reflecting the specific needs of patients in a particular post-acute setting (Table 1).

Support for known themes across mHealth	New themes specific to post-acute setting
• Bring your own device/loans • Security and privacy • Useful to the patient, supporting engagement • Algorithmic feedback when appropriate • Choice of communication methods based on context • Photos (or other sensor-based data): show is better than tell • Meeting information needs without overwhelming	• Response time commensurate with severity • Predictable patient-provider communication within a transparent process • Personalized, more efficient, more acceptable treatment plan based on patient-reported data • Usability while in a cognitively or physically impaired state

Table 1: Common themes across mHealth vs new themes for post-acute care mHealth. Key themes identified from patient interviews have been categorized based on whether they are already known in the mHealth literature or appear novel to a post-acute setting.

The most important insights about patient expectations for acute care mHealth apps relate to patient-provider communication and resulting management of care concerns. This prioritization likely reflects the challenges and frustrations participants faced when managing their post-discharge complications. For example, after developing a wound concern, participants described a lack of control over their situation. Related to communication, they could not easily or quickly reach a familiar provider. Related to management, they often felt unnecessarily directed to the emergency department as the default option. Through an mHealth application, patients wanted to be empowered to choose how and when they would be contacted, and wanted to be satisfied that the provider managing their care made a personalized recommendation based on all available information (e.g. through review of serial symptom logs and wound photos). Because issues of patient-provider communication and management are essential to addressing many acute concern, future work could explore the generalizability of these themes across a variety of acute and post-acute conditions.

Although our findings are limited to *patient* views, one of the key differentiating elements of acute mHealth is the relative importance of other stakeholders, most notably *providers*. Acute mHealth apps must be designed to satisfy two very different user groups who almost certainly have competing priorities. In our previous work, a needs assessment of providers for a post-discharge wound tracking app[15], providers expressed concern over additional time requirements, workflow disruption, issues surrounding receipt of photos (e.g. liability, poor quality), and EMR integration. Patient and provider expectations differed on such things as frequency of and trigger for wound tracking: patients expected to track their wound routinely even in the absence of an obvious problem, while providers envisioned less frequent use, generally only if a problem was suspected. Similarly, many of the key patient expectations identified in this paper are subject to provider buy-in and will have to be negotiated between patients

and providers. Acute care mHealth apps might ultimately be disruptive, catalyzing a shift from provider-driven to patient-centered care processes.

Figure 3 depicts one such mHealth app we are developing to facilitate wound-tracking and patient-provider communication by surgical patients after hospital discharge. We demonstrate how the 11 guiding design considerations have informed the most recent prototype. This prototype is an interactive mockup currently undergoing heuristic evaluation and user testing; concurrently, we are using agile techniques for development of the patient-facing app and provider-facing dashboard.

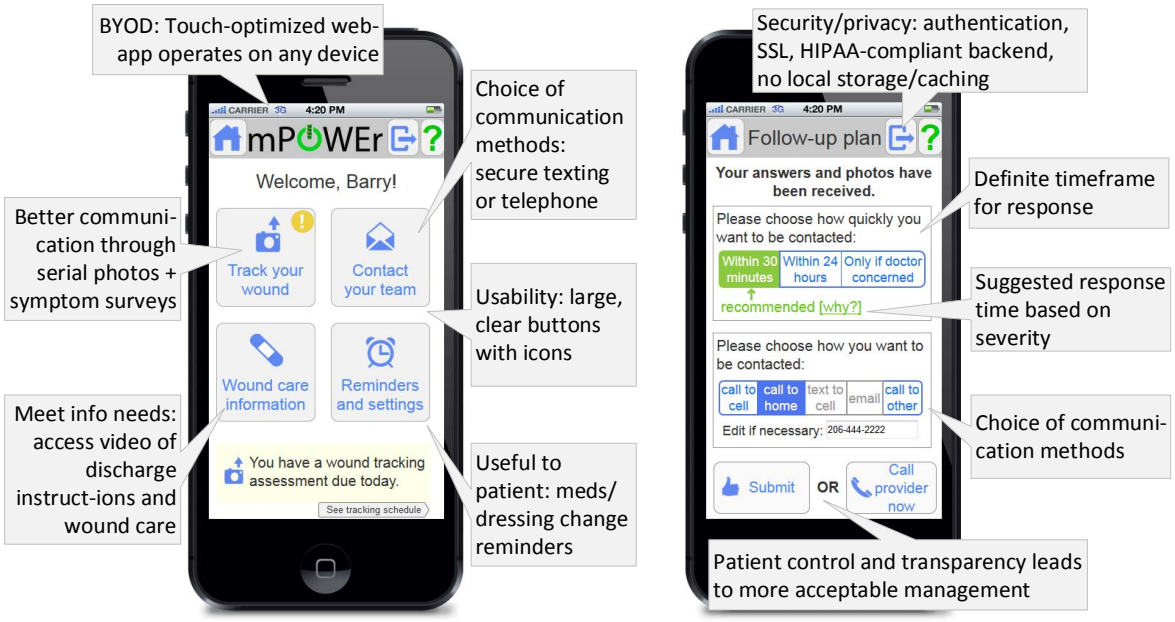


Figure 3: Application of design considerations to a wound tracking app. Each callout represents one of the 11 guiding themes that emerged from interviews with patients who experienced post-discharge complications.

Though the chronic vs. acute distinction is widely used, it may be more useful in generalizing mHealth design considerations to distinguish health conditions along two axes: short-term vs long-term and high-intensity vs low-intensity usage (Figure 4). Typical chronic illnesses tend to be long-term and low intensity while acute concerns tend to be short-term and high intensity, but other conditions (occupying the adjacent quadrants) can have elements of both chronic and acute conditions. Some conditions may even shift around, e.g. short-term, high-intensity surgical wound monitoring may shift to long-term, low-intensity chronic wound monitoring, or stable diabetes may become uncontrolled, shifting toward higher intensity. In designing apps for individual conditions or groups of conditions, it makes sense to consider both intensity and duration, and how these change over time. For example, short-term use may require simplicity and easy learnability, whereas long-term use may allow the possibility of more complexity and customization; high intensity use will require consideration of how to facilitate timely patient-provider communication, whereas low intensity use may not require provider involvement, using algorithmic feedback to patients instead.

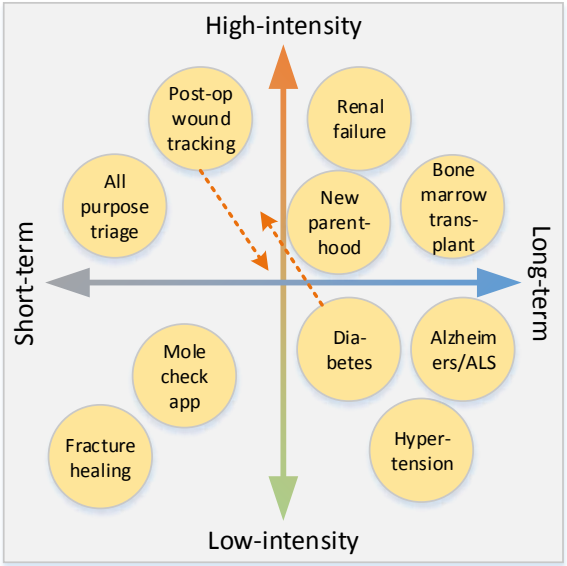


Figure 4: Example mHealth apps by duration and intensity. Dotted orange lines indicate possible shifts based on disease course or progression.

Our research has a number of strengths, including identification of new themes relevant to post-acute mHealth and affirmation of other themes that have been previously reported in the context of chronic illness mHealth. Our findings provide a sound basis for a patient-centered approach to software development, uncommon in the health domain, yet key to developing applications that patients will actually use.

Despite these strengths, this study has several limitations. First, we only interviewed surgical patients. Patients affected by other conditions may have different needs and preferences. We believe ours is a good initial test population due to the challenging and often eventful post-discharge experience following major surgery. Second, we only interviewed patients who experienced post-discharge complications. We believe that patients who experienced problems are the most likely users of the app and have the most insight into current system failings. Lastly, we interviewed a relatively small number of patients from two very different, but related, hospital settings. As is customary in qualitative research, the sample size was based on reaching saturation. In the future we will address some of these issues through user-centered development of our wound-tracking mHealth platform and examine its impact on patient satisfaction, quality of life, clinical outcomes, and healthcare utilization.

Conclusion

Through interviews with patients who experienced post-discharge complications, we explored the design space of a post-acute care mHealth app. Patients described lack of information at discharge, lack of control over communication and mistrust about management decisions made by providers about their care. In response, they envisioned design qualities of an mHealth app that could empower patients through meeting information needs and facilitating predictable communication, and empower their providers with information to make the best decisions about their care. We present a set of design considerations for post-acute care apps and propose a new model for differentiating mHealth apps by the intensity and duration of illness. These contributions incorporate key patient preferences to expand the mHealth landscape with apps that patients will embrace.

Acknowledgments

We thank our participants, the iMed research group, Sarah Han, Cheryl Armstrong, and Mary Ko for their significant contributions. This work was supported by the University of Washington Department of Surgery, the University of Washington AHRQ Comparative Effectiveness Research Award (K12 HS019482) and University of Washington NCATS Translational Science Training Grant (TL1 TR000422).

References

1. Rhodes KV. Completing the play or dropping the ball? the case for comprehensive patient-centered discharge planning. 2013:24–25.
2. Gibson A, Tevis S, Kennedy G. Readmission after delayed diagnosis of surgical site infection: a focus on prevention using the American College of Surgeons National Surgical Quality Improvement Program. Am J Surg. 2013.
3. Kazaure HS, Roman SA, Sosa JA. Association of postdischarge complications with reoperation and mortality in general surgery. Arch Surg. 2012;147(11):1000–7.
4. Sanger P, Hartzler A, Han SM, Lober WB, Evans HL. Patient Perspectives on Post-Discharge Surgical Site Infections: Towards a Patient-Centered mHealth Solution. Under review. 2014.
5. Patel RA, Klasnja P, Hartzler A, Unruh KT, Pratt W. Probing the benefits of real-time tracking during cancer care. AMIA Annu Symp Proc. 2012;2012:1340–9.
6. Hayes G, Abowd G, Davis J. Opportunities for pervasive computing in chronic cancer care. In: Pervasive ’08 Proceedings of the 6th International Conference on Pervasive Computing.; 2008:262–279.
7. Mamykina L, Mynatt E, Davidson P, Greenblatt D. MAHI: investigation of social scaffolding for reflective thinking in diabetes management. In: Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI ’08. New York, New York, USA: ACM Press; 2008:477.
8. Harris LT, Tufano J, Le T, et al. Designing mobile support for glycemic control in patients with diabetes. J Biomed Inform. 2010;43(5 Suppl):S37–40.

9. Klasnja P, Hartzler A, Powell C, Pratt W. Supporting cancer patients' unanchored health information management with mobile technology. AMIA Annu Symp Proc. 2011.; 2011:732–41.
10. Arsand E, Tufano JT, Ralston JD, Hjortdahl P. Designing mobile dietary management support technologies for people with diabetes. J Telemed Telecare. 2008;14(7):329–32.
11. Liu LS, Hirano SH, Tentori M, et al. Improving communication and social support for caregivers of high-risk infants through mobile technologies. Proc ACM 2011 Conf Comput Support Coop Work - CSCW '11. 2011:475.
12. Kientz JA, Arriaga RI, Abowd GD. Baby steps: evaluation of a system to support record-keeping for parents of young children. In: Proceedings of the 27th international conference on Human factors in computing systems - CHI 09. New York, New York, USA: ACM Press; 2009:1713.
13. Chell E. Critical Incident Technique. In: Symon G, Cassell C, eds. Qualitative Methods and Analysis in Organizational Research: A Practical Guide. London: Sage Publications; 1998:51–72.
14. Strauss A, Corbin J. Basics of qualitative research: grounded theory— procedures and techniques. Newbury Park: Sage Publications; 1990.
15. Sanger P, Hartzler A, Lober WB, Evans HL. Provider needs assessment for mPOWEr: a Mobile tool for Post-Operative Wound Evaluation. In: Proc. AMIA.; 2013.

Motivating the Additional Use of External Validity: Examining Transportability in a Model of Glioblastoma Multiforme

Kyle W. Singleton[1,2], William Speier[1,2], Alex AT Bui PhD[1,2], William Hsu PhD[1,2]
[1]Department of Bioengineering, University of California, Los Angeles
[2]Medical Imaging Informatics, Department of Radiological Sciences, University of California, Los Angeles

Abstract

Despite the growing ubiquity of data in the medical domain, it remains difficult to apply results from experimental and observational studies to additional populations suffering from the same disease. Many methods are employed for testing internal validity; yet limited effort is made in testing generalizability, or external validity. The development of disease models often suffers from this lack of validity testing and trained models frequently have worse performance on different populations, rendering them ineffective. In this work, we discuss the use of transportability theory, a causal graphical model examination, as a mechanism for determining what elements of a data resource can be shared or moved between a source and target population. A simplified Bayesian model of *glioblastoma multiforme* serves as the example for discussion and preliminary analysis. Examination over data collection hospitals from the TCGA dataset demonstrated improvement of prediction in a transported model over a baseline model.

Introduction

Substantial amounts of time, money, and effort are exerted to run scientifically sound experiments in the form of randomized controlled trials (RCTs) to determine the efficacy of medical therapies. In addition, systematic reviews and meta-analysis of the findings in RCTs and other research has been made paramount to demonstrating the validity of findings across the large body of medical research. The major focus to date has been in support of calculating the statistics of these experimental endeavors. Utilizing the combined findings from experimental studies and collected observational data, researchers attempt to develop prognostic disease models to aid in clinical decision-making. However, these models often show decreased performance when used to predict results for new data that were not a part of the original modeling dataset.

RCTs, systematic reviews, and meta-analyses investigate the internal validity of data obtained from one or more trials. Despite the rigors employed to ensure statistical accuracy and understanding, the derived knowledge still only represents a level of certainty in regards to the population examined during experimentation. Previous work has demonstrated that the internal validation of results is not suggestive of applicability to future patients[1]. The ability to apply obtained knowledge to new cases is still somewhat nebulous and rarely applies despite expectations[2]. This issue is related to the complexity of disease and treatment, the issues involved in working with populations without incorporating bias, and the difficulty in completely observing any population. The ability to apply data between populations by determining the external validity of findings is a growing area of study. The primary function of validated data is the ability to apply, or *transport*, experimental or observational results to future domains. Transportability theory is a recently developed method suggested for evaluating when the findings for a population meet the proper constraints to be considered externally valid and therefore are applicable to another population[3].

In this work we: 1) develop a limited Bayesian belief network (BBN) disease model for prediction of glioblastoma multiforme (GBM) survival; 2) examine the ability of transportability theory to provide information concerning the external validity of data collected for the model; and 3) test the transport of data between different contributing hospitals in our dataset. Publicly available data from The Cancer Genome Atlas (TCGA)[4] initiative of the National Cancer Institute (NCI) were obtained to form our training and test set populations. This work is meant to demonstrate the use of transportability and some of the complexities involved in moving data (and the resultant models) between populations for predictive purposes. The model used in this paper is simplified to provide an opportunity to examine the characteristics of a disease model and transportability without the complications a large set of predictive variables may add.

Background

Internal and external validity

Prognostic modeling research is largely driven by evidence-based medicine (EBM) tasks: RCTs, subsequent systematic reviews of RCTs, and meta-analysis derived from completed systematic reviews play a critical role in establishing accepted

clinical practice. Physicians update their understanding of disease, as well as new treatments or changes to existing treatment options, by reviewing these studies. These efforts establish important variables and design parameters for collecting data from a controlled population of patients and healthy controls. Despite attempts to use a standard design, it is often difficult to compare across RCTs, even when both studies examine the same drug's effect on a given condition; (confounding) differences can exist in the sampling size, collection constraints, patients lost to follow-up, etc. Moreover, randomized trials are designed to maximize internal validity (i.e. consistency within a cohort by controlling for potentially confounding variables). These studies validate the efficacy of an intervention under ideal conditions but do not necessarily address its clinical effectiveness across a real-world population (i.e., the external validity/generalizability of the intervention to routine practice)[5]. Though more "practical" or "pragmatic" clinical trials are now promoted to relax subject eligibility requirements (thereby broadening the test population), a given investigation may still encompass assumptions about the underlying study group and environment that are difficult to overcome.

The concept of internal validity allows claims about the treatment methods tested. With robust clinical trial design one can determine if true statistical differences exist between intervention and control groups. However, these statistical claims are only valid for the population observed in the RCT. Systematic reviews gather related RCTs, standardizing the differences between trials and providing documentation of the trends of findings for a particular disease or target therapy. Efforts such as the Cochrane Collaboration recognized the need for maintaining unbiased systematic review and meta-analysis for all of medical research. Contributions by members of the community to collaborations ensure that more medical treatments are reviewed than in the past.

Meta-analyses generate a further statistical evaluation of the internal validity of a set of RCTs linked through the systematic review process. In essence, meta-analysis aggregates the statistical findings of a group of RCTs into the semblance of a single, larger study. This analysis is not a combination of all the raw data from the individual populations of each trial, but an examination of the outcome statistics calculated in the results of each trial by performing a weighted average. This revised statistical evaluation of the individual findings is used to make claims about the consensus of medical research on the given topic examined. The conflicting information from RCTs can lead to inconclusive findings concerning a treatment with the suggestion of further research. Significant findings from a meta-analysis are suggestive of strong associations of the experimental findings; however, these conclusions are still decidedly internal as the analysis steps examine the consensus of medical research and do not directly evaluate the external applicability by applying findings from the RCTs to different populations. Thus, it remains difficult to apply the knowledge to future cases where not all circumstances of the RCT-defined environment will hold.

For this reason, a new set of research efforts are examining the concept of *external validity*, and the direct application of past studies' results to future analysis and prediction. Unlike the significant body of work for evaluating internal validity, few techniques have been developed to test external validity; and internal validity methods are not directly applicable to understanding how findings generalize. External validity is a growing concern in epidemiological research[1,6,7], and is encouraging a move towards testing all data findings with methods relevant to both internal and external validity so that evidence and study conclusions are fully contextualized and applied appropriately[8].

Causal models and transportability

Graphical notation is an increasingly prevalent technique for modeling probabilistic and causal relationships across observation and outcome variables of a disease in order to represent disease's etiology, presentation, interventions, and ultimate course. Graphical notation provides a visual interpretation using well-defined sets of vertices and edges: vertices are representative of the variables chosen for a domain of interest, and edges describe relationships that exist between a pair of variables. The relationships are descriptive of the inferences derived from experiments, the belief the model developer has about the variables from past experience or observations, and other sources of knowledge such as domain experts. For example, disease model graphs can make use of RCT and meta-analytical findings, as they provide a set of presumed belief in the relationships between disease variables.

When graphs contain only directional edges and do not provide any cyclical pathway between vertices they are termed directed acyclic graphs (DAGs). The special constraints of a DAG, in addition to only representing causal relationships between variables, form the basis for a *causal model* or *causal network*. Representation using DAGs has proven effective for working within a Bayesian framework as graph notation is able to provide an efficient means for describing the environment of variables and the associated probabilities of their states in the environment. Figure 1 contains an example of a graphical causal model for treatment effect on lung tumor progression.

Pearl and Barenboim introduced transportability theory as a basis for using relations expressed through a causal model to describe which variables' probability distributions "transport", or move, between populations[9]. For example, a physician in a

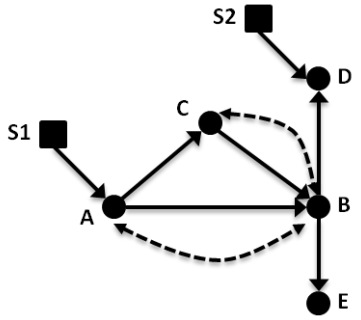


Figure 1. Example causal diagram for lung cancer treatment. Variables: (A) treatment; (B) tumor progression; (C) tumor biopsy gene expression; (D) clinical history; and (E) CT imaging findings. In this causal diagram, solid circles represent standard variables (such as in a Bayesian belief network), and solid arrows between these nodes represent causal relationships. Dashed arrows/arcs indicate confounding influences between two variables that may exist when considering other populations. Selection nodes, shown as squares, provide a means to sub-select or filter a given variable so that the evidence is comparable between two groups.

rural setting might wish to apply the results of an RCT conducted at a large research hospital to decision making for patients locally under his/her care. The RCT findings can be understood in the context of a causal graph, per Figure 1, as a treatment (A) with effect on a patient outcome (B), with additional measured factors such as clinical history, imaging, or genetics (C, D, E). Transportability allows a researcher to identify potential confounding evidence between variables (represented by dotted lines) and population differences (indicated by square nodes, S1 and S2) that are known or believed to exist between two cohorts. The influence of these constraints can be used to determine what data from the RCT can be applied to the rural patients in a principled way. For instance, the physician may not have enough genetic information for his population to build a model; applying transportability can help ascertain whether the genetic information collected in the RCT can be reused (i.e., transported) to the local group (and if not, under what different graphical circumstances such data transport would be valid). Similarly, differences between the hospital and local populations (e.g., demographics) can be accommodated via transportability. In general, if all existing differences can be accounted for, then the external validity of the findings makes the model variables transportable to the new population.

A distinction should be made concerning the differences between causal and probabilistic models (e.g., BBNs). Application of transportability theory is performed on a causal model where an arrow from node X to node Y denotes that X is used in the function that determines Y. Connections are representative of the process "X causes Y" seen in nature, conveying an inherent ordering of events and representing direct functional relationships. Within the probabilistic (Bayesian) context, edges between nodes are often interpreted in a similar fashion as causal connections, but relationships between variables are encoded only by conditional probability tables and statistical relationships. However, arrows in a causal model are meant not only to represent probabilistic dependence but also direct causation. Therefore a causal graphical model is a robust description of the assumptions made by the modeler.

To properly describe the full set of causal connections in the graph, additional information not commonly captured in a Bayesian belief network must be explicitly represented. First, unmeasured confounding information expected to exist between any two nodes needs to be marked accordingly. These confounders are represented by bi-directional dashed edges and cover the counterfactual circumstances of variables that may be impossible to observe or measure. An example of this situation could be the potential interaction of a non-prescription pain-killer and treatments prescribed by the physician (Figure 1, the dashed arc between A and B). Patients may not report their non-prescription drug use or there may be unmeasurable interactions even if the physician knows both drugs are being taken. Second, when population differences are suspected or known to exist for a particular variable, a selection node is added that embeds a method to control for this variation. A selection node serves this purpose by explicitly identifying population differences in the mechanism (e.g., disparities in demographics, socioeconomic status) that are responsible for assigning a value to that variable. By way of illustration, if age differences were significant between two populations, a selection node could be used to indicate the need to select patients who are age-matched. We discuss these points further in our methods and describe them in the context of the simplistic Bayesian model of GBM shown in Figure 2.

Motivating disease and data resources

Nearly half of the 45,000 newly diagnosed cases of adult brain tumor seen annually in the United States are cases of glioblastoma multiforme, an aggressive malignant primary brain tumor. When receiving targeted care at large research hospitals, GBM patients have an average survival time of 12-24 months. In addition, NCI SEER (Surveillance, Epidemiology, and End

Results) data over the last 20 years show little effective change in the survival of GBM patients[10]. A combination of surgical resection, radiation treatment, and chemotherapy are the current standard of care in modern brain cancer treatment. The attempt to improve patient survival time increases the need to study new chemotherapeutic agents and other potential interventions. A growing body of genetic research on the many types of cancer demonstrates the intricate variations that exist between cancer cells, both across tissue types and within individual cancer groups[11,12]. To aid future decision making tasks, statistical examination of cancer findings strive to provide predictive models of treatment and outcomes. However, statistical analysis of cancer studies have been unable to reach a consensus on the most effective predictive variables for GBM patients[13–16]. As our present understanding of cancer and effective treatments are limited, the field continues to work towards integrative models of disease to better employ the influx of experimental data towards improving clinical decision making tasks.

A number of multi-institutional efforts now exist to establish observational databases, supplementing experimental datasets. Two efforts focused on building databases for GBM research are The Cancer Genome Atlas (TCGA) and the Repository for Molecular Brain Neoplasia Data (REMBRANDT). TCGA is a public database of clinical and genetic information for 20 different types of cancer. Containing primarily clinical and genomic (copy number, DNA methylation, gene expression, single nucleotide polymorphisms) data, the TCGA dataset has ongoing efforts to also make radiological and pathological images available. The REMBRANDT database is focused specifically on data obtained for all types of brain glioma (astrocytoma, GBM, mixed, oligodendroglioma) with a limited number of unmatched non-tumor controls.

Methods

Many potential modeling variables from the clinical, treatment, imaging, and genetic domains have been explored in GBM research. A brief list of example variables from existing public GBM datasets are shown in Table 1. A subset of the available data related to a chosen clinical question is selected to facilitate discussion of transportability in the network. In this work, the number of variables considered is substantially limited in order to build a simple BBN. In this way, focus is given to introducing core concepts involved with the theory without discussing advanced causal situations prematurely.

Table 1: Partial list of potential predictive variables from among two multi-institutional data sources, TCGA and REMBRANDT.

Variable	
Demographics	Total radiation dosage
Presenting age	Other drug name
Family & social history	Other drug Frequency
Environmental exposure	Other drug Dosage
Tumor location	Steroid drug name
Tumor size	Steroid frequency
Tumor grade	Steroid dosage
VEGF	Karnofsky score
EGFR VIII	Other performance score
PTEN	Tumor volume (on imaging)
TP53	Necrosis imaging finding
MGMT	Contrast enhancement imaging finding
DNA methylation	Non-contrast enhancing region
Chemotherapy drug name	Tumor multi-focality
Chemotherapy frequency/dosage	Edema volume (on imaging)
Number of chemotherapy cycles	Mass effect
Type of surgical resection procedure	Satellites
Extent of resection	ADC map (imaging)
Type of radiation therapy	Time to progression (TTP)
Radiation therapy fractionation	Time to survival (TTS; death)

Predictive Model

Our example model (Figure 2) contains four variables; 1) a demographic variable, *age* 2) a cognitive assessment variable, *Karnofsky performance score* (KPS) 3) a genetic variable, a *9-gene metagene score* derived from Colman, et al.[17], and 4) an outcome variable, *overall survival* (survival past median of 12 months). Age and KPS were chosen due to their predictive significance in previous GBM models[18–20]. Overall survival is the most common outcome variable used for prediction in previous GBM models and is most commonly predicted using median survival time cutoff[18–20].

The incorporation of a genetic variable relates to the growing interest in genetic prediction variables for cancer. For example, a number of papers in GBM treatment discuss O6-methylguanine-DNA-methyltransferase (MGMT) methylation and tumor protein 53 (TP53) gene expression as potential predictive markers for GBM patient survival. Previous work has found a significant up-regulation of MGMT expression in the tumor tissue when treated with O6-alkylating agents such as temozolomide (Temodar), indicating a potential benefit for patient's survival[21–26]. Similarly, up/down regulation of TP53 factors into cell apoptosis; reduced rates of apoptosis are characteristic of many types of cancer and can contribute to large growth rates of cancerous cells[27].

Table 2. Selected model variables.

Variable	Range/Categorical values
Age	0 (<40); 1 (40<65) ; 2 (65<80) ;3 (>80)
Karnofsky Performance Score	20,40,60,70,80,90,100 - 7 Category Assignment
Metagene Score	0 (Low, Score <= 0), 1 (High, Score > 0)
Survival past median	0 (No); 1 (Yes)

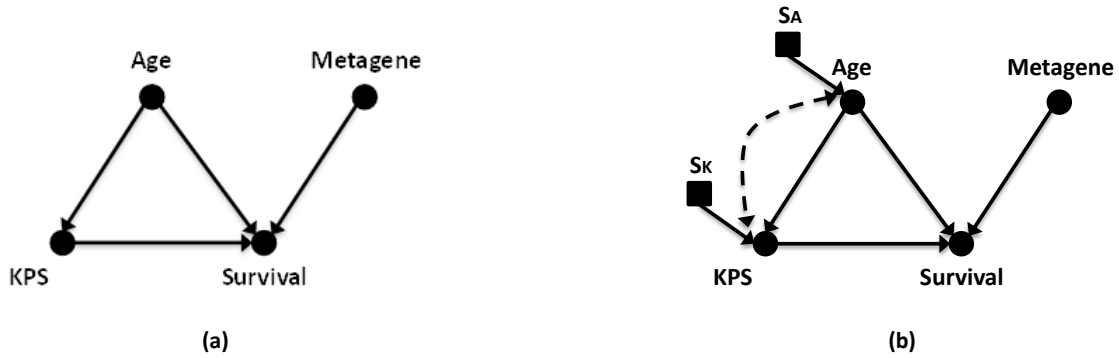


Figure 2: Example causal diagram for (a) GBM survival prediction and (b) the same causal diagram of GBM with links and nodes representing expected confounding information and population differences for variables. In the diagram, solid circular nodes represent observed variables; while square nodes indicate selection nodes controlling for population differences. Causal links are represented with solid lines with directional arrows. Bi-directional dashed lines indicate a variables linked by confounders. The selected observational variables are Patient Age (Age); Karnofsky Performance Score (KPS); 9-gene metagene score (Metagene); and Patient Survival at Population Median (Survival). Unique selection nodes for Age and KPS are shown as S_A and S_K.

Genetic testing is not available to most clinical locations and smaller locations will not have a large enough sample of cases to estimate the rates of expression for their population. Therefore, a genetic variable is a suitable example of an item from a model that would benefit from transportation from a source to target location.

As previously mentioned, our genetic variable is a metagene score derived from nine gene expression values measured in the TCGA dataset. The selection of these significant genes and the metagene scoring technique were derived from Colman, et al.[17]. Briefly, the metagene score is calculated by summation of the weighted expression levels of the genes for each patient and then discretized into high and low score classes. Discretized ranges for all variables in the model are shown in Table 2.

Data selection

Data was obtained from the TCGA public data repository. In total, 579 cases exist in the TCGA database with clinical information. Variable selection and preprocessing steps were performed on the full set of available TCGA cases with available clinical and genomic data. Cases were first selected for no evidence of prior glioma to match our analysis with the methods of Colman et al. and provide similar metagene analysis, reducing the number of available cases to 544. Because of the small number of variables in our Bayesian network we further sought to avoid missing data and the need for any imputation. A portion of the desired KPS variables were blank or unavailable for 143 cases. Finally, only cases observed until death or past the median survival cutoff are appropriate for the analysis in a Bayesian framework, so censored patients were removed. This reduced the final count of available cases for analysis to 346.

The final selected population for TCGA was then discretized based on the categories in Table 2. The dataset was divided into three source and target subsets based on contributing hospital location. Source cohorts serve as our "previous location" used for the majority of model construction and the Target cohorts are "new patients" for prediction at another location. Three target splits were made into a large, medium, and small set of cases to examine the effects of prediction when varying amounts of target data is available. The final TCGA hospitals chosen to serve as targets for analysis were Hospitals 2, 6, and 19 with contributed subjects of 84, 65, and 18 respectively. All remaining hospital locations supplied data to the complementary source cohort (262, 281, and 328 cases respectively).

Applying transportability

In the example GBM model in Figure 2a, we have a set of four variables and their causal connections. The example network comprises no connective links other than the direct causal connections derived from literature. Causal assumptions have been made in constructing this graph, and we must consider the differences that may exist between source nodes and nodes of a target population. Confounders and selection nodes are likely involved in most graphical networks and must be considered and dealt with when problematic. An example graph with a number of these issues, such as in Figure 2b, is a case where data may not be transportable unless certain constraints can be met either by transportability rules (do-calculus/d-separation mentioned below) or a valid belief that removal of the connections can be made without affecting the outcomes. The goal is to map between the messy real-world graph in Figure 2b and the ideal causal graph in Figure 2a to enable the transport of information. We review two issues below:

1) *Unobserved and confounding variables*. Let us consider the dotted connections in the example GBM network. Bidirectional dotted lines represent latent confounding variables in the causal graph. For this discussion, let us say that a connection represents interactions between age and Karnofsky performance score mediated by unmeasured variables (i.e., data that could not be observed). The addition of this link denotes that there is belief that complex biology explains the interaction between KPS and age and could mask our causal assumption. For example, KPS is derived from an examination of a patient's current mental and physical status. This status derives from a combination of the current symptomatic state of disease in the patient and some mix of other past disease. The patient's symptoms might be tied to a damaged hip from osteoporosis, causing a decreased score due to lost mobility, or be tied to a past stroke, causing a decreased score due to aphasia. These kinds of effects would mask our attempt to measure age's effect on KPS values caused exclusively by GBM. If proof exists in the literature that such an interaction is common, it might be required that additional variables be added to the model to correct the confounding before a proper transportability assessment can be made. This particular confounding example seems farfetched and we would remove the confounding link as we have a reasonable belief that clinicians are considering past injury when quantifying the KPS value.
2) *Population differences*. In addition to confounders, consideration must be given to the population differences that exist in the collected data. For example, the number of patients treated with given chemotherapies may vary between two locations depending on physician treatment preferences/experience, hospital practices, and availability of the potential drugs. Selection nodes in Figure 2b represent potential cohort differences in age and KPS scoring. Age often varies depending on the type and location of hospital where data is collected. KPS scoring can vary depending on factors such as amount of clinician training in performance scoring, the overall experience with patients in the domain, and the standard variability seen between different examiners. Adding new selection nodes changes how we consider these variables as we evaluate the ability to transport the network findings. Unless a belief or measurement can be made about invariance between the populations, selection nodes may indicate that stratification or re-estimation of variables may be warranted.

Having described the links assigned to the graph, the application of Pearl's work with transportability is possible for a given graph[28]. A set of algebraic rules called do-calculus[9,28] enables a formal mathematical statement to be derived that determines what elements of information are transportable with the given variables, relationships, confounders, and selection nodes. The do-calculus allows for links in the graph to be broken based upon forced experimental constraints. Further graphical analysis via d-separation, and front- and back-door criteria, can help determine which variables of the model are identifiable. Identification entails the evaluation of the graph edges that remain when observational data is used to set a variable to a specific state and then determining when the network is not directly affecting the transportation of findings. Thus, when a causal graph is not identifiable, its findings are not transportable. For example, KPS, a scoring of the neurological performance of a patient based on symptoms, can serve as a surrogate measure for imaging findings of brain tumor growth, which is influenced by population differences. When KPS can be determined as conditionally independent of population differences, it can serve as a replacement for the imaging information using the front door criterion of d-separation and unblock a situation where imaging findings may not be available. A full description of the do-calculus and d-separation can be found in Pearl's work[9,28]. Further individual examples can be drawn for situations involving back door paths and bidirectional counterfactual edges; many are reviewed in more detail in the available work from Pearl and Barenboim[9,29].

Network evaluation

Our Bayesian belief network predictive model for GBM was tested using custom code in MATLAB (version 7.10.0, MathWorks, Inc, Natick, MA). Source and target cohorts were built by splitting the TCGA dataset by contributing location; one TCGA participating location was held out as the Target set while all remaining data formed the Source set. For example, Hospital 2 contributed 84 cases to the 346 total TCGA dataset. A target dataset for the Hospital 2 split contains these 84 cases. Then, a source dataset was made from all remaining TCGA sites (6, 8, 12, 14, 15, 19, etc) containing 262 cases. In this way, the Source cohort acts as a previous location used for model construction and the Target cohort is a location with new patients in need of prediction. For this analysis, three splits were performed targeting locations with a large (Hospital 2, n=84), medium (Hospital 6, n=65), and small (Hospital 19, n=18) number of cases in the TCGA set. Four model considerations were used while varying the training and test cohorts from the three source-target splits. The four model considerations are: Source versus Source (SS), Target versus Target (TT), Source versus Target (ST), and Transported Source versus Target (TrST). Each consideration

Model	Training Data	Test Data
SS	All Source	Source
TT	All Target	Target
ST	All Source	Target
TrST	Age: Source, KPS: Target, Metagene: Source, Survival: Source	Target

Table 3: Description of training and test data used in the model considerations. Test data is cross validated using the leave-one-out cross validation method.

	Model							
	SS		TT		ST		TrST	
Hospital 2 (262,84)	0.69	(2.7E-08)	0.76	(1.5E-05)	0.74	(1.1E-04)	0.76	(1.7E-05)
Hospital 6 (281,65)	0.72	(1.6E-11)	0.68	(0.007)	0.63	(0.056)	0.63	(0.059)
Hospital 19 (328,18)	0.71	(4.7E-12)	0.94	(0.004)	0.68	(0.248)	0.94	(0.004)

Table 4: Leave-one-out validation results of transportability analysis. Values represented are Area under the curve (AUC) and Mann-Whitney U p-value for significant difference between survival prediction classes. Three hospitals in the TCGA dataset are compared to demonstrate the effects of target cohort size. Karnofsky performance score (KPS) was held out as missing/unmeasured data in this model.

describes the Training-Test setup used for modeling. Leave-one-out validation was performed on test cases to determine the prediction rate of the models. Mann Whitney U-tests were used to test for significant difference between prediction classes of the model.

The SS and TT examinations represent the gold standard evaluation of a model built using data from source and target locations respectively. This emulates the current state of practice where each research location builds a model rather than pooling data or using a past model. The ST examination tests the external validity of the Source model at predicting new cases from the Target cohort. The ST examination represents the special case where all variables are assumed to be trivially transportable (i.e., there are no differences affecting the Source variables, a very rare circumstance). All model probabilities are obtained from the original Source cohort with no training input from Target patients. Finally, the TrST split examines a more realistic transport where the KPS variable is trained by the Target data under the assumption that the Source KPS data is too different from the Target patients. In this model, all other variables are trained using transported values from the Source cohort. The expectation is that the joint use of information from the Target and Source datasets will outperform the ST method where differences of target information are not taken into account. Table 3 provides a full breakdown of the training and test data used in each of the four model considerations.

Results

After variable selection and application of selection criteria, 346 TCGA cases were available for analysis and were split into three source-target cohorts of Hospital 2 (262 source, 84 target), Hospital 6 (281 source, 65 target) Hospital 19 (328 source, 18 target). Each source-target cohort was then used for model training, followed by testing using leave-one-out cross validation (LOOCV) across the four described training variations (SS, TT, ST, TrST). LOOCV was chosen in order to maximize the number of cases available for the training steps, as the available hospital sample sizes are small. Complete results of the Bayesian analysis are shown in Table 4 and are discussed in more detail below.

Prediction using source training data (SS and TT)

Results from the analysis of standard methods of source and target modeling demonstrate moderate rates for LOOCV. Area under the curve (AUC) values ranged from 0.69-0.72 (LOOCV) for SS models. TT models ranged from 0.68-0.94 (LOOCV). TT models outperformed SS models in most considerations. However, the sample sizes of each TT model are smaller, reducing confidence in AUC values holding steady under all circumstances as seen by the weaker p-values in Mann-Whitney testing. We refer to the TT model AUC values as the standard accuracy target for the subsequent model applications, ST and TrST.

Prediction using outside training data (ST)

When the source model is applied as training data for a target location in the ST model, a decrease in performance is seen for all hospital combinations when compared to the TT score. In LOOCV, AUC values drop by 2.7%, 5.3%, and 26% respectively for the three hospital splits. These lower AUC values indicate that the training data from the source model is not able to predict cases in the target set as accurately as a target trained model. Results are most externally valid when populations are the same and variations between variables are minimal. Only in the large target cohort split, Hospital 2, do we see a significant differentiation ($p = 1.1E-04$) between prediction classes and therefore see potential external validity. For the other splits, external validity does not hold for the source model on outside data as p-values do not reach significance (0.056 and 0.248).

Prediction using transported probabilities (TrST)

By transporting appropriate information from the source location and using information for appropriate variables from the target location, the intention is to improve the model by making it more accurate against the differing variables and distributions at the target location. In our TrST model, the KPS value is assumed to vary between Source and Target populations. Therefore, KPS probabilities are trained using Target patient data while other variables use transported data from the Source.

In two applications of the TrST model, we see an improvement of AUC over the application of the source trained model, ST. Performance for Hospital 2 and Hospital 19 improved to match the original prediction accuracy seen in the TT model for this validation. In the case of Hospital 19,the smallest target cohort, the Mann-Whitney U test statistic also changed from being insignificant (ST p=0.248) to significant (TrST p=0.004). These improvements suggest that data from the KPS variable in the Target was able to better model local cases. Hospital 6, however, showed no improvement in accuracy between ST and TrST attempts, suggesting no significant difference between the Source and Target KPS data for this split. When these values are similar, little new information is added to improve predictions and the external validity of a source must be high to match prediction accuracies reached using target data. Detecting these cases is important to using transportability for improving model accuracies.

Discussion

The transport of probabilities for prediction from a source model to a target cohort imparted an increase in the predictive power of the model over an original source model for two of the three sites in our evaluation. These results demonstrate at a basic level the potential power of transportability theory to assess a model and determine appropriate variables for transport. Consideration of confounding and population difference that are possible in new cases is important when determining whether external validity applies to a model.

Transport of data for Hospital 2 and Hospital 19 demonstrated improvements over source models. This result indicates that the recovered target information works in concert with the transported source information to bring model predictions back to a rate similar to training a full model from scratch. Transporting data in this way can lead to more accurate models when data is unmeasurable or unavailable. In addition, it can be beneficial to future studies by reducing the number of variables that must be measured at the new location, saving time and money. Recall, in our simplified model for this work (Figure 2) we transported age, metagene, and survival information from the source. In doing so, the target was not required to provide information to estimate probabilities for these variables: only KPS data was collected from the target.

Overall, the predictive power of the models in this work is moderate, a common issue for current GBM models. While the highest AUC for a gold standard model in our analysis is 0.94, this score is for a very small dataset. Nevertheless, our results demonstrate the effect transportability can have in making an outside model useful to a target location. In two of the three splits of TCGA data, performance was improved over directly applying the source model.

One limitation within this work is the overall simplification of the problem. A probabilistic model of GBM should include a number of variables covering clinical, treatment, imaging, and genetic factors. The presented model only examines two such facets (clinical and genetic) and minimized the feature set to four specific nodes for the prediction task. This simplification was necessary for an introductory discussion of transportability, but is unrealistic as we consider the proper model options for testing external validity in the future. More realistic models with 10-20 features will complicate the ability to analyze the transportability of findings. Future analysis must examine the computational sophistication of larger disease models in order to find tractable solutions to transportability questions.

Also, low prediction rates for the current model are likely tied to the simplistic model representation chosen in order to facilitate discussion. Other statistical models have reported higher levels of time to survival prediction in GBM (AUC 0.81)[20]. In addition, we only explore a limited number of contributions from confounding information and population differences related to this model. Additional examination of location differences in the TCGA dataset might elicit the variable(s) that cause poor improvement in a situation such as Hospital 6 and indicate problems assumed away incorrectly. Providing a robust examination of factors that can disturb the external validity of data is necessary to support the claims made when completing evaluations with transportability rules. Such factors may lead to inaccurate decisions that findings are applicable externally. Overlooked or ignored confounders may destroy this capability in practice.

Another limitation of the current model is the use of a 9-gene metagene score that summarizes a set of gene expression values. The 9-gene metagene was used in this work based upon a previous assessment against TCGA data by Colman et al.[17]. An examination that treats each gene as a variable of the model rather than a summary statistic might yield improved results. In addition, many other gene expression rates are measured for these populations and more statistical examination of the predictive involvement of these genes is warranted.

Application of transportability theory to increasingly complex model designs will be important to expand the utility of this approach. For instance, adding a new variable to the model can cause a number of complications in the causal network not fully discussed here. A few considerations that might be made when adding items to the network include: how the variable was measured, what other variables it is causally connected to, how the addition affects the previous assumptions of the links in the model (changes to independence), and if measurement of the variable introduces difference into the population. As model complexity rises, it appears that the number of considerations may become difficult to appreciate.

While this work has a simplified model and assumptions, the potential utility of transportability theory is clearly demonstrated. In our working example, the step of adding confounding arcs and selection nodes in the simplified GBM model imparts that there are additional factors to consider. Honestly evaluating where these links and nodes can exist enables a discussion that faithfully considers the causal nature of the relationships described in a graph. In this way, a researcher can use transportability theory to demonstrate that issues have been considered and the assumptions made when attempting to claim findings are externally valid. With the proper examination, a mathematical description of this fact is derivable via the do-calculus.

Another potential area for future investigation is the addition of better descriptions for model design, parameters, and causal graph assumptions. One mechanism for providing this information is through the Predictive Model Markup Language (PMML), an XML-based language for describing models to improve interoperability[30]. However, PMML currently lacks descriptions for Bayesian networks, so extensions will be necessary for application. By including model information and graphs designs to experiments and RCTs, investigators can explicate the study's experimental target and the specific assumptions and decisions involved with the chosen design. The provided graph might then act as a template for outside researchers to test the model transportability against their own datasets. With the time and cost investment involved with running RCTs and cleaning observational data, these explicit descriptions could aid in the discovery process and also influence future investigations if transportable findings from a previous study mean that time and resources can be spent targeting previously unexamined variables in the domain.

Despite the strengths of transportability, there is difficulty in describing the method using more than basic examples with a minimized set of considerations such as those used above. Pearl and Bareinboim have incorporated more complex graph models in their work[9,28,29], but those models are not always contextualized in a way that is clear to the layman. In this work we have attempted to begin bridging this gap and make an attempt at introducing the core concepts of transportability. However, future work must address the best means to introduce the complex methods involved in determining external validity in this fashion. Further efforts must be made to provide descriptions of the theory that are accessible to a broader audience with an interest in testing external validity. This will require communication between computer scientists, statisticians, and informaticians to balance the descriptive language used and ensure that papers related to transportability and external validity can be published more widely.

Conclusions

Testing the external validity of scientific findings is important for the application of knowledge across populations. Transportability theory provides a robust method for describing the causal relationships of experimental variables and the circumstances that allow findings to be transported to additional populations. We have described core concepts of transportability in the context of a GBM model that describes transportability of a metagene biomarker for gene expression between two cohorts. The increase in AUC in testing for two of the three examined test cases is indicative of the utility of transporting information. The need to perform robust analysis of the potential confounders and population differences using a technique like transportability is important and future work will focus more heavily on these restrictions. The simplified model provides an understandable introduction to transportability and the examination of how poor assumptions and population differences may encourage use of models prematurely. Additional work in the area of transportability can provide a useful tool set for examining the causal relationships of experimental work and calculating the circumstances where the findings are externally valid with another population.

Acknowledgements This work was supported by National Cancer Institute grant R01 CA157553.

References

1. Bleeker S., Moll H., Steyerberg E., *et al.* External validation is necessary in prediction research: *J Clin Epidemiol* 2003;**56**:826–32. doi:10.1016/S0895-4356(03)00207-5
2. Singleton KW, Hsu W, Bui AA. Comparing Predictive Models of Glioblastoma Multiforme Built Using Multi-Institutional and Local Data Sources. *AMIA Annu Symp Proc* 2012;**2012**:1385–92.
3. Madhavan S, Zenklusen J-C, Kotliarov Y, *et al.* Rembrandt: Helping Personalized Medicine Become a Reality through Integrative Translational Research. *Mol Cancer Res* 2009;**7**:157–67. doi:10.1158/1541-7786.MCR-08-0435
4. McLendon R, Friedman A, Bigner D, *et al.* Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 2008;**455**:1061–8. doi:10.1038/nature07385
5. Rothwell PM. External validity of randomised controlled trials: 'To whom do the results of this trial apply?' *The Lancet* 2005;**365**:82–93. doi:10.1016/S0140-6736(04)17670-8
6. Petersen ML. Compound Treatments, Transportability, and the Structural Causal Model. *Epidemiology* 2011;**22**:378–81. doi:10.1097/EDE.0b013e3182126127

7. König IR, Malley JD, Weimar C, *et al.* Practical experiences on the necessity of external validation. *Stat Med* 2007;**26**:5499–511. doi:10.1002/sim.3069
8. Wiens J, Guttag J, Horvitz E. A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. *J Am Med Inform Assoc* 2014;:amiajnl–2013–002162. doi:10.1136/amiajnl-2013-002162
9. Pearl J, Bareinboim E. Transportability of Causal and Statistical Relations: A Formal Approach. In: *2011 IEEE 11th International Conference on Data Mining Workshops (ICDMW)*. 2011. 540 –547. doi:10.1109/ICDMW.2011.169
10. Howlader N, Noone AM, Krapcho M, *et al.* SEER Cancer Statistics Review 1975-2009 (Vintage 2009 Populations). http://seer.cancer.gov/csr/1975_2009_pops09/index.html (accessed 4 Aug2014).
11. Aghi M, Gaviani P, Henson JW, *et al.* Magnetic Resonance Imaging Characteristics Predict Epidermal Growth Factor Receptor Amplification Status in Glioblastoma. *Clin Cancer Res* 2005;**11**:8600–5. doi:10.1158/1078-0432.CCR-05-0713
12. Cahill DP, Levine KK, Betensky RA, *et al.* Loss of the Mismatch Repair Protein MSH6 in Human Glioblastomas Is Associated with Tumor Progression During Temozolomide Treatment. *Clin Cancer Res* 2007;**13**:2038–45. doi:10.1158/1078-0432.CCR-06-2149
13. Pope WB, Sayre J, Perlina A, *et al.* MR Imaging Correlates of Survival in Patients with High-Grade Gliomas. *Am J Neuroradiol* 2005;**26**:2466–74.
14. Chaichana K, Parker S, Olivi A, *et al.* A proposed classification system that projects outcomes based on preoperative variables for adult patients with glioblastoma multiforme. *J Neurosurg* 2010;**112**:997–1004. doi:10.3171/2009.9.JNS09805
15. Lacroix M, Abi-Said D, Fourney DR, *et al.* A multivariate analysis of 416 patients with glioblastoma multiforme: prognosis, extent of resection, and survival. *J Neurosurg* 2001;**95**:190–8.
16. Zinn PO, Majadan B, Sathyan P, *et al.* Radiogenomic Mapping of Edema/Cellular Invasion MRI-Phenotypes in Glioblastoma Multiforme. *PLoS ONE* 2011;**6**:e25451. doi:10.1371/journal.pone.0025451
17. Colman H, Zhang L, Sulman EP, *et al.* A multigene predictor of outcome in glioblastoma. *Neuro-Oncol* 2010;**12**:49–57. doi:10.1093/neuonc/nop007
18. Helseth R, Helseth E, Johannesen TB, *et al.* Overall survival, prognostic factors, and repeated surgery in a consecutive series of 516 patients with glioblastoma multiforme. *Acta Neurol Scand* 2010;**122**:159–67. doi:10.1111/j.1600-0404.2010.01350.x
19. Gutman DA, Cooper LAD, Hwang SN, *et al.* MR Imaging Predictors of Molecular Profile and Survival: Multi-institutional Study of the TCGA Glioblastoma Data Set. *Radiology* 2013;**267**:560–9. doi:10.1148/radiol.13120118
20. Mazurowski MA, Desjardins A, Malof JM. Imaging descriptors improve the predictive power of survival models for glioblastoma patients. *Neuro-Oncol* 2013;:nos335. doi:10.1093/neuonc/nos335
21. Hegi ME, Diserens A-C, Gorlia T, *et al.* MGMT Gene Silencing and Benefit from Temozolomide in Glioblastoma. *N Engl J Med* 2005;**352**:997–1003. doi:10.1056/NEJMoa043331
22. Karayan-Tapon L, Quillien V, Guilhot J, *et al.* Prognostic value of O6-methylguanine-DNA methyltransferase status in glioblastoma patients, assessed by five different methods. *J Neurooncol* 2010;**97**:311–22. doi:10.1007/s11060-009-0031-1
23. Wiewrodt D, Nagel G, Dreimüller N, *et al.* MGMT in primary and recurrent human glioblastomas after radiation and chemotherapy and comparison with p53 status and clinical outcome. *Int J Cancer* 2008;**122**:1391–9. doi:10.1002/ijc.23219
24. Chinot OL, Barrié M, Fuentes S, *et al.* Correlation Between O6-Methylguanine-DNA Methyltransferase and Survival in Inoperable Newly Diagnosed Glioblastoma Patients Treated With Neoadjuvant Temozolomide. *J Clin Oncol* 2007;**25**:1470–5. doi:10.1200/JCO.2006.07.4807
25. Esteller M, Garcia-Foncillas J, Andion E, *et al.* Inactivation of the DNA-Repair Gene MGMT and the Clinical Response of Gliomas to Alkylating Agents. *N Engl J Med* 2000;**343**:1350–4. doi:10.1056/NEJM200011093431901
26. Rivera AL, Pelloski CE, Gilbert MR, *et al.* MGMT promoter methylation is predictive of response to radiotherapy and prognostic in the absence of adjuvant alkylating chemotherapy for glioblastoma. *Neuro-Oncol* 2010;**12**:116–21. doi:10.1093/neuonc/nop020
27. Kang H-C, Kim C-Y, Han J, *et al.* Pseudoprogression in patients with malignant gliomas treated with concurrent temozolomide and radiotherapy: potential role of p53. *J Neurooncol* 2011;**102**:157–62. doi:10.1007/s11060-010-0305-7
28. Pearl J. *Causality: Models, Reasoning, and Inference*. Cambridge University Press 2000.
29. Bareinboim E, Pearl J. Transportability of Causal Effects: Completeness Results. In: *Twenty-Sixth AAAI Conference on Artificial Intelligence*. 2012. http://www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/view/5188 (accessed 14 Mar2014).
30. Data Mining Group - PMML version 4.2. http://www.dmg.org/pmml-v4-2.html (accessed 4 Aug2014).

Machine Learning for Risk Prediction of Acute Coronary Syndrome

Jacob P. VanHouten, MS[1,2], John M. Starmer, MD[1], Nancy M. Lorenzi, PhD[1], David J. Maron, MD[3], Thomas A. Lasko, MD, PhD[1]

[1]Departments of Biomedical Informatics and [2]Biostatistics, Vanderbilt University School of Medicine, Nashville TN
[3]Department of Medicine, Division of Cardiovascular Medicine Stanford University School of Medicine, Stanford, CA

Abstract

Acute coronary syndrome (ACS) accounts for 1.36 million hospitalizations and billions of dollars in costs in the United States alone. A major challenge to diagnosing and treating patients with suspected ACS is the significant symptom overlap between patients with and without ACS. There is a high cost to over- and under-treatment. Guidelines recommend early risk stratification of patients, but many tools lack sufficient accuracy for use in clinical practice. Prognostic indices often misrepresent clinical populations and rely on curated data. We used random forest and elastic net on 20,078 deidentified records with significant missing and noisy values to develop models that outperform existing ACS risk prediction tools. We found that the random forest (AUC = 0.848) significantly outperformed elastic net (AUC=0.818), ridge regression (AUC = 0.810), and the TIMI (AUC = 0.745) and GRACE (AUC = 0.623) scores. Our findings show that random forest applied to noisy and sparse data can perform on par with previously developed scoring metrics.

Introduction

Coronary heart disease is the leading cause of death worldwide, accounting for 7.0 million (11.2%) of all deaths annually. It is also the most common cause of death in the United States[1]. There are many negative sequelae of coronary heart disease, including acute coronary syndrome (ACS), which is an umbrella term for unstable angina and myocardial infarction[2]. Many patients with ACS seek treatment in the emergency departments of hospitals, where chest pain is the second most common presenting complaint[3].

There is significant variation in presentation for patients experiencing ACS. The variability includes often atypical presentations, and there is a large overlap between symptoms of a patient with ACS and those of a patient with diseases of non-cardiac etiology[4]. To appropriately rule in or out ACS, physicians rely on many different testing modalities, such as cardiac markers, electrocardiographic readings, stress tests and coronary angiography.

Despite the advancements made in diagnostic technology, between 2 and 5% of true ACS cases are reported to have findings indicative of non-cardiac disease and mistakenly discharged from the emergency room[5]. This error confers nearly twice the risk of 30-day mortality[6]. Furthermore, a missed diagnosis of ACS can be costly for a physician or hospital. Data from the Physician Insurers Association of America show that 26% of all malpractice claims against emergency departments from 1985 to 2003 were for patients evaluated for chest pain[6]. It also is estimated that greater than 25% of all medical malpractice dollars paid out to plaintiffs result from missed cases of ACS[7].

In response to the very real threat of life-endangering medical errors and malpractice litigation, emergency room doctors often perform a full gamut of tests on every patient who presents with chest pain. Of patients worked up and admitted to the hospital for ACS, only 30% are ultimately determined to have a cardiac origin for their symptoms[9]. The costly workup for suspected ACS is the main driver of the 10 to 15 billion dollars spent to accurately diagnose patients with unclear chest pain symptoms[10].

Risk stratification, or formal prediction of a patient's ACS risk at the time of presentation, is recommended by American College of Cardiology and American Heart Association guidelines[11]. Many risk stratification tools have been developed over the past 50 years to provide structure to the often unstructured process of prognostication, and these tools rely largely on the same individual tests employed in diagnosing ACS. Some of these tools, such as TIMI[12], were derived from clinical trials data. Others, such as GRACE[13], came from global registries. Still others, like HEART[14], were created by expert consensus without a defined study population. These types of models have performed with varying degrees of success (Table 1). A catalog of the tools used to risk stratify patients with potential ACS, and the performances of these various approaches, is beyond the scope of this paper. Than, Flaws, Cullen and Deely provide a useful, though not comprehensive, reference[12]. Even with these predictive tools, the rates of over- and under-treatment have remained largely unchanged.

Table 1. Selected Examples of ACS Risk Stratification Tools, Included Features, and First Performance

Tool	Source	Included Features	Performance
TIMI[13]	Clinical Trial	age, risk factors for CAD, prior coronary stenosis, ST deviation on ECG, severe anginal symptoms, aspirin use in past week, elevated serum cardiac markers	AUC = 0.74
GRACE[14]	Registry	Killip heart failure class, systolic blood pressure, heart rate, age, creatinine, presence of cardiac arrest, ST-segment deviation, elevated cardiac enzymes	AUC = 0.84
HEART[15]	Expert Opinion	history, ECG, age, risk factors, troponin	AUC = 0.89

Several limitations related to current risk models may explain this unimproved clinical performance. For instance, use of these risk scores is likely affected by their ease of application. Some, like GRACE, require the use of a computer for the complicated risk score calculation. Others, like TIMI or HEART, use as input clinical findings that are subjective in nature, which may require additional time on the part of the physician. Either of these issues might limit the use of such risk stratification tools in a busy emergency department.

Generalizability to clinical practice is another shortcoming of these models, as clinical trials or registries may not represent the typical emergency department patient population[16]. For clinical trials, strict inclusion and exclusion criteria cause older and multi-morbid patients to be underrepresented. Inclusion in many registries is dependent on final diagnosis, which can lead to problems in model development due to a lack of negative examples. In both clinical trials and registries, standardized data procedures lead to largely complete and error-free data from these sources, which is also not representative of clinical practice.

From a model development perspective, simpler model building approaches have also led to a lack of generalizability. Feature selection via stepwise selection or recursive partitioning, which have been used in several previous models, are not robust to idiosyncrasies in training data, which leads to poor performance in real emergency departments[17].

Based on these limitations, there is a need for ACS risk stratification models that can utilize data as it exists in the electronic medical record (EMR), imperfections and all. Such models would be derived from data reflective of

medical practice, and thus generalize to a broader population. They would not rely on a priori knowledge or simple methods of feature selection, but effortlessly handle large numbers of variables to make the best predictions. Ideally, these models would also not require direct human calculation or input of subjective features, but could be run in the background of an EMR system, automatically updating their predictions as new information is entered into the system, and support the clinician at the time of decision making.

While there are many machine learning approaches available for building such models, elastic net and random forest are excellent choices because they easily include many variables, resist overfitting, and provide generalizable performance. The elastic net algorithm, which combines the lasso and ridge penalties on top of regularized logistic regression, avoids overfitting and induces sparsity[18]. Random forest creates an ensemble of decision trees built using different random subsets of the features available for training, and votes for the most likely outcome between the trees to achieve generalizability of the model to future predictions[19]. In this work, we explore the performance of both of these methods.

Materials and Methods

We obtained the data for this study from Vanderbilt University Medical Center's Synthetic Derivative (SD), which is a deidentified copy of the main hospital record databases created for research purposes. The SD contains over 1.5 million deidentified electronic records with up to 15 years of data per patient with no defined exclusions. The deidentificaton of SD records is achieved through the application of a commercial program which removes HIPAA identifiers and shifts dates by a random amount of time up to one year that is unique for each individual's record. This study was approved by the Vanderbilt University Medical Center Institutional Review Board.

From this large dataset, we selected records from patients who had presented to the emergency department between the shifted dates of January 1, 2007 and May 31, 2012. Criteria for inclusion were that the patient be at least 18 years of age at the time of presentation and that they have a troponin measurement and ECG recorded during their visit to the emergency department. This selection schema is similar to an approach adopted by Vaidya, Shapiro, Lovett and Kuperman, who demonstrated their approach in identifying an ACS cohort[20]. Our method yielded similar results on our data set as theirs, and as a result we believe that we captured the vast majority of suspected ACS cases to present to the emergency department during this time period. In total, 20,078 patients were included in this data analysis, and no records were excluded for missing data.

We defined our reference standard label identifying the presence or absence of ACS using billing and procedure codes found in the patient record. We used the collection of codes proposed by one of the value-based care teams for Vanderbilt University, in which they define a patient with ACS as one who visited the emergency department for chest pain and who received any ICD-9 code listed as an ACS diagnostic code, or any ICD-9 or CPT procedure code for percutaneous coronary intervention (PCI) or coronary artery bypass graft (CABG) over the thirty days following the admission date (Table 2). If a patient was admitted twice within a short amount of time, and within thirty days of both visits received a code consistent with ACS, both admission instances were considered positive for an ACS event. Of the 20,078 encounters, we determined 8,408 to be positive for ACS by this rule (41.9%).

Table 2. Codes Use for Identification of ACS Events

Parent Code	Individual Codes
410: Acute Myocardial Infarction	410, 410.0, 410.00, 410.01, 410.02, 410.1, 410.10, 410.11, 410.12, 410.2, 410.20, 410.21, 410.22, 410.3, 410.30, 410.31, 410.32, 410.4, 410.40, 410.41, 410.42, 410.5, 410.50, 410.51, 410.52, 410.6, 410.61, 410.62, 410.7, 410.70, 410.71, 410.72, 410.8, 410.80, 410.81, 410.82, 410.9, 410.90, 410.91, 410.92
411: Other Acute/Subacute Ischemic Heart Disease	411, 411.1, 411.8, 411.81, 411.89
413: Angina Pectoris	413, 413.0, 413.1, 413.9
414: Other forms of Chronic Ischemic Heart Disease	414, 414.0, 414.00, 414.01, 414.02, 414.03, 414.04, 414.05, 414.2, 414.3, 414.8, 414.9
ICD Procedure Codes: PCI or CABG	0.66, 36.03, 36.04, 36.06, 36.07, 36.09, 36.10, 36.11, 36.12, 36.13, 36.14, 36.15, 36.16, 36.20, 36.31, 36.32, 36.3, 36.34
CPT Codes: PCI or CABG	33510, 33511, 33512, 33513, 33514, 33516, 33517, 33518, 33519, 33521, 33522, 33523, 33533, 33534, 33535, 33536, 92973, 93980, 92981, 92982, 92984, 92995, 92996

Features used in the model were included based on the structure of the data, the frequency with which they occurred, or whether they had been proven as predictors in previous models. Features were included only if they were unambiguous in their interpretation and recorded in a structured (numeric or positive/negative) form. As such, many features that we expected to be predictive were excluded, such as quality of pain, length of pain symptoms, or qualitative interpretation of ECG readings. From the subset of structured findings, we collected many known risk factors for ACS, including vital signs, laboratory values that were recorded for at least 10% of our study population, and select past medical history. For all laboratory findings we included only the first instance of that finding within their medical encounter, except for troponin and ECG measurements, where we included the first two measurements if possible. We collected 88 variables for use in our models (Table 3). No data were assessed or modified for quality assurance, regardless of the potential severity of misrepresentation or nonsense. We separated out and set aside a validation cohort (4015 records, 20%) for final testing of the single best performing model. On the remaining data, or modeling data, repeated the following procedure 100 times. We sampled with replacement to create a training set of the same size as the modeling data (16063 records), keeping the instances that were not included in this training set as a test set. Missing values in the data were then replaced by the median value for that feature in the test and training sets separately to avoid information leak.

Table 3. Features, Distributions and Missingness of Included Features, by Label

Feature	Label Negative (n=11670; 59.1%) Median (IQR)	Missing (%)*	Label Positive (n=8,408; 41.9%) Median (IQR)	Missing (%)*
Vitals:				
Age	61 (49 - 74)	0.0	66 (56 - 76)	0.0
Gender	48.05 Male	0.0	61.4% Male	0.0
Race	64.6% White	0.0	74.0% White	0.0
Height	170.2 (162.6 - 177.8)	78.4	171.3 (163.2 - 180.3)	59.8
Weight	81.0 (66.9 - 99.0)	78.5	84.6 (71.4 - 101.5)	59.9
Glucose	114 (98 - 146)	<1.0	121 (101 - 163)	<1.0
Systolic	130 (113 - 149)	28.2	130 (114 - 148)	10.5
Diastolic	71 (60 - 82)	28.2	70 (60 - 81)	10.5
Pulse	86 (74 - 100)	88.6	80 (70 - 93)	88.5
Resp Rate	18 (16 - 20)	85.3	18 (16 - 20)	87.2
1st ECG:				
HR from ECG	84 (71 - 100)	<1.0	79 (68 - 94)	<1.0
PR Interval	156 (140 - 179)	<1.0	163 (140 - 184)	<1.0
QRS Duration	92 (84 - 104)	<1.0	98 (88 - 118)	<1.0
QT Interval	454 (432 - 478)	<1.0	459 (435 - 488)	<1.0
P Wave	47 (16 - 65)	1.2	45 (4 - 63)	1.3
Initial 40 ms	28 (4 - 50)	<1.0	19 (-5 - 48)	<1.0
Mean QRS	19 (-14 - 53)	<1.0	14 (-23 - 52)	<1.0
Terminal 40 ms	28 (-27 - 87)	<1.0	30 (-32 - 96)	<1.0
ST Segment	78 (32 - 152)	<1.0	118 (55 - 176)	<1.0
2nd ECG:				
HR from ECG	83 (70 - 100)	63.8	78 (66 - 94)	45.0
PR Interval	160 (136 - 180)	64.1	164 (144 - 188)	45.9
QRS Duration	92 (84 - 104)	63.8	96 (86 - 112)	45.0
QT Interval	455 (433 - 480)	63.8	460 (433 - 488)	45.0
P Wave	45 (12 - 64)	64.4	44 (6 - 62)	46.5
Initial 40 ms	28 (4 - 50)	63.8	18 (-5 - 46)	45.1
Mean QRS	18 (-15 - 51)	63.8	14 (-21 - 50)	45.0
Terminal 40 ms	25 (-27 - 86)	63.8	30 (-28 - 93)	45.0
ST Segment	81 (31 - 159)	63.8	116 (51 - 177)	45.0
Past Medical Hx:				
Smoking	16.6%		10.7%	
Diabetes	21.1%		12.8%	
Hyperlipidemia	28.0%		20.7%	
Hypertension	41.9%		23.2%	
CABG	1.3%		3.4%	
PCI	3.3%		9.5%	
MI	22.3%		22.1%	
CAD	22.6%		22.1%	
Stroke	14.9%		9.8%	
Heart Failure	18.7%		13.8%	
Past Medications:				
Aspirin	37.8%		24.2%	
Metoprolol	38.0%		24.2%	
Nitroglycerine	20.5%		18.5%	

Table 3. (Continued)

Feature	Label Negative (n=11670; 59.1%) Median (IQR)	Missing (%)*	Label Positive (n=8,408; 41.9%) Median (IQR)	Missing (%)*
Labs:				
WBC	8.5 (6.5 - 9.9)	1.7	8.5 (6.6 - 11.1)	<1.0
PCV	38 (34 - 42)	1.5	39 (34 - 42)	<1.0
Platelet Ct	228 (179 - 287)	1.6	221 (176 - 276)	<1.0
RBC	4.3 (2.8 - 4.8)	1.7	4.3 (3.8 - 4.8)	<1.0
MCV	89 (85 - 93)	1.7	90 (86 - 93)	<1.0
MCH	29.7 (28.1 - 31.2)	1.7	29.8 (28.2 - 31.3)	<1.0
MCHC	33.2 (32.2 - 34.2)	1.7	33.1 (32.1 - 34.1)	<1.0
RDW	14.2 (13.3 - 15.8)	1.7	14.3 (13.4 - 15.7)	<1.0
RDWSD	46.5 (43.2 - 51.3)	23.5	46.6 (43.4 - 51.2)	19.7
Abs Neuts	5.7 (3.9 - 8.5)	11.7	5.6 (4.1 - 8.1)	8.6
% Neuts	69.4 (59.0 - 79.2)	11.7	68.7 (59.6 - 77.3)	8.6
Abs Lymphs	1.58 (1.05 - 2.24)	11.7	1.59 (1.10 - 2.24)	8.6
% Lymphs	19.4 (11.4 - 28.8)	11.7	19.8 (12.7 - 28.0)	8.6
Abs Monos	.64 (.47 - .88)	11.7	0.66 (0.50 - 0.89)	8.6
% Monos	7.7 (5.9 - 9.7)	11.7	7.9 (6.3 - 9.8)	8.6
Abs Eos	0.11 (0.04 - 0.20)	13.7	0.13 (0.06 - 0.23)	10.3
% Eos	1.3 (0.5 - 2.6)	13.6	1.6 (0.7 - 2.9)	10.3
Abs Basos	0.03 (0.02 - 0.04)	13.0	0.03 (0.02 - 0.04)	9.7
% Basos	0.3 (0.2 - 0.5)	13.0	0.3 (0.2 - 0.5)	9.7
% NRBC	0.0 (0.0 - 0.0)	24.9	0.0 (0.0 - 0.0)	20.6
NRBC Ct	0.0 (0.0 - 0.0)	25.0	0.0 (0.0 - 0.0)	20.6
BUN	16 (11 - 26)	<1.0	19 (13 - 29)	<1.0
HGB	12.7 (11.0 - 14.2)	1.7	12.7 (11.1 - 14.2)	<1.0
Na	138 (136 - 140)	1.0	138 (136 - 140)	<1.0
K	3.9 (3.6 - 4.3)	<1.0	4.0 (3.6 - 4.4)	<1.0
Cl	103 (100 - 106)	1.0	103 (100 - 106)	<1.0
CO2	25 (23 - 27)	1.0	25 (23 - 27)	<1.0
BNP	214 (69 - 628)	64.5	314 (106 - 843)	48.3
Creatinine	1.08 (0.86 - 1.54)	<1.0	1.20 (0.95 - 1.68)	<1.0
Ca	9.1 (8.7 - 9.4)	1.7	9.1 (8.7 - 9.4)	1.1
Anion Gap	9 (8 - 11)	1.9	9 (8 - 11)	<1.0
Total Protein	6.5 (5.8 - 7.1)	90.0	6.3 (5.8 - 6.9)	90.0
Albumin	3.3 (2.8 - 5.0)	87.1	3.5 (3.1 - 3.8)	87.3
Alk Phosp	78 (61 - 104)	47.8	75 (60 - 98)	47.1
AST	27 (21 - 41)	47.0	27 (21 - 39)	46.4
ALT	21 (15 - 34)	47.6	21 (15 - 31)	46.9
CK-MB	2.3 (1.5 - 4.0)	22.9	2.8 (1.7 - 4.9)	12.7
Creat Phos Kin	88 (50 - 162)	3.4	92 (56 - 161)	1.4
Lipase	26 (19 - 36)	75.3	26 (20 - 37)	77.7
MB Ratio	2.3 (1.4 - 3.8)	24.6	2.8 (1.8 - 4.7)	13.7
INR	1.2 (1.1 - 1.7)	58.9	1.2 (1.0 - 1.7)	52.0
Urine Sp Gr	1.01 (1.01 - 1.02)	46.1	1.01 (1.01 - 1.02)	1.3
TSH	1.8 (1.0 - 3.2)	74.1	1.8 (1.0 - 3.1)	31.5
EGFR	67.1 (43.8 - 88.4)	1.9	59.8 (39.9 - 79.6)	<1.0
1st Troponin	0.02 (0.01 - 0.05)	0	0.04 (0.02 - 0.13)	0
2nd Troponin	0.02 (0.01 - 0.05)	10.4	0.04 (0.02 - 0.17)	6.9

* Features without missing percentage were either present or absent in medical record.

We compared the elastic net and random forest algorithms against logistic regression with only the ridge penalty, a common method of benchmarking. We performed parameter tuning by an additional step of ten-fold cross-validation. For ridge regression, we selected the lambda penalization parameter in this way. For elastic net, we selected the lambda penalization parameter, as well as the alpha mixing parameter, by grid search from a possible range of 0 to 1 by intervals of 0.1. For random forests, we selected the values for the number of trees (500) and the percentage of features (square root of total predictors) explored at each random split by grid search and human guided search.

We compared our models to the TIMI and GRACE scores. For both TIMI and GRACE scores, available features were included and appropriately transformed to the scale of these scores. Our data did not include all variables needed to populate these models, but a prior study demonstrated a method for imputing missing features for these scores, though their study population experienced significantly fewer missing values[21]. In accordance with the methods of this study, data that were missing were either replaced by a reasonable surrogate, such as any history of past aspirin use replacing aspirin use within the past seven days for the TIMI score, or they were replaced with the most common clinical finding, such as all records being considered negative for cardiac arrest on admission. As a result, our calculated TIMI and GRACE scores had possible point totals of 5 and 246, compared to fully populated maximums of 7 and 372.

All five models were evaluated by visual comparison of receiver operating characteristic (ROC) curve plots and by calculation of the area under the curve (AUC) for their performance on the test data sets. We calculated the median performance and 95% confidence intervals for AUC over the 100 replicates for each algorithm. After identifying a clearly best-performing algorithm, we built a model using all of the modeling data, and tested that model on the validation set to obtain a generalizable estimate of performance. Analysis was carried out using R version 3.0.1 and the packages randomForest, glmnet, ROCR, Epi and rms downloaded from the CRAN website[22–27].

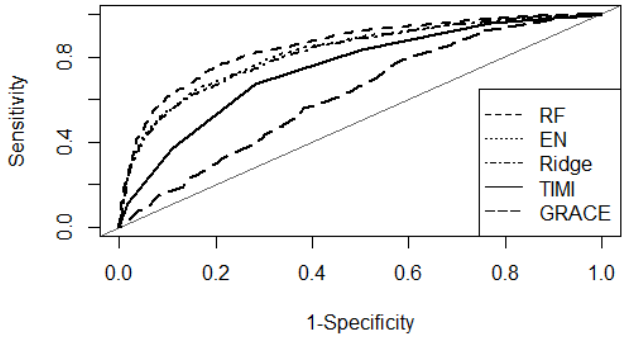
Results

Over all 100 replicates, the random forest outperformed all other methods examined (Figure 1, Table 4). Figure 1 represents a typical run of these five algorithms in our experiment.

Table 4. Performance of Compared Model in this Study

Model	Median [95% CI] AUC over 100 Replicates
Random Forest	0.848 [0.841, 0.857]
Elastic Net	0.818 [0.808, 0.828]
Ridge Regression	0.810 [0.801, 0.820]
Modified TIMI	0.745 [0.737, 0.755]
Modified GRACE	0.623 [0.615, 0.634]

Figure 1: ROC for Prediction of ACS on Test Set



The elastic net, while not performing as well as the random forest, also did not significantly reduce the complexity of the model, retaining on average 75% of the coefficients in the full regression model. There was no significant difference between the elastic net and the ridge penalized logistic regression. Random forest, elastic net and ridge regression all outperformed the modified TIMI and GRACE scores.

The generalization performance of the random forest on the validation set was good, with AUC = 0.849.

Discussion

We developed models using the random forest and elastic net algorithms utilizing a novel clinical database that improved upon existing ACS risk prediction tools and which used only structured inputs from an EMR. This work represents the first attempt that we know of to use data from the uncontrolled environment of a non-trial, non-registry emergency department for the development of such a model. These data more fairly represent the heterogeneous patient population that comprises the cohort of suspected ACS cases nationwide. We have also demonstrated that computationally advanced models can potentially outperform simpler statistical methods. While our random forest and elastic net models outperformed the TIMI and GRACE models, we found random forest to significantly outperform all other models.

Random forests have many advantages that make them useful for classification. They are able to use all of the available data without overfitting, and can also utilize data that is largely missing or that contains multiple recording errors. In the case of these experiments, we worked with an available large dataset which contained many outliers and no apparent quality control. As an example, height and weight were recorded with a mixture of metric and standard units. For some features, as much as 90% of the data were missing. Even with un-curated and sparse data, we were able to produce a classifier that outperformed previously developed methods in a head to head comparison.

While the random forest automatically explores complex interactions between variables, these interactions must be manually specified in regression-type approaches. As it was not clear a priori which interactions were important, and it would be intractable to include all possible interactions, we elected not to include interaction terms in our elastic net and ridge regression models. This may account for a portion of the apparent outperformance of the regression methods by random forest, while also demonstrating the capability of random forests to automatically select complex interactions which would not be available to other methods without user specification. We further explored this capability of random forest using only the features used to calculate the TIMI score. Unbinning the discretized variables from TIMI improved predictive ability, but random forest remained superior to logistic regression. We also found that creating flags for missing features did not significantly affect our results (data not shown).

One key limitation of this study is the availability of data in the system with which we worked. Many of the timestamp entries, in addition to being shifted through the de-identification process used in creating the SD, were truncated to one-day resolution. While temporal specificity down to the minute is not important in some studies, such as genome-wide or phenome-wide association studies, knowing the correct relationship between event times in the complex and sometimes brief encounter with a patient suspected of ACS is critical. Despite the lack of temporal clarity, our model still performed on par with existing methods. Additionally, we were unable to determine which patients discharged from Vanderbilt presented again within 30 days to an outside hospital, resulting in a possible underestimation of the total number of positive events. Use of identified data in the future would allow us to consider patients whose ACS event might be unobservable in the current research paradigm. In the current version of the SD, we were unable to capture the outcome of death, which may have also led to undercounting.

Another limitation of this study is in the definition of positive events in our cohort. We believe there may be label bias due to use of ICD-9 and CPT codes to define an ACS event, but were unable to identify other approaches that would solve this issue. Increasing the number of ICD-9 codes required for a positive label, which has improved phenotype identification from electronic medical records in previous studies, did not significantly improve our performance here (data not shown). Our definition of an ACS event relies heavily on work done by the value-based care group at Vanderbilt University Medical Center. There are several other definitions, including the World Health Organization-accepted universal definition of myocardial infarction. However, alternative methods of identifying ACS-positive records would have required significant manual chart review.

The purpose of our model is to help physicians stratify patients by risk of ACS, in order to more appropriately deliver care. In this context, missing an ACS case would be extremely undesirable. We believe that the majority of suspected ACS patients were included in our analysis, due to the low threshold for inclusion in our study. It is possible, however, that some atypical patients were not included.

Our model has good discriminative ability. However, in order to be clinically useful, such a model must also perform well on the lowest risk individuals, which can be assessed via a calibration curve. Our model had good calibration, but improvement is needed before we physicians could confidently use this model as a decision support tool.

With the continued growth of electronic medical records, automated approaches such as this one will gain traction as providers seek to leverage the expanse of data. If confirmed, this tool could ultimately improve the quality of care by reducing unnecessary admissions and discharges of patients with acute chest pain, and help shift the cost curve for healthcare that has so heavily weighed on the medical system by addressing one extreme case of resource overutilization.

Acknowledgements

This work was partially supported by the National Library of Medicine Training Grant 5T15LM007450-09. Clinical data was provided by the Vanderbilt Synthetic Derivative, which is supported by institutional funding and by the Vanderbilt CTSA grant ULTR000445.

References

1. WHO | The top 10 causes of death. World Health Organization; Available from: http://who.int/mediacentre/factsheets/fs310/en/
2. Kumar A, Cannon CP. Acute coronary syndromes: diagnosis and management, part I. Mayo Clin Proc . 2009 Nov;84(11):1021–36.
3. Hess EP, Wells G a, Jaffe A, Stiell IG. A study to derive a clinical decision rule for triage of emergency department patients with chest pain: design and methodology. BMC Emerg Med . 2008 Jan;8:3.

4. Scirica B. Acute coronary syndrome: emerging tools for diagnosis and risk assessment. J Am Coll Cardiol . Elsevier Inc.; 2010 Apr 6;55(14):1403–15.
5. Goodacre SW, Cross E, Arnold J, Angelini K, Capewell S, Nicholl J. The health care burden of acute chest pain. Heart . 2005 Feb;91(2):229–30.
6. Pope J, Aufderheide TP, Ruthazer R. Missed diagnoses of acute cardiac ischemia in the emergency department. Engl J . 2000;342(16):1163–70.
7. Rusnak RA, Stair TO, Hansen K, Fastow JS. Litigation against the emergency physician: common features in cases of missed myocardial infarction. Ann Emerg Med . Elsevier; 1989 Oct 1;18(10):1029–34.
8. Strehlow M. Chest Pain : Observation & Rapid Rule Outs. 2011.
9. Ekelund U, Nilsson H-J, Frigyesi A, Torffvit O. Patients with suspected acute coronary syndrome in a university hospital emergency department: an observational study. BMC Emerg Med . 2002 Oct 3;2(1):1.
10. Limkakeng A, Gibler WB, Pollack C, Hoekstra JW, Sites F, Shofer FS, et al. Combination of Goldman risk and initial cardiac troponin I for emergency department chest pain patient risk stratification. Acad Emerg Med . 2001 Jul;8(7):696–702. t
11. Anderson JL, Adams CD, Antman EM, Bridges CR, Califf RM, Casey DE, et al. 2011 ACCF/AHA Focused Update Incorporated Into the ACC/AHA 2007 Guidelines for the Management of Patients With Unstable Angina/Non-ST-Elevation Myocardial Infarction: A Report of the American College of Cardiology Foundation/American Heart Association Tas. Circulation . 2011 Mar 28;410–528.
12. Than M, Flaws DF, Cullen L, Deely JM. Cardiac Risk Stratification Scoring Systems for Suspected Acute Coronary Syndromes in the Emergency Department. Curr Emerg Hosp Med Rep . 2013 Jan 8;1(1):53–63.
13. Antman EM, Cohen M. The TIMI risk score for unstable angina/non–ST elevation MI. JAMA J Am Med Assoc . 2000;284(7):835–42t
14. Granger CB, Goldberg RJ, Dabbous O, Pieper KS, Eagle K a, Cannon CP, et al. Predictors of hospital mortality in the global registry of acute coronary events. Arch Intern Med . 2003 Oct 27;163(19):2345–53.
15. Six a J, Backus BE, Kelder JC. Chest pain in the emergency room: value of the HEART score. Neth Heart J. 2008 Jun;16(6):191–6
16. Yan AT, Jong P, Yan RT, Tan M, Fitchett D, Chow C-M, et al. Clinical trial--derived risk model may not generalize to real-world patients with acute coronary syndrome. Am Heart J . 2004 Dec;148(6):1020–7.
17. Yan AT, Yan RT, Tan M, Casanova A, Labinaz M, Sridhar K, et al. Risk scores for risk stratification in acute coronary syndromes: useful but simpler is not necessarily better. Eur Heart J . 2007 May;28(9):1072–8.
18. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B (Statistical Methodol . 2005 Apr;67(2):301–20.
19. Breiman L. Random forests. Mach Learn . Springer; 2001 May;45(1):5–32.
20. Vaidya SR, Shapiro JS, Lovett PB, Kuperman GJ. Acute coronary syndrome cohort definition: troponin versus ICD-9-CM codes. Future Cardiol . Future Medicine Ltd London, UK; 2010 Sep 8;6(5):725–31.
21. Goodacre SW, Bradburn M, Mohamed A, Gray A. Evaluation of Global Registry of Acute Cardiac Events and Thrombolysis in Myocardial Infarction scores in patients with suspected acute coronary syndrome. Am J Emerg Med . Elsevier B.V.; 2012 Jan;30(1):37–44.
22. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria; 2013. Available from: http://www.r-project.org/
23. Liaw A, Wiener M. Classification and Regression by randomForest. R News [Internet]. 2002;2(3):18–22. Available from: http://cran.r-project.org/doc/Rnews/
24. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw . 2010;33(1)./
25. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics . 2005;21(20):7881
26. Carstensen B, Plummer M, Laara E, Hills M. {Epi}: A Package for Statistical Analysis in Epidemiology [Internet]. 2013. Available from: http://cran.r-project.org/package=Epi
27. Harrell F. rms: Regression Modeling Strategies [Internet]. 2013. Available from: http://cran.r-project.org/package=rms

Enabling claims-based decision support through non-interruptive capture of admission diagnoses and provider billing codes

Colin G. Walsh, MD, [1,2] David K. Vawdrey, PhD[1,3], Peter D. Stetson, MD, MA [2,1] Matthew R. Fred, MD[3], George Hripcsak, MD, MS[1]

[1]Department of Biomedical Informatics, Columbia University, New York, NY;
[3]Department of Medicine, Columbia University, New York, NY
[3]Department of Information Technology, NewYork-Presbyterian Hospital, New York, NY

Abstract

The patient problem list, like administrative claims data, has become an important source of data for decision support, patient cohort identification, and alerting systems. A two-fold intervention to increase capture of problems on the problem list automatically – with minimal disruption to admitting and provider billing workflows – is described. For new patients with no prior data in the electronic health record, the intervention resulted in a statistically significant increase in the number of problems recorded to the problem list (3.8 vs 2.9 problems post- and pre-intervention respectively, p value $2x10^{-16}$). The majority of problems were recorded in the first 24 hours of admission. The proportion of patients with at least one problem coded to the problem list within the first 24 hours increased from 94% to 98% before and after intervention (chi square 344, p value $2x10^{-16}$). ICD9 "V codes" connoting circumstances beyond disease were captured at a higher rate post intervention than before. Deyo/Charlson comorbidities derived from problem list data were more similar to those derived from claims data after the intervention than before (Jaccard similarity 0.3 post- vs 0.21 pre-intervention, p value $2x10^{-16}$). A workflow-sensitive, non-interruptive means of capturing provider-entered codes early in admission can improve both the quantity and content of problems on the patient problem list.

Introduction

Administrative claims remain the bulwark of medical billing, and they are also frequent elements of decision support systems including clinical alerts, comorbidity capture, and predictive models. At the same time, the patient problem list has evolved from the "Problem-Oriented Medical Record" defined by Dr. Lawrence Weed in 1968 to an area of research and application of clinical informatics.(1, 2) Both administrative claims abstracted by billers and patient problems derived from provider documentation or provider-entered codes may fall into similar classification schema like the International Classification of Diseases (ICD).(3) But while the literature on claims and on the problem list has expanded markedly since the 1990s, studies evaluating the intersection between these codes are less common.

Because of their ubiquity and classification standards primarily through ICD9/10, administrative claims and their secondary use touch on domains across quality, patient safety, decision support, prediction, personalized medicine, and more. A review of all of these applications would be exhaustive. A cogent example exists in the interplay of diagnostic and pharmacy claims data on medication management.(4-6) A Dutch study in 2013 demonstrated up to 38% of drug therapy alerts failed to appear because of missing information in the electronic patient record; of the 442 records considered, disease information was missing in 83%.(7) A systematic review outlined statistically significant reductions in medication errors in patients with renal insufficiency and in pregnant patients in studies of alerting systems in the electronic medical record.(8)

Biller-assigned administrative claims do come with their own limitations and biases. From predicting mortality to identifying complications particularly in work led by Iezzoni, administrative claims alone may be insufficient data sources for particular tasks.(9-11) Code "creep" – overbilling for more codes than are supportable by documentation – is well-described.(12-14) However, there remains another critical limitation of systems relying on administrative claims; these claims are not coded until after a patient has been discharged and therefore are not available to any of the panoply of systems waiting to use them until days post discharge.

The patient problem list offers some of the advantages of administrative claims – structured data, easily integrated into decision support or quality reporting. Indeed, a coded problem list is a core objective of Meaningful Use, Stage I.(15, 16) A small amount of research has linked problem lists to higher quality care such as increased rates of appropriate prescription of ACE inhibitors or Angiotensin Receptor Blockers for patients with more accurate problem lists; similarly, adding chronic health concerns like obesity to the problem list increase rates of providers

addressing these problems with patients.(17, 18) A number of studies since the 1990s have outlined approaches to maximize the accuracy, completeness, and ease of populating problem lists through methods as varied as direct provider documentation of problems, natural language processing, inference rules, and wikis.(19-29) Some of these approaches are computationally intensive, and others may alter workflows.

Study Aims

The aim of this study is to evaluate a two-fold intervention built into existing provider workflows to increase documentation of problems on patient problem lists. One intervention is the conversion of the "Admitting Diagnosis" Field in the Admit Patient Order Set from a free text field to a structured data entry field using a diagnostic synonym lookup table. The second intervention is the alignment of a daily provider workflow – Evaluation/Management Billing (E/M) by providers on their own documentation – with the population of the patient problem list using a tool locally known as iCharge.

Methods

Admitting Diagnosis Problem List Intervention

The initial order set for every patient admitted to the medical center starts with the Admit Patient Order in the inpatient electronic medical record, Allscripts Sunrise Clinical Manager. The admitting diagnosis must be completed, and until fall 2013, this field was a free text entry field that was pre-filled with free text data obtained through the Admit/Discharge/Transfer Registration System. In November 2013, the field was converted into a structured order entry field that asked providers to continue inputting an admitting diagnosis but provided a means to enter a diagnosis that correlated with an ICD9 code on the patient problem list.(Figure 1) A synonym lookup table provided by Intelligent Medical Objects (IMO), a third party interface terminology,(32) permitted the entry of common abbreviations and natural language to obtain a structured diagnosis for the reason for admission. The order set was not changed in any other way and minimal education around this intervention was required.

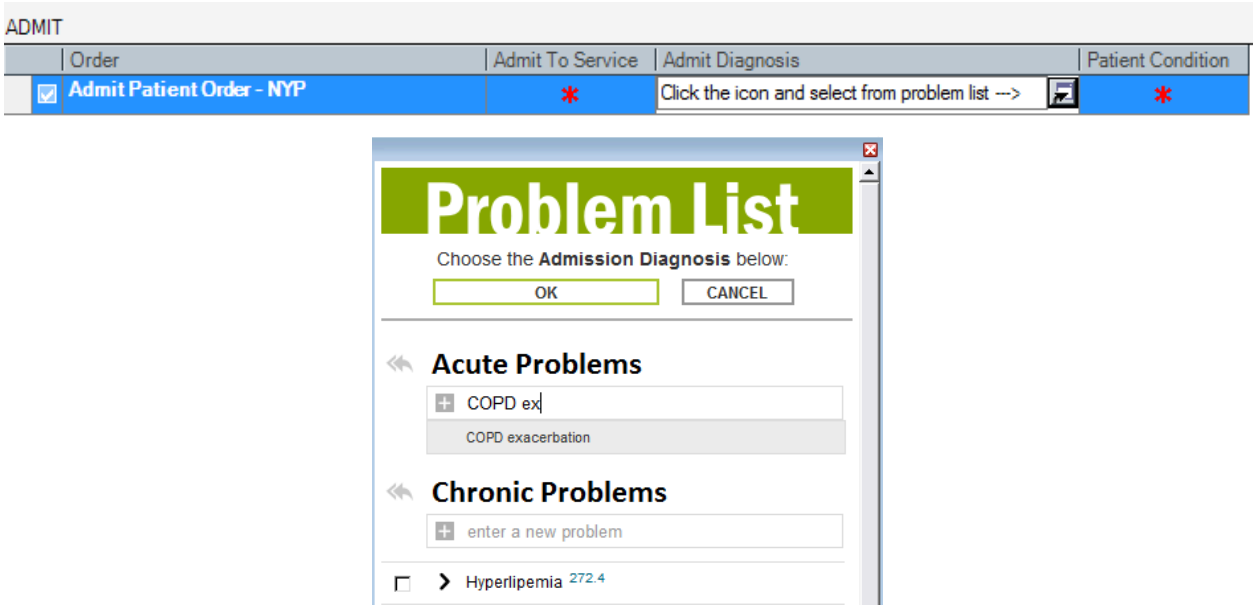


Figure 1. Activation of the custom Problem List Tool (bottom) via the Admit Order Admit Diagnosis Field (top)

In activating the problem list tool, providers were also able to enter secondary diagnoses at the same time. These diagnoses were categorized as acute, chronic, and prior, to reflect common clinical categorizations of a patient's past medical history. The problem list tool was also available at the time of inpatient documentation entry and permitted pasting already coded diagnoses into notes directly under "Past Medical History" (Figure 2).

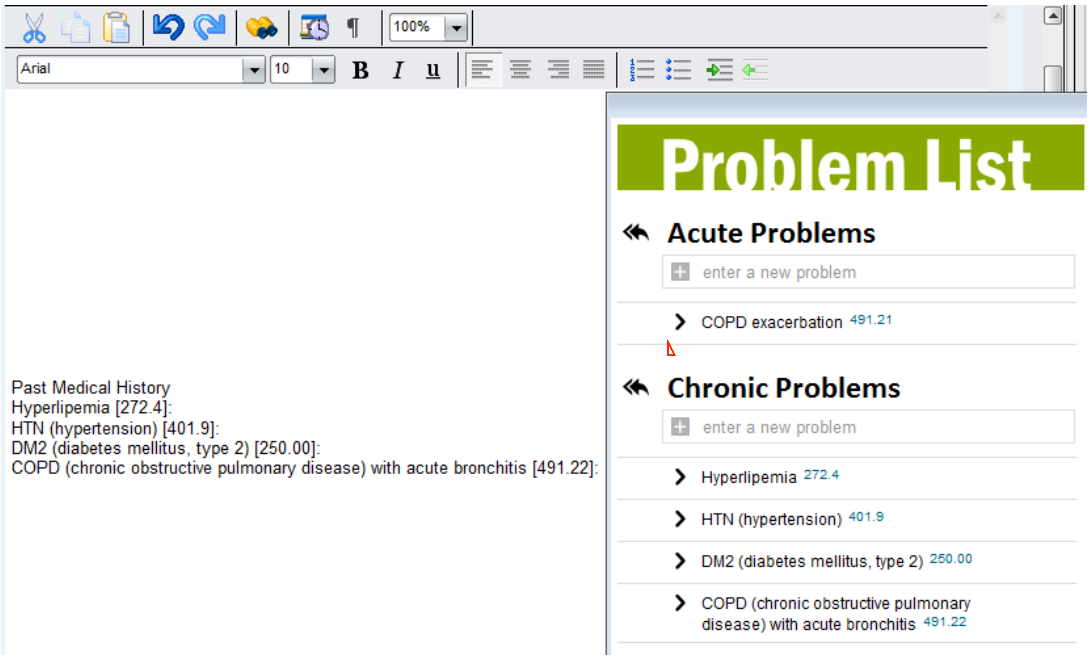


Figure 2. Problem List Integration into Inpatient Notes. Clicking the arrow (circled in red) pastes the diagnoses into the free text note shown on the left

iCharge – Provider E/M Billing Tool with Transparent Connection to the Patient Problem List

In January 2011, the iCharge tool was implemented into provider billing workflows. The prior billing process incorporated paper billing slips submitted by providers after they'd completed requisite inpatient documentation. iCharge converted this process into an electronic "widget" that was readily accessible in the electronic health record. The iCharge workflow started with the provider selecting the inpatient note for which a bill was to be generated. A link in the iCharge tool opened the note itself for providers to review details and to verify the correct note would generate the bill. iCharge was applicable to initial and follow-up provider documentation throughout a patient's admission.

Once a bill and an encounter had been selected, iCharge directed providers to link at least one coded diagnosis to the encounter. From 2011 through November 2013, iCharge permitted but did not require the storage of coded diagnoses to patient problem lists. An option was provided to save coded diagnoses for an encounter back to the problem list, known as "Health Issues".

On November 7, 2013, iCharge was upgraded with a core functionality change – the automatic addition of coded diagnoses to the patient's longitudinal problem list – this will be referred to as "auto-save" for purposes of this work. The option for providers to manually save coded diagnoses back to the problem list was removed but the remainder of iCharge functionality, while enhanced incrementally, was intact.

Throughout the intervention time period, traditional administrative medical billing was conducted by the billing department of New York Presbyterian Hospital. The billing codes assigned to each inpatient admission post discharged were stored in the clinical data warehouse for the hospital.

Data Collection

All health issues/problem list data were extracted from problem list tables in the clinical data warehouse. Administrative claims data assigned to inpatient admissions from January 1, 2011, to March 1, 2014, were also extracted for comparison to the problem list. The admissions occurred at Columbia University Medical Center in New York, NY. Admissions to affiliate hospitals or to other hospitals in New York were not included.

In addition to the Problem List Tool and iCharge, it is possible for providers to enter problem list codes into the problem list via "Health Issues". This step is an additional and rare action in the provider clinical workflow, although precise frequency data for this means of problem entry was not separable from the main data sources in this study.

Evaluation Methods: Change in Problem List Codes Before and After Intervention

The evaluation of the problem list interventions - 1) Admitting Problem List Tool; 2) iCharge and Problem List Integration via auto-save – is based in a comparison of the quality and quantity of problem list codes before and after these interventions. Problem List auto-save was activated on November 7, 2013. The four months prior to auto-save activation was compared to the period following its activation. Attributes of problem list entry were collected including provider type, the time of data entry, and the coded problems themselves.

An important subgroup of this analysis remains patients who have never been encountered in the medical center before. While these patients will not have administrative claims assigned until after they are discharged home, they will have problems assigned to the problem list via the Admitting Diagnosis, iCharge, and the custom Problem List tool.

One subset of ICD9 coding, "V Codes" (V00-V91), correspond to diagnostic circumstances beyond disease definitions. These codes range from "Normal Pregnancy, V22", to "Artificial Opening Status, V44" (gastrotomy tubes, etc) to "Housing, household, and economic circumstances, V60". Factors of mental illness such as suicidal or homocidal ideation are similarly captured by these codes. While not a complete surrogate to psychosocial determinants of health, V codes do capture aspects of a patients' clinical history that are not well-defined in other areas of the ICD9 schema. The number and type of V Codes entered via the problem list interventions were recorded and compared.

Evaluation Methods: Comparison of Problem List to Administrative Claims

Problem list data entered by providers was compared to administrative claims data entered by medical billers using the Charlson/Deyo Comorbidity Index.(30, 31) The Charlson/Deyo comorbidity index aggregates ICD9 coding into clinical categories.(31) ICD9 codes 250-250.3 and 250.7, for example, are aggregated into "Diabetes" while ICD9 codes 250.4-250.6 are aggregated into "Diabetes with Chronic Complications." Both ICD9s derived from the problem list over the course of an admission and those derived from administrative claims data after discharge were aggregated according to Deyo/Charlson comorbiditiy indices. While there is no gold standard indicating how similar problems on the problem list should actually be, it is possible to compare the similarity of the vectors of Deyo/Charlson comorbidity indices between the problem list and administrative claims.

To compare these categorical binary vectors (the presence or absence of ICD9 codes corresponding to the Deyo/Charlson comorbidities without weighting), the metric of Jaccard similarity was used.(36) In the Jaccard index of similarity, a typical approach for two vectors, A and B, is:

$$J = \frac{M_{11}}{M_{01} + M_{10} + M_{11}}$$

where M_{11} is the number of attributes present in both A and B, M_{10} is the number of attributes present only in A, and M_{01} is the number of attributes present only in B. A higher value of J corresponds to greater similarity between A and B.

Data was processed and analyzed in the R statistical environment using packages, "plyr" and "data.table".(33-35)

This study was approved by the Institutional Review Board.

Results

From January 1, 2010, through March 1, 2014, there were 271,615 inpatient admissions corresponding to 140,538 unique medical record numbers. There was an average of 5,400 inpatient admissions per month. In any given month, an average of 2,800 new medical record numbers were assigned corresponding to patients who have not had prior encounters in the medical center and to a very small number that are inadvertently given a second medical record number at registration on a subsequent admission.

Administrative Claims Data

There were 2,785,658 ICD9 codes assigned to the entire study cohort since 2010. There were an average of ten ICD9 codes assigned to each admission. The number of ICD9 codes assigned ranges from a minimum of 1 code to a maximum of 51.

Problem List auto-save was activated on November 7, 2013. The four months prior to auto-save activation was compared to the period following its activation. With respect to medical billing assigned codes irrespective of the problem list, there was a decrease in the average number of ICD9 coded by medical billers in period after auto-save compared to the period before (10.8 ICD9 codes assigned prior to auto-save compared to 9 codes on average after, p value $2x10^{-16}$).

Administrative claims are linked to inpatient admissions in the clinical data warehouse, but they are not assigned until patients have been discharged from the hospital.

Problem List Data

Over the entire study period, there were 319,420 problems entered onto patient problem lists. All providers either completing billing through iCharge (physicians) or entering admission orders through the electronic health record could impact the problem list. There was also an option to bring up the problem list independently in the electronic health record and any member of the care team could do so. Table 1 outlines the usage of the problem list by provider type over the study time period.

Provider Type	Number of Health Issues Entered	Percentage
Physicians – attendings, fellows	270,958	84.8%
Physicians – residents	19,797	6.2%
Physician Assistants	15,012	4.7%
Nurse Practitioners	5,719	1.8%
Medical Students	496	0.2%
Physical or Occupational Therapists	1,308	0.4%
Nurses	159	0.04%
Other or Not Recorded	5,971	1.8%

Table 1. Problem List Entry Count by Provider Type

The timing of entry of health issues was considered in aggregate and specifically for the period before and after the implementation of Problem List auto-save. The vast majority of problems were coded to the problem list at the time immediately before or after the admission event.

A histogram illustrating the timing of problem list diagnosis entry in the first 24 hours of admission is shown with a comparison between the four months preceding and the four months following the activation of Problem List auto-save (Figure 4).

The highest frequency of problem list entry events occurred in the time *before* patients were officially registered as admitted; that registration time correlates with the arrival of patients on the medical ward. In the time before this registration, providers are preparing for patients' arrival on the wards by entering pending admission orders, writing admission notes, and billing on admission documentation. The increase in coded problems after the activation of auto-save is apparent in the plot.

The proportion of patients with at least one problem in the problem list within 24 hours of admission increased from 94% before auto-save to 98% after auto-save activation (chi square 344, p value $2x10^{-16}$).

To assess whether the increased frequency of problem list entry events was sustained throughout admission, all problem list entry timestamps were normalized to length of stay for their respective admissions. The following histogram illustrates that the impact of health issue auto-save is maintained throughout the admission via subsequent billing and documentation events and not solely on the day of admission itself (Figure 5).

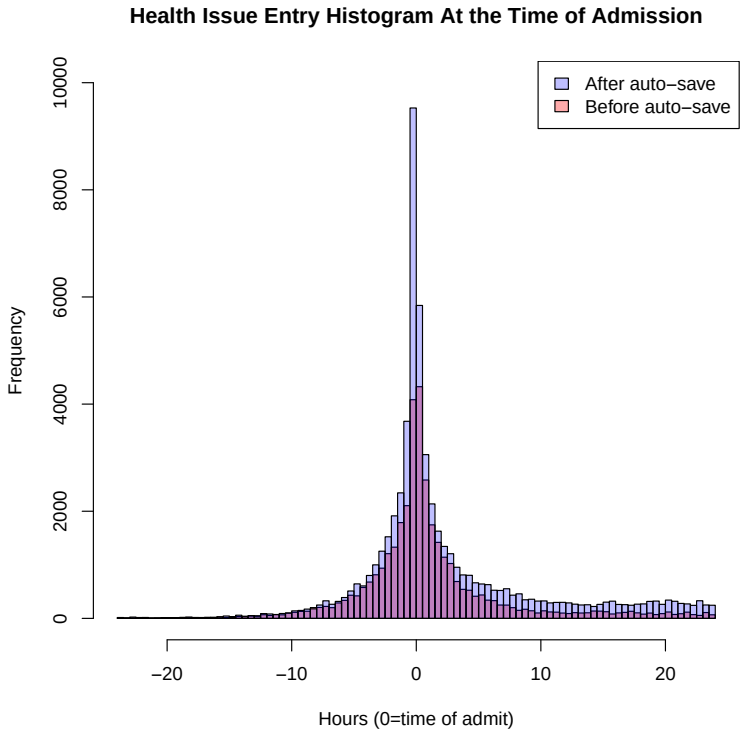


Figure 4. Histogram comparing problem list entry in the first twenty-four hours of admission, comparison between the four months preceding and following the activation of Problem List auto-save

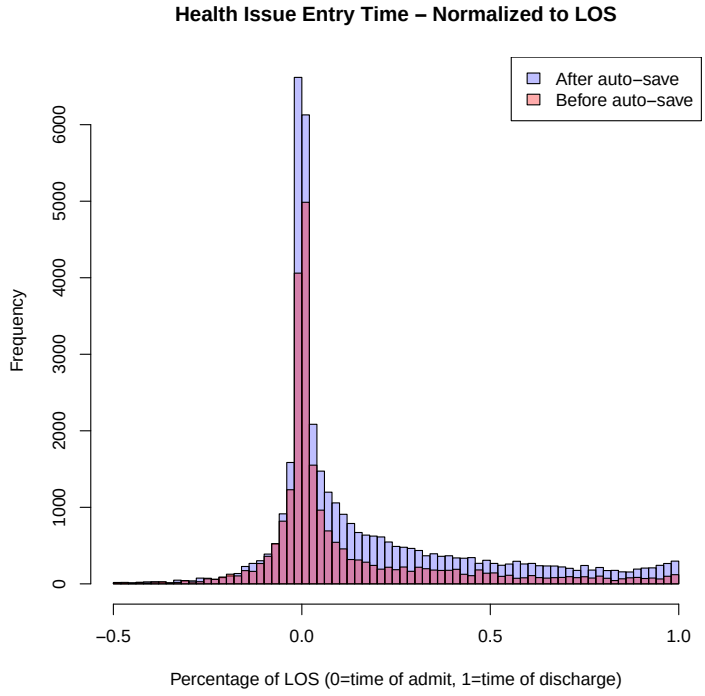


Figure 5. Histogram showing time of problem list diagnosis entry normalized to percentage of length of stay (LOS) - time of admission is 0% and time of discharge is 100% or 1.

Change in problem list entry for patients never before admitted to the medical center

The number of problems on the problem list for new patients increased to 3.8 problems from 2.9 problems on the problem list in the period after activation of auto-save compared to before; this result was statistically significant, p value $2x10^{-16}$. A similar result was seen for all patients in the study period, but particular emphasis is placed on those patients who would not have had problems in the problem list or prior administrative claims.

Content of problem list codes

In the period following the activation of Problem List auto-save, there was a marked increase in the number of V codes being captured (Figure 6). Some of the largest differences are seen in V45 and V42 (corresponding to organ transplantation), V30 (single liveborn), V27 (outcome of delivery), V22 (normal pregnancy). The quantity of V code counts before and after auto-save activation were different (chi square 920, p value 0.002).

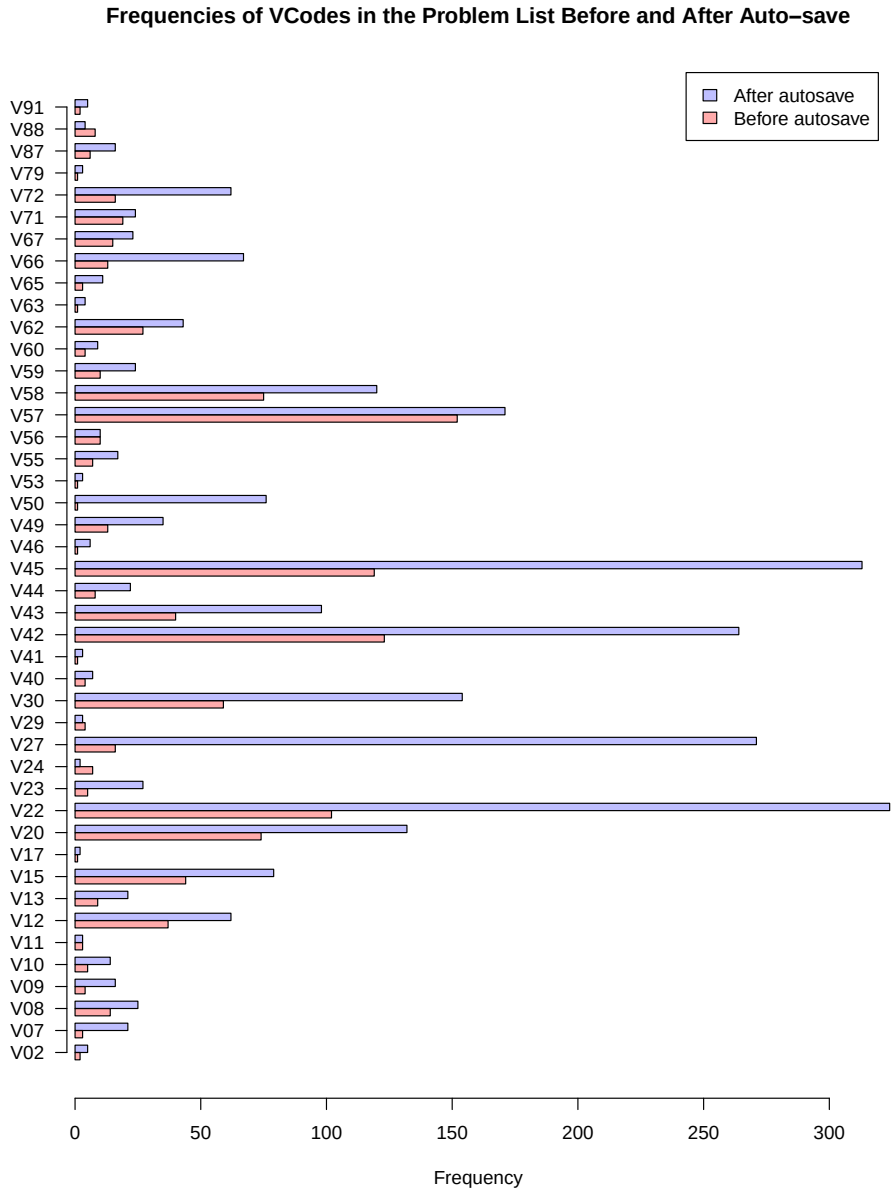


Figure 6. Barplot showing frequency of ICD9 “V code” entry before and after auto-save

Comparison of Administrative Claims Data Compared to Problem List

Jaccard similarity was calculated for Deyo/Charlson comorbidities derived from claims compared to those derived from the problem list before and after auto-save. The similarity of Deyo/Charlson comorbidities was higher for problems entered after auto-save activation compared to the period before, Jaccard similarity of 0.3 compared to 0.21, respectively, p value $2x10^{-16}$.

Discussion

The primary findings of this work demonstrate that interventions to automatically populate problem lists through extant admitting and billing workflows can increase the number of problems codified to those lists. This effect was achieved without altering or interrupting provider workflows. The majority of problems coded through this method are obtained at the time of admission, not the time of discharge, making them available for use while the patient is still admitted to the hospital. A number of decision support tools, such as drug-condition alerts, are enabled by this finding. Another important finding relates to the content of automatically populated problems on the problem list. V codes, a class of ICD9 codes connoting social, mental health, and even economic factors in a patient's clinical history, were captured at a higher rate through Problem List auto-save. Finally, a comorbidity index was applied to both administrative claims and to problem list ICD9 codes. While similarity was low overall, it increased between claims and problem list codes with the introduction of auto-save.

This study extends the work of others by demonstrating the impact of problem list interventions that are integrated mid-stream with existing provider workflows and without development of or validation of methods in natural language processing or machine learning. Collecting problem list entries through admit orders and through billing workflows also permitted other members of the care team to contribute to the problem list itself. While physicians contributed the vast majority of problems, the potential remains for a multi-disciplinary problem list incorporating biopsychosocial determinants of health. The effect on V codes of this intervention is one small step in this direction.

Limitations of this study include the length of time that auto-save was active compared to the overall study period. Generalizability is limited by differences in clinical and billing workflows across different sites. Deyo/Charlson comorbidity indices were not validated on problem list data, so the results of similarity measurements should be considered in this light. The effect of concomitant educational interventions within departments or specific units were not recorded and may confound these results; the relatively short length of time comparing the periods before and after auto-save have the secondary benefit of minimizing such effects. Frequency of problem entry via Health Issues, outside of the iCharge and admitting order workflows, could not be measured precisely in this study although the overall contribution was felt to be small. It is also worth noting that the ability to enter problems via Health Issues was not modified during the study time period so any baseline rate of use for some providers would be reflected in the data for the pre- and post-intervention periods.

Future research should compare the integration of problem list data along with and in place of administrative claims data. Predictive models built on problem list data, for example, would need to be validated separately and not simply used in place of claims data. The underlying workflows behind administrative claims and problems lists are not the same. The effect of Problem List auto-save should be evaluated to determine if it is sustained over time. Also, the impact of problem list evolution over the course of an admission and over the course of a patient's longitudinal care can be examined in subsequent work. Finally, as problem list capture improves, issues seen in administrative claims could also arise. "Problem creep" or "problem clutter" may be unintended consequences of these efforts.

Conclusion

A workflow-sensitive, two-fold intervention to increase capture of the patient problem list resulted in a statistically significant increase in recorded problems. These problems were recorded even before patients have left the emergency room to arrive on the hospital ward. Decision support integrating coded diagnostic data is enabled early in admission for all patients, not only those who have been admitted before. Problems recorded in this way improved in both quantity and content.

References

1. Weed LL. Medical records that guide and teach. The New England journal of medicine. 1968;278:593-600.
2. Weed LL. The problem oriented record as a basic tool in medical education, patient care and clinical research. Annals of clinical research. 1971;3:131-4.
3. (WHO) WHO. International Classification of Diseases (ICD) 2014 [cited 2014 3/1/2014]. Available from: http://www.who.int/classifications/icd/en/.
4. Gonzalez CJ, Rivera CA, Martin RJ, Mergian GA, Cruz H, Agins BD. Using computer-based monitoring and intervention to prevent harmful combinations of antiretroviral drugs in the New York State AIDS Drug Assistance Program. Joint Commission journal on quality and patient safety / Joint Commission Resources. 2012;38(6):269-76.
5. Stockl KM, Le L, Harada AS, Zhang S. Use of controller medications in patients initiated on a long-acting beta2-adrenergic agonist before and after safety alerts. American journal of health-system pharmacy : AJHP : official journal of the American Society of Health-System Pharmacists. 2008;65(16):1533-8.
6. Noirot LA, Reichley R, Dunagan WC, Bailey TC. Using outpatient prescription claims to evaluate medication adherence in an acute myocardial infarction population. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2005:1062.
7. Floor-Schreudering A, Heringa M, Buurma H, Bouvy ML, De Smet PA. Missed drug therapy alerts as a consequence of incomplete electronic patient records in Dutch community pharmacies. The Annals of pharmacotherapy. 2013;47(10):1272-9.
8. Ojeleye O, Avery A, Gupta V, Boyd M. The evidence for the effectiveness of safety alerts in electronic patient medication record systems at the point of pharmacy order entry: a systematic review. BMC medical informatics and decision making. 2013;13:69.
9. McCarthy EP, Iezzoni LI, Davis RB, Palmer RH, Cahalane M, Hamel MB, et al. Does clinical evidence support ICD-9-CM diagnosis coding of complications? Medical care. 2000;38(8):868-76.
10. Weingart SN, Iezzoni LI, Davis RB, Palmer RH, Cahalane M, Hamel MB, et al. Use of administrative data to find substandard care: validation of the complications screening program. Medical care. 2000;38(8):796-806.
11. Iezzoni LI, Daley J, Heeren T, Foley SM, Hughes JS, Fisher ES, et al. Using administrative data to screen hospitals for high complication rates. Inquiry : a journal of medical care organization, provision and financing. 1994;31(1):40-55.
12. Carter GM, Newhouse JP, Relles DA. How much change in the case mix index is DRG creep? Journal of health economics. 1990;9(4):411-28.
13. Seiber EE. Physician code creep: evidence in Medicaid and State Employee Health Insurance billing. Health care financing review. 2007;28(4):83-93.
14. Hsia DC, Krushat WM, Fagan AB, Tebbutt JA, Kusserow RP. Accuracy of diagnostic coding for Medicare patients under the prospective-payment system. The New England journal of medicine. 1988;318(6):352-5.
15. Services C-CfMaM. Meaningful Use 2014. Available from: http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/Meaningful_Use.html.
16. Holmes C. The problem list beyond meaningful use. Part I: The problems with problem lists. Journal of AHIMA / American Health Information Management Association. 2011;82(2):30-3; quiz 4.
17. Hartung DM, Hunt J, Siemienczuk J, Miller H, Touchette DR. Clinical implications of an accurate problem list on heart failure treatment. Journal of general internal medicine. 2005;20(2):143-7.
18. Banerjee ES, Gambler A, Fogleman C. Adding obesity to the problem list increases the rate of providers addressing obesity. Family medicine. 2013;45:629-33.
19. Warren JJ, Collins J, Sorrentino C, Campbell JR. Just-in-time coding of the problem list in a clinical environment. Proceedings / AMIA Annual Symposium AMIA Symposium. 1998:280-4.
20. Meystre S, Haug PJ. Automation of a problem list using natural language processing. BMC medical informatics and decision making. 2005;5:30.
21. Meystre SM, Haug PJ. Comparing natural language processing tools to extract medical problems from narrative text. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2005:525-9.
22. Meystre S, Haug PJ. Natural language processing to extract medical problems from electronic clinical documents: performance evaluation. Journal of biomedical informatics. 2006;39:589-99.
23. Chen ES, Wright A, Maloney FL, Van Putten C, Paterno MD, Goldberg HS. Enhancing clinical problem lists through data mining and natural language processing. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2008:901.

24. Solti I, Aaronson B, Fletcher G, Solti M, Gennari JH, Cooper M, et al. Building an automated problem list based on natural language processing: lessons learned in the early phase of development. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2008:687-91.
25. Wright A, Chen ES, Maloney FL. An automated technique for identifying associations between medications, laboratory results and problems. Journal of biomedical informatics. 2010;43:891-901.
26. Pacheco JA, Thompson W, Kho A. Automatically detecting problem list omissions of type 2 diabetes cases using electronic medical records. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2011;2011:1062-9.
27. Wright A, Pang J, Feblowitz JC, Maloney FL, Wilcox AR, McLoughlin KS, et al. Improving completeness of electronic problem lists through clinical decision support: a randomized, controlled trial. Journal of the American Medical Informatics Association : JAMIA. 2012;19:555-61.
28. Mehta N, Vakharia N, Wright A. EHRs in a Web 2.0 World: Time to Embrace a Problem-List Wiki. Journal of general internal medicine. 2013;29:434-6.
29. Mowery DL, Jordan P, Wiebe J, Harkema H, Dowling J, Chapman WW. Semantic annotation of clinical events for generating a problem list. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2013;2013:1032-41.
30. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. Journal of chronic diseases. 1987;40(5):373-83.
31. Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. Journal of clinical epidemiology. 1992;45(6):613-9.
32. Kottke TE, Baechler CJ. An algorithm that identifies coronary and heart failure events in the electronic health record. Preventing chronic disease. 2013;10:E29.
33. RStudio. RStudio. 0.98.501 ed2012.
34. Wickham H. The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software. 2011;40(1):1-29.
35. M Dowle TS, S Lianoglou, A Srinivasan. data.table: Extension of data.frame. 1.9.2 ed2014.
36. Jaccard P. Nouvelles recherches sur la distribution florale. Bull Soc Vaudoise Sci Nat. 1908(44):223-70.

A Mobile/Web App for Long Distance Caregivers of Older Adults: Functional Requirements and Design Implications from a User Centered Design Process

Steven S. Williamson, BS[1], Paul N. Gorman, MD[1], Holly B. Jimison, PhD[2]
[1]Oregon Health & Science University, Portland, OR; [2]Northeastern University, Boston, MA

Abstract

Recent trends of population aging and globalization have required an increasing number of individuals to act as long distance caregivers (LDCs) to aging family members. Information technology solutions may ease the burden placed on LDCs by providing remote monitoring, easier access to information and enhanced communication. While some technology tools have been introduced, the information and technology needs of LDCs in particular are not well understood. Consequently, a needs assessment was performed by using video conferencing software to conduct semi-structured interviews with 10 LDCs. Interviews were enriched through the use of stimulus materials that included the demonstration of a prototype LDC health management web/mobile app. Responses were recorded, transcribed and then analyzed. Subjects indicated that information regarding medication regimens and adherence, calendaring, and cognitive health were most needed. Participants also described needs for video calling, activity data regarding sleep and physical exercise, asynchronous communication, photo sharing, journaling, access to online health resources, real-time monitoring, an overall summary of health, and feedback/suggestions to help them improve as caregivers. In addition, all respondents estimated their usage of a LDC health management website would be at least once per week, with half indicating a desire to access the website from a smartphone. These findings are being used to inform the design of a LDC health management website to promote the meaningful involvement of distant family members in the care of older adults.

Introduction

2011 marked a critical milestone for Americans; the first set of baby boomers reached the age of retirement. For the first time in the United States' history, the number of adults age 65 and older exceeded the number of children under the age of 5. By 2014, the percentage of the population age 65 and older will reach an all-time high of 14%; double the proportion that was seen in the 1940's [1]. As this process of population aging unfolds, the problems associated with caring for unprecedented numbers of older adults become increasingly apparent. Unparalleled demand will be placed not only upon the US healthcare system, but also, upon the millions of family members, friends, and neighbors that provide unpaid care to elderly loved ones. These individuals, often referred to as informal caregivers[2], form "the backbone for much of the care that is received by older adults in the United States"[3]. In an increasingly global society, geographic separation presents a significant challenge to many as they strive to provide care from afar. Challenges such as inadequate methods of communication, living in different time zones, and lack of familiarity with a loved one's surroundings may all combine to prevent a long distance loved one from providing care[4]. Such separation often increases the burdens of time, cost, and emotional strain upon the caregiver[5]. In the last few years, it has been suggested [6,7]that internet technologies have matured insomuch that they may prove to be viable options for providing support to long distance caregivers (LDCs). Such an approach however, remains understudied.

Some early research has focused on identifying information needs of caregivers for individuals with dementia [8,9], while others have focused on providing appropriate information regarding how to care for other specific illnesses/conditions[10,11]. One promising study, conducted by the National Alliance of Caregiving and United Healthcare [12], investigated caregivers ranking of various health IT tools to support them in their care. This study, measured perceived benefits and barriers of 12 technologies for both in-home and out-of-home caregivers. Systems that allowed for personal health record tracking, caregiving coordination and medication support had high levels of perceived benefits and lower levels of perceived barriers. We suggest that these and other caregiving tools may be especially useful in the context of a smart home in which older adults are monitored using unobtrusive sensors to track various health metrics.

Recently, work[13], has been undertaken to provide a better understanding of the prevalence of technology use by out-of-home caregivers in the United States. Current estimates indicate that about one third of out-of-home caregivers use health IT in their caregiving activities. An interesting contrast is found however, in that even among "technology nonusers", over 70% of LDCs expressed an interest in using technology in their caregiving

responsibilities. The incongruence between interest to use health IT tools and actual usage may be explained by barriers such as perceived cost, potential resistance by the care recipient[12], and a lack of user-centered focus in the design and implementation of current LDC systems[13].

In the interest of promoting higher levels of usage and utility, we resolved to use a user-centered-design approach to assess information needs and discern important usability principles in the design and development of health IT tools for LDCs. The research outlined below is innovative, due to the fact that no studies have specifically looked at the information needs and technology preferences of LDCs by providing caregivers an opportunity to openly discuss their needs and preferences. Furthermore, our study is unique in that we are investigating information needs in the context of a smart home, containing multiple sensors that provide important data streams about activity, cognition, and physiologic parameters. Investigating LDC needs from this perspective provides us with additional information that will enable us to better understand the emerging needs of caregivers living in an increasingly "electronic" world.

Methods

Due to the exploratory nature of this research, qualitative methods were chosen. Since our subject recruitment pool contained individuals throughout the United States, we chose to conduct semi-structured interviews via Skype as the primary method of data collection. We chose Skype over a traditional telephone as we felt that the face-to-face interaction would help subjects to feel more at ease when talking to an unfamiliar person. The use of Skype also helped us to detect any visual cues that may not have been as apparent via a phone call and allowed us to visually present questions and stimulus materials to subjects as we spoke with them. Skype also served as an ideal platform for data collection due to the fact that all communications are encrypted using robust encryption algorithms.

Prototype Development

In keeping with other qualitative research, [14,15] we elected to develop basic prototypes of a caregiver web (shown in Figures 1 & 2) and mobile app as stimulus materials to facilitate and further enrich our discussions. This approach was chosen due to the limited availability of remote caregiving systems and the anticipated lack of familiarity with the types of data that may be collected in a smart home environment. In an attempt to intelligently develop an initial prototype that delivered an optimal user experience, we drew upon the following four data sources for guidance:

1. Well established usability principles from the human computer interaction literature
2. Scientific articles that specifically described caregiving systems/prototypes
3. Existing commercial systems designed to be used by caregivers
4. Usability experts within our institution

Based upon our review of the literature and existing systems, we begun development of an initial mockup using

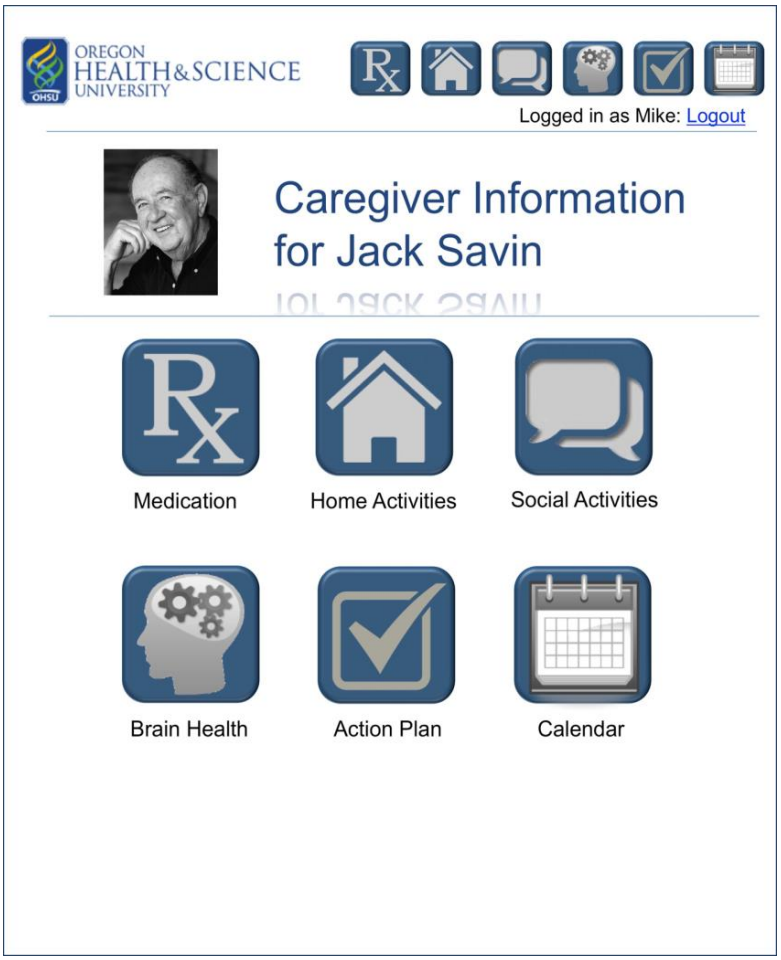


Figure 1: Screen shot showing splash page from prototype caregiver web app

Microsoft PowerPoint. During this process, usability experts within our institution were also consulted. To provide caregivers with a better understanding of system functions, we opted to make parts of the mockup interactive. This was accomplished using transitions and animations within the PowerPoint slideshow and allowed us to demonstrate behaviors and actions when we clicked on various elements within the user interface.

Interview Guide Development

Careful thought and attention was given to development of an interview guide[16]. Our guide was developed with the primary goal of facilitating open conversation with the caregivers in our study. As such, we elected to start by asking very broad, open-ended questions which were then followed up with more focused questions and probes when greater clarification was needed. Great care was taken to ensure that questions would be easily understood and would not lead study participants towards a specific idea or thought, but rather to allow them to express their own thoughts freely. Due to the fact that interviews were conducted via Skype, we also elected to display each question on the participants' screens as they were asked. This allowed each subject to both hear each question audibly as well as see the questions visually. It was our hope that doing so would help to ensure that each question was better understood and would allow respondents to re-read the question as they formulated their response.

After an initial draft of the interview guide was completed, two scholars were asked to review each question for simplicity, readability and neutrality and to suggest changes when necessary. As a final step in development, the interview guide was then used to conduct two mock interviews. This process not only allowed for minor changes to the guide but also helped the research team to practice interviewing techniques before the start of formal data collection.

Study Setting

A network of older adults living in smart homes throughout the Portland, Oregon region has been established as part of our existing cognitive health coaching platform[17]. Older adults that participate in this health coaching platform are continuously monitored using various health tracking sensors. Areas of study include:

- Medication adherence and reminding - measured by a camera embedded pillbox
- Socialization - measured by phone, Skype and email monitors
- Sleep quality - measured using mattress pressure sensors
- Cognitive health - measured by cognitive computer games

Because the socialization module encourages the use of the telephone, email and Skype video calling, each older adult in the socialization intervention had previously chosen a remote partner with whom they would regularly communicate. These individuals in turn agreed to provide remote support to the older adults in our project. All remote partners lived in a location different than the older adult and were generally a close friend or family member. A group of 11 subjects was recruited from within this pool of remote partners as participants in our needs assessment.

Data Collection

Each participant was contacted initially via telephone and then later interviewed for approximately 45 minutes. Due to geographic separation between the subjects and the researchers and as all enrolled LDCs were familiar users of the Skype video conferencing software, interviews were conducted remotely through the use of this system. Initially, a short introduction was given in which the purpose of the study and each subject's role was clearly explained. An emphasis was placed on the fact that subjects could ask questions or make suggestions at any time. Next, subjects were asked to introduce themselves and to describe some of the challenges that they had encountered as they strived to provide care from a distance. Respondents were then asked which types of information are most important to them as caregivers. Next, subjects were asked to identify ways in which technology might serve to ease some of the burdens encountered by LDCs.

After respondents answered these questions, stimulus materials including sample screen-shots for web and smartphone based health management were displayed using Skype's screen sharing feature. These materials enriched discussion and provided subjects with a real world example of ways in which technology could help them to provide care. In particular, subjects were shown a prototype website in which sensor data regarding medication adherence, socialization, calendaring, sleep quality, and cognitive health was presented using easily understood language and graphics. Tips and suggestions for how to help the older adult were also displayed. After presenting the prototype, discussion was facilitated by the presentation of thoughtful questions designed to promote feedback about key areas of interest (*e.g.* estimates regarding level of usage, importance of mobile devices, design recommendations). Finally, each subject was asked for any additional comments or suggestions regarding site design and types of information available. Each interview was recorded, transcribed, and subsequently analyzed by grouping similar thoughts and concepts into appropriate themes and ideas. The findings of our needs assessment will then be used to inform the development of version 2.0 of our prototype. This new and improved prototype will then be used in a usability study investigating the way that real world users interact with the proposed application.

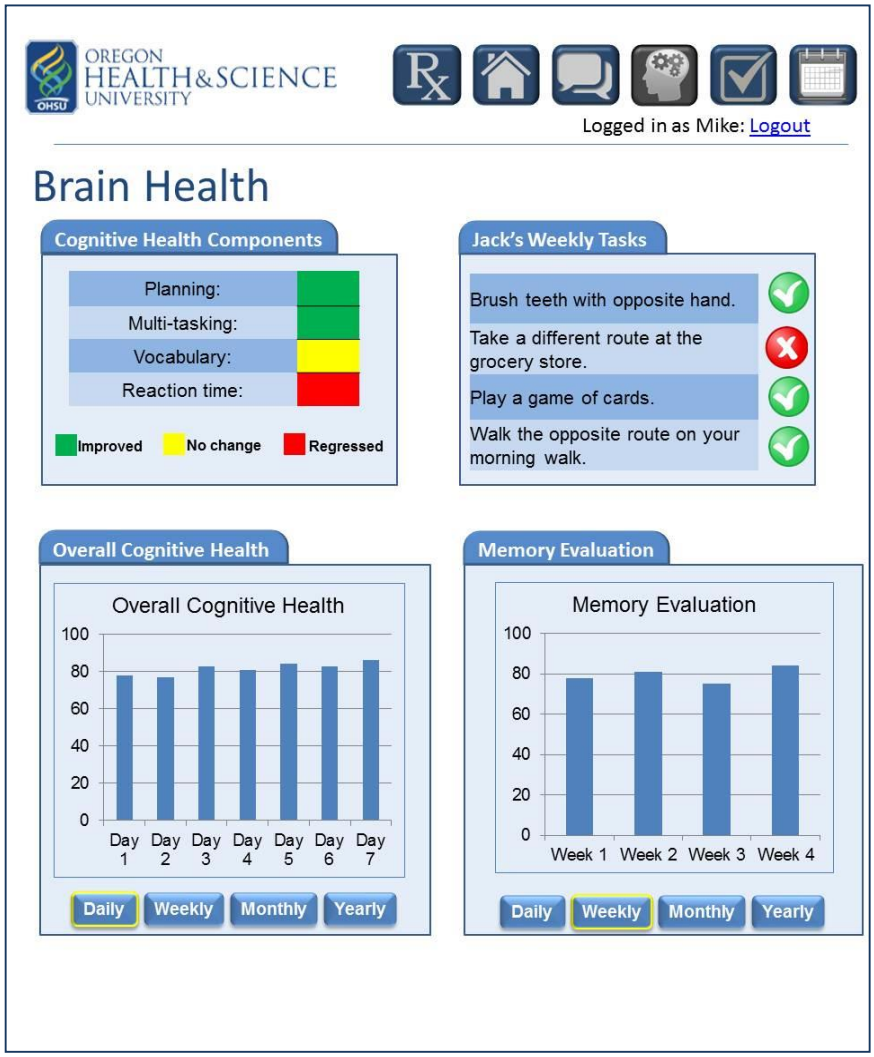


Figure 2: Screen shot showing the brain health page from prototype caregiver web app

Results

Of the 11 subjects that were initially recruited, 10 individuals were successfully contacted and interviewed, with one participant unable to proceed due to lack of a sufficiently reliable internet connection. Of these 10 individuals, 6 were female and 4 were male.

Desired Functionality

Subjects in our study reported that LDCs desire 14 different basic functions: video calling, calendaring, medication tracking, cognitive health tracking, sleep tracking, physical exercise tracking, access to medical records, asynchronous communication, photo sharing, journaling, online health resources, real-time monitoring, an overall summary of wellness, and guidance/feedback regarding the care they provide. These 14 functions are described in Table 1.

Table 1: Fourteen Basic Functions for Long Distance Caregiving	
Video Calling	Nearly all of the individuals interviewed spoke about the benefits of using videoconferencing software such as Skype to communicate with the older adult under their care. Four individuals spoke of the value of nonverbal communication that is not available over a regular telephone call. The participants' thoughts regarding this matter are well summarized by the comment "Now, instead of hearing how she's doing, I can see how she's doing. It's one thing to tell someone how you're doing but it's a little harder to look at someone and tell them that you're feeling good when you're not." In addition, one interviewee talked about the benefits of being able to show objects over video rather than simply describing them. Some frustration was expressed that this was only a valuable form of communication when there were not technological barriers such as unreliable internet connections, audio dropouts or pixelated video.
Calendaring	Six individuals indicated that a shared calendar would be useful to them in their caregiving responsibilities. Respondents were especially interested in being able to view upcoming doctors' appointments and any planned trips or outings. One person commented that "to [her] the calendar would not be at all useful because digital calendars are cumbersome". Another commented that "it might take a bit of switching going from a paper calendar to an electronic one but I think I can convince my mom to switch". The comment was also made that it would be useful for older adults to see a very high level version of the caregiver's calendar so that the older adults could be reminded of times that the LDC would not be available and the care recipient would know to contact somebody else if a concern arose. The idea that the calendar could be used to coordinate care by multiple LDCs was also mentioned. This would allow multiple individuals to share the responsibilities of caregiving rather than a single individual being expected to carry the burden for the majority of care.
Medication Tracking	Four individuals indicated that information regarding medication adherence was very important to them. In addition, the importance of knowing the older adult's medication list and regimen was also mentioned. Respondents made comments such as "medications are a big concern" and "if you're not taking your medication, everything else would fall apart". One individual, however, said that medication information was the least important of all the types of information presented. He commented that this was due to the fact that using images obtained from a camera embedded inside a pill box did not really indicate if the medication had actually been taken. In his words "they could take it out of the box but then not really take it".
Cognitive Health Tracking	Four interviewees suggested that data regarding cognitive health was very important to them. Two of these individuals indicated that this information would be especially interesting to them if it could be presented over a long period of time allowing the caregiver to track any problems. In the words of one subject, "as he gets older, I especially worry about his brain and memory".
Sleep Tracking	Three respondents spoke of the importance of knowing if and when an older adult was experiencing difficulty sleeping. Each of them expressed concern that inadequate sleep can then lead a large number to other problems/concerns. One interviewee described the utility of a system that would automatically alert her after her loved one had experienced multiple consecutive nights of poor sleep so that she could call and check on the older adult and then intervene if necessary.
Physical Exercise Tracking	The importance of knowing whether or not an older adult is regularly exercising was also mentioned. Caregivers wanted to know that the older adult in their care was able to regularly exercise. Along with this information need one caregiver also mentioned the importance of knowing certain metrics of physical ability such as strength and balance.
Medical Records Access	Two LDCs asked about the possibility of being able to access the older adult's medical information and test results. They expressed a desire to be more informed and involved in the older adult's medical care because "sometimes if we don't go with him/her, then his/her story doesn't make sense".
Asynchronous Communication	Many of the LDC's interviewed spoke about the need and value that comes from asynchronous communication. Whether this communication was via e-mail, text message, instant message *etc.* seemed to be less important than the ability to communicate asynchronously. This was important to

Asynchronous Communication (continued)	them because it allowed them to communicate with the older adults without having to worry about the time of day (*e.g.* too early, too late, while the older adult was busy). One respondent also said that this form of communication allowed the care recipient to communicate with him without worrying about disrupting him at work.
Photo Sharing	The ability to share photographs was mentioned in a few different interviews as an important form of communication. Two LDCs spoke about the value of being able to send pictures back and forth. These individuals mentioned that seeing pictures helped to bridge the gap between caregiver and care recipient and made them feel more involved in each other's lives.
Journaling	Two caregivers spoke about some sort of electronic caregiving journal that would allow for note taking and could be used to keep track of items that may not be included within the caregiving application. One suggested that the journal could be tied to a calendar so that reminders could also be integrated within the journaling feature.
Online Health Resources	Multiple caregivers talked about the importance of being able to access reliable health information electronically. Caregivers described medical websites as an important resource that they could use to research a specific condition or illness and then share the pertinent information with the older adult.
Real-time Monitor	The need for a real-time indication of an older adults status was described throughout our conversations with LDCs. Caregivers were especially interested to know if an older adult had fallen or was in immediate need of help. Conversely, caregivers also wanted to know when the older adult was doing well and no intervention was needed on their part. One caregiver talked about a system that could not only communicate when their help was needed but also "how badly [the care recipient] needed help".
Summary Metric of Overall Wellness	While many caregivers saw the value in providing data regarding individual items (*e.g.* medication, sleep, *etc.*) they also expressed a need for a summary metric that could be an overall indicator of wellness. This would allow caregivers the ability to look at a single graph and see a general trend of wellness over time.
Feedback /Guidance	Over half of the respondents talked about the importance of providing guidance and feedback. Not only is it important to provide monitoring data to caregivers, but it is essential to also provide suggestions of what they as the caregiver can do to provide better care and encourage healthy behaviors by the older adult. One caregiver also spoke about the importance of providing encouragement to caregivers when they logged into the system and tried to play a more active role.
Other Suggestions	Some respondents also suggested other types of information that would be useful to them as long distance caregivers. One caregiver suggested the inclusion of "information about hobbies and interests". He went on to suggest a page in which the older adult could share pictures and information regarding hobbies with the caregiver. Another caregiver was interested in the possibility of including information regarding diet through the use of a "smart refrigerator to track if she needs milk and that sort of thing". One final suggestion was the ability to send an alert to the older adult. He commented that "Dad hasn't been drinking enough water lately. It would be really nice if there was some way to remind him with a beep or something."

Design Implications

In addition to describing desired functions of an LDC web/mobile app, study participants also shared insights that have important design implications for those seeking to develop such a system. These design implications concern usage patterns, device preferences, data sharing preferences, and the presentation of longitudinal tracking data, described further in Table 2.

Table 2. Design Implications for a Mobile/Web App for Long Distance Caregivers.	
Usage Patterns	Every individual interviewed expressed optimism about their usage of the proposed system and felt that they would use it on a fairly regular basis. All participants indicated that they would likely use the system at least once per week with three participants indicating that they thought they would use the system "a couple times per week" and two respondents suggesting that they would use the system on a daily basis. Two individuals indicated that they would be much more willing to use the system regularly if "the system had the ability to alert me when there was something that needed my immediate attention". Some LDCs estimated that their usage would be heavily tied to the health of the older adult under their care. They suggest that when the older adult was healthy they would be less likely to have any concerns and would not use the system as regularly. In contrast, they felt they would use the system much more frequently when the older adult's help was concerning to them. While not as valuable as actual usage data, these expected usage patterns provide valuable information regarding the overall flow and design of a caregiver website. Such high frequency of usage would suggest the need to design a dashboard that would allow the caregiver to quickly check an older adults condition without the need to click on each individual category. Also, as noted by two of our participants, an intelligent alerting system that drew the caregiver's attention to potentially worrisome data would be ideal. If alerts are to be used however, the authors urge that a great deal of care be taken so as to not inundate caregivers with false alarms as this is likely to lead to alert fatigue.
Device Preferences	Of our sample, half of the respondents indicated that they would be likely to access the LDC website from a smartphone. This closely mirrors smartphone adoption data for the US population during the time that the interviews were performed. As such, we expect an increasing proportion of caregivers to request smartphone compatibility for a caregiver website. Of those that desired smartphone compatibility, many talked of the convenience and importance of having access to the system while traveling either to/from work or while on vacation. These participants described use cases in which a smartphone would be used while on the go but a traditional PC would still be the preferred choice if available (i.e. when at home). Such usage in which both a smartphone and a traditional PC are used interchangeably requires a consistent look and feel, as well as similar functionalities and feature sets regardless of which device is used to access the site. In addition, due to respondents reporting high levels of expected usage, a mobile app is recommended in lieu of a smartphone compatible website. Such an approach allows caregivers to view historical data even when no data connection is available and allows for more sophisticated alerts to be displayed when necessary. Identified barriers to using a smartphone to access the LDC website were a small screen and relatively high costs of ownership and usage. However, we expect these concerns to fade somewhat as smartphone manufacturers/providers continue to shift towards larger screen sizes and lower cost devices/services.
Data Sharing Preferences	A few caregivers expressed concern that due to the sensitive nature of health data, their older adult may not be willing to share all of the different types of information with them. Though this was not confirmed by discussing data sharing preferences with older adults in our study, we suggest that any such system provides a way in which older adults are able to control the visibility of the data collected. It was also suggested that older adults may be more willing to share monitoring data if the system is implemented before they are facing serious health challenges. In the words of one subject "It might be better to start them when they don't need it because if you start too late then they may not want to do it. I guess it's one of those things that if they sense that they aren't doing well then they will resist that." As such, we recommend the early implementation of health monitoring systems as a possible way to mitigate this challenge. Such an approach also has the added benefit of collecting longitudinal data while an individual is still healthy so that there is a greater likelihood of early detection when problems arise.
Longitudinal Tracking	Caregivers reacted favorably to the idea that longitudinal data could be presented in a meaningful way. They recognized the value of being able to look at data over different periods of time to identify potential areas of concern. As one subject put it "looking by week or month that data is very useful. What I keep trying to get in my head is the progression …seems like this is a great tool to know their decline". While we agree that there is likely value in providing longitudinal data to caregivers, we also expressed concern about the possibility of either misinterpretation or over interpretation of these data. Contributing to this is the fact that many data streams from current smart home environments are fairly noisy. These concerns may be mitigated somewhat through the use of data smoothing algorithms and clear indications to the caregiver when specific scores are vastly different from an individual's baseline.
Caregiving Terminology	Very few individuals that we interviewed identified themselves as either a long distance caregiver or a caregiver in general. Many individuals expressed that they thought of a caregiver playing a more hands-on role that was not possible from a distance and instead viewed themselves as a helper, friend or family member. While we feel that it would be healthy to help redefine the lay definition of what constitutes a caregiver, we also recognize a need to properly frame any communication with LDCs in language that they understand and can relate to.

Overall Impressions

The overall reaction from caregivers was very positive with many making comments such as "I think this is a great idea" and "This is going to be really helpful for people like me". Though individuals suggested the improvements detailed above, none of the participants thought that building a web site/app for LDCs was a generally bad idea. In addition to the expected benefits of being able to ease the burden of providing care and improve involvement of LDCs, a few other benefits were suggested. One caregiver remarked that the system would "help [her] not feel so guilty for living so far away". It was also suggested that such a system would help older adults because "having us involved helps her to feel loved and valued". Even when an older adult already lives near family members, one individual suggested that "I can help my mom and uncle by alleviating some of their stress. If there's something going on I can let them know and have them go visit her". At the conclusion of one of the interviews, one caregiver became emotional as he spoke of the privilege of being able to care for his aging parents as they "experience this amazing process of the end of life" and suggested that a LDC website would allow him to do that more effectively.

Discussion

After conducting qualitative semi-structured interviews with 10 subjects, we identified 14 different functions that LDCs desire (video calling, calendaring, medication tracking, cognitive health tracking, sleep tracking, physical exercise tracking, access to medical records, asynchronous communication, photo sharing, journaling, online health resources, real-time monitoring, an overall summary of wellness, and guidance/feedback regarding the care they provide). We also identified 4 important design implications concerning LDC usage patterns, device preferences, data sharing preferences, and the presentation of longitudinal tracking data. Overall, we found that participants reacted very positively to the proposed system.

These results are concordant with previous studies that have investigated the role of technological solutions for caregivers. Our findings are similar to those of The National Alliance for Caregiving (NAC)[12] who also identified health record tracking, medication support tools, caregiving coordination tools, interactive systems for physical, mental and leisure activities, a symptom monitor and transmitter, a video phone system and a caregiving decision support tool as some of the most important tools for caregivers. Though described by our respondents using different terminology, many of the desired features identified in our study are functionally very similar. One feature listed in the NAC that was not reported by our study is the need for caregiver training simulations. While it is possible that this feature is also desired by LDCs, we note that the NAC study involved both long distance and in-home caregivers and this feature in particular may be more important to individuals serving as traditional “hands-on” caregivers. We also identified some desired features that have not been suggested previously and identified design implications that we believe are important for those looking to develop successful LDC systems.

During the planning of this study, there was some concern that presenting the prototype to interview participants may bias our findings. We expected that we might lead interviewees to talk about the information needs that we anticipated them to have rather than actual information needs. We were surprised to find that while some participants did not speak of some types of information until prompted, at least one participant spoke of each type of information need before the prototype was presented to them. This leads us to believe that our approach was indeed appropriate and the prototype served as a probe to elicit deeper understanding rather than serving to bias our respondents. This is reaffirmed by the high level of agreement between our study and previous work.

The overwhelmingly positive reaction towards our prototype system also follows trends found by other researchers[13]. Our results however, indicate an even higher level of acceptance with 100% of subjects expressing enthusiasm for an LDC system. While somewhat explained by the limitations listed below, we also suggest that such high levels of enthusiasm are the result of high levels of caregiver burden, with many caregivers desperately looking for assistance as they struggle to provide for loved ones.

Limitations

While we feel that the findings of this study are indeed useful, our choice of methodology and sample population created some important limitations that should be considered. These include:

1) Small Sample/Lack of Diversity - Though not as important due to our use of qualitative methods, our sample size was still very small (N = 10) and had limited inclusion of ethnic minorities. Also, the LDCs in our study only provided care to older adults that had displayed little to no cognitive impairment

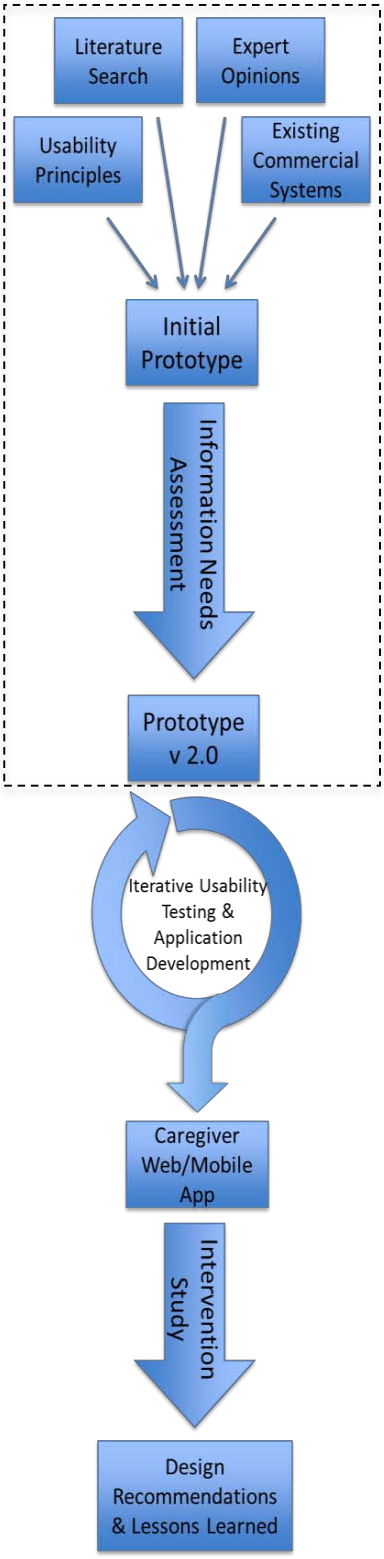


Figure 3: Long term research plan for development and evaluation LDC web/mobile app (dashed box designates work described in this paper)

2) Pro-technology - Study participants were drawn from a pool of technology using seniors and LDCs and are likely more receptive to new technologies than the general population.

3) Hesitant to criticize - Study participants may have been unwilling to provide a critical analysis of our prototype for fear of offending members of the research team.

4) Unable to Discern Needs - It is unclear if subjects are entirely aware of their own needs. As with many needs assessments, there is concern that individuals are unable to identify specific needs, choosing instead to be content with their currently available toolset. This may be especially true when discussing a new technology that subjects have not had the opportunity to use in the real world. During our interviews, this was evident when individuals responded that they were not sure whether or not they would need a particular type of information.

As such, we stress the importance of conducting future research to address these limitations by investigating what other information needs are required by other caregiving populations and by trying to answer our research questions with complementary methodologies.

Despite these limitations, many valuable themes emerged that we hope will prove useful as we strive to provide LDCs with new information technology tools. It was very encouraging to find that all 10 individuals interviewed suggested that building a LDC web/mobile app would be accepted positively. Equally encouraging were indications that the proposed application may be used on a regular basis. However, perceived usefulness and usage may not be accurate indicators of actual system usage and utility.

While this study has identified information types that are likely to be useful and valuable to individuals providing care from afar, the best methods for presenting these data to caregivers warrants further exploration. In addition, while the ability to access an older adult's medical information and test results has been suggested as a useful feature, limitations regarding privacy of medical data may prove to be substantial hurdles.

Future Research

The work described here is the initial step in a larger effort to better understand the role that heath IT can play in assisting caregivers. Our future research will use these findings to improve our existing prototype. As shown in Figure 3, these improvements, along with making the prototype fully interactive will allow us to enter an iterative usability testing and development phase during which users will be asked to use our prototype to perform various tasks. This iterative cycle of development will then allow us to create a final caregiver web/mobile app that is both useful and user-friendly. This out will then be evaluated by means of an intervention study. Both quantitative and qualitative methods will be used to determine the impact that our caregiver application makes in the real world. These data will then be synthesized into design recommendations and lessons learned for future researchers interested in the field of technology enhanced caregiving.

Conclusion

The information needs of long distance caregivers are extensive and may vary somewhat depending upon the health problems of the care recipient. LDCs described needs for video calling, calendaring, data regarding medication, sleep, physical exercise and cognitive health, asynchronous communication, photo sharing, journaling, access to online health resources, real-time monitoring, an overall summary of health, and feedback/suggestions to help them improve as caregivers. We feel confident that we have obtained sufficient preliminary data to justify the continued development of a long distance caregiver application with the final goal of conducting a field trial of such a system in the real world.

References

1. Administration on Aging: Department of Health & Human Services. Projected Future Growth of the Older Population [Internet]. Administration on Aging. [cited 2014 Jan 15]. Available from: http://www.aoa.gov/aoaroot/aging_statistics/future_growth/future_growth.aspx

2. Donelan K, Hill CA, Hoffman C, Scoles K, Feldman PH, Levine C, et al. Challenged To Care: Informal Caregivers In A Changing Health System. Health Aff (Millwood). 2002 Jul 1;21(4):222–31.

3. Retooling for an Aging America: Building the Health Care Workforce - Institute of Medicine [Internet]. [cited 2012 Nov 13]. Available from: http://www.iom.edu/Reports/2008/Retooling-for-an-Aging-America-Building-the-Health-Care-Workforce.aspx

4. Mazanec P, Daly BJ, Ferrell BR, Prince-Paul M. Lack of Communication and Control: Experiences of Distance Caregivers of Parents With Advanced Cancer. Oncol Nurs Forum. 2011 May;38(3):307–13.

5. Bevan JL, Sparks L. Communication in the context of long-distance family caregiving: An integrated review and practical applications. Patient Educ Couns. 2011 Oct;85(1):26–30.

6. Benefield LE, Beck C. Reducing the distance in distance-caregiving by technology innovation. Clin Interv Aging. 2007 Jun;2(2):267–72.

7. Rosland A-M, Heisler M, Janevic MR, Connell CM, Langa KM, Kerr EA, et al. Current and potential support for chronic disease management in the United States: the perspective of family and friends of chronically ill adults. Fam Syst Health J Collab Fam Healthc. 2013 Jun;31(2):119–31.

8. Czaja SJ, Rubert MP. Telecommunications Technology as an Aid to Family Caregivers of Persons With Dementia. Psychosom Med. 2002;64(3):469–76.

9. Topo P. Technology Studies to Meet the Needs of People With Dementia and Their Caregivers. J Appl Gerontol. 2009;28(1):5–37.

10. Koenig KN, Steiner V, Pierce LL. Information Needs of Family Caregivers of Persons With Cognitive Versus Physical Deficits. Gerontol Geriatr Educ. 2011;32(4):396–413.

11. Washington KT, Meadows SE, Elliott SG, Koopman RJ. Information needs of informal caregivers of older adults with chronic health conditions. Patient Educ Couns. 2011;83(1):37–44.

12. National Alliance for Caregiving. e‐Connected Family Caregiver: Bringing Caregiving into the 21st Century. 2011;

13. Zulman DM, Piette JD, Jenchura EC, Asch SM, Rosland A-M. Facilitating Out-of-Home Caregiving Through Health Information Technology: Survey of Informal Caregivers' Current Practices, Interests, and Perceived Barriers. J Med Internet Res [Internet]. 2013 Jul 10 [cited 2014 Mar 11];15(7). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3713893/

14. Jimison HB, Sher PP, Appleyard R, LeVernois Y. The Use of Multimedia in the Informed Consent Process. J Am Med Inform Assoc. 1998 May 1;5(3):245–56.

15. Crilly N, Blackwell AF, Clarkson PJ. Graphic elicitation: using research diagrams as interview stimuli. Qual Res. 2006;6(3):341–66.

16. Boyce C, Neale P. Conducting in-depth interviews: A guide for designing and conducting in-depth interviews for evaluation input. 2006.

17. Jimison HB, Pavel M. Integrating computer-based health coaching into elder home care. Technology and aging: Selected papers from the 2007 International Conference on Technology and Aging. 2008. p. 122–9.

Does Sustained Participation in an Online Health Community Affect Sentiment?

Shaodian Zhang[1], Erin Bantum, PhD[2], Jason Owen, PhD, MPH[3], Noémie Elhadad, PhD[1]
[1]Columbia University, New York, NY; [2]University of Hawai'i Cancer Center, Honolulu, HI; [3]VA Palo Alto Health Care System, Menlo Park, CA

Abstract

A large number of patients rely on online health communities to exchange information and psychosocial support with their peers. Examining participation in a community and its impact on members' behaviors and attitudes is one of the key open research questions in the field of study of online health communities. In this paper, we focus on a large public breast cancer community and conduct sentiment analysis on all its posts. We investigate the impact of different factors on post sentiment, such as time since joining the community, posting activity, age of members, and cancer stage of members. We find that there is a significant increase in sentiment of posts through time, with different patterns of sentiment trends for initial posts in threads and reply posts. Factors each play a role; for instance stage-IV members form a particular sub-community with patterns of sentiment and usage distinct from others members.

Introduction

Online health communities, such as forums, blogs, and health-related Facebook or Yahoo groups, have become popular places for patients to exchange information with and seek psychosocial support from their peers[1–3]. Content analysis of online health communities has shown that there are primarily two types of support members provide to each other: informational and emotional support[4–7]. For patients with chronic or life-threatening diseases, there is evidence that psychological distress and anxiety related to medical decision making process as well as daily coping with the disease can be alleviated through the emotional support obtained in a community[8]. Like for other conditions, patients with breast cancer as well as caregivers of patients, rely on cancer-specific online health communities for both informational and emotional support[4,9–11]. Observational studies of breast cancer communities based on analysis of questionnaires and surveys of their members have indicated a positive association between a member's community participation and emotions such as empathy and satisfaction[10,12–15]. Meanwhile, content analysis of online patient-authored text has provided new perspectives on the health impact of online social networking[16,17], but such analysis usually requires manual annotations which can be costly when the contents are in large scale. As such, automated solutions that can be leveraged to study outcomes of online participations are needed.

More recently, automatic sentiment classification methods have been exploited to investigate sentiments of forum posts published by patients. For instance, studies found that thread originators change their sentiment in a positive direction through reviewing others' replies and self-replying[18], and such changes are largely resulting from postings of influential users[19]. The studies also find that within threads, sentiment changes are correlated with several factors such as number of self-replies, number of replies by others, and length of replies. In the general natural language processing community, sentiment analysis has been carried out on various genres of texts such as product reviews[20], news and blogs[21], and tweets[22].

In this study, we focus on a large online breast cancer community and seek to understand the effect of changes in post sentiment overall through sustained participation in a community. We leverage automated sentiment analysis to conduct large-scale analysis over all the posts in the community. But instead of examining sentiment changes within threads, we examine changes of sentiment from a longitudinal standpoint. We seek to answer the following two research questions: (1) does member participation in the community over different periods of time have an impact on the member posts' sentiment? And (2) do the following factors contribute to changes in posts' sentiment: age of members, cancer stage of members, duration of membership, and amount of posting affect?

Methods

To explore changes in post sentiments in an online health community, we carried out the following steps. First, we collected all the posts in a large, public community. We trained and evaluated an automated sentiment analysis tool, specific to the community at hand. We applied the sentiment analysis to all posts and assessed the changes in

sentiment through various factors of interest both in a static and longitudinal fashion. The study was reviewed and approved by the Columbia University Medical Center IRB.

Dataset

We crawled, collected, and analyzed data from the publicly available discussion board of breastcancer.org, one of the most popular online breast cancer communities. The discussion board is organized in several forums, each with threads and posts. At the time of collection, dataset consisted of 291,528 posts in 31,034 threads, published by 12,819 community members between May 2004 and September 2010[23]. Metadata including user profiles was also extracted, consisting of self-reported demographics, diagnosis histories, and treatment histories.

Automated Sentiment Analysis – Annotation, Training, and Testing

Since this study relies on automated sentiment analysis on all posts, we built our own sentiment analysis classifier, specific to our dataset, to ensure accuracy and robustness on this particular community.

Sentiment Annotation. A random sample of 1,000 posts from the dataset was manually annotated by two annotators according to the sentiment they conveyed overall[24]. To ensure annotators chose a polarity, we restrained the annotation to positive or negative only (no neutral), and provided guidelines and examples to the annotators. Overall, a post was considered positive if its author conveyed typical positive emotions, like joy, happiness, gratitude, as well as curiosity, independently of the topic discussed. Conversely, a post was considered negative if it conveyed negative emotions, such as anger, anxiety, sadness, and hopelessness. Disagreements between the two annotators were adjudicated, resulting in a dataset of 1,000 posts annotated as either positive or negative sentiment.

Sentiment Classification. The annotated 1,000 posts were used to train and test binary sentiment classifiers. We experimented with three established robust classifiers: Maximum Entropy[25], Adaboost[26], and Support Vector Machine (SVM)[27]. Among them, Adaboost outperformed other models in a similar sentiment classification task[18], but over a dataset different from ours and with different features. Each classifier was evaluated through 5-fold cross validation according to accuracy, AUC of the ROC curve, and F measures of positive and negative classes.

Table 1. Features used for sentiment classification.

General linguistic features	
Words: number of words	*Emarks*: number of exclamation marks
PosWords: number of emotionally positive words	*NltkProb*: probability of being positive generated by the online NLTK based sentiment classifier
NegWords: number of emotionally negative words	*NltkProbNtr*: probability of being neutral generated by the online NLTK based sentiment classifier
AvgWdLen: average word length	
Sen: number of sentences	
Qmarks: number of question marks	
Domain-specific features	
Symp: number of domain-specific symptoms mentioned	*Meds*: number of domain-specific medication or treatment methods mentioned
Genre-specific features	
PosEmo: number of positive emoticons	*Person*: number of person names
NegEmo: number of negative emoticons	

We exploited several types of features to build the sentiment classifiers, from general linguistics to domain and genre specific features, as listed in Table 1. Features specific to the genre of online community and social media included emoticon lists for extracting *PosEmo* and *NegEmo* features (from http://en.wikipedia.org/wiki/Emoticon), as well as presence of people's names (personal names were tagged automatically using the Stanford Named Entity Recognizer [28] over the dataset as part of the pre-processing). General linguistic features included number of words in the post, and dictionary matching based features like number of emotionally positive/negative word stems. *PosWords* and *NegWords* were extracted by looking up two adjective lists: *glad, happy, relieved, grateful, excited, thrilled, thankful, great, lucky, pleased, blessed, fortunate, hopeful, inspiring, encouraging;* and *scared, sad, anxious, embarrassing, disappointing, confused, heartbreaking, frightened, frustrated, angry, upset, distress, stress, discouraging,* as well as their morphological variants (e.g. frustrated -> frustrating). Finally, to include other general linguistic features, we leveraged the output of a robust sentiment classifier which uses the NLTK package[29] and returns the probability of a post to be negative or otherwise, its probability of being positive.

For domain-specific features, we focused on mentions of medical terms in the posts, like treatments and side effects[30]. As such, recognizing these domain-dependent medical terms, which form a sublanguage of breast cancer communities, is a critical step in our analysis[31]. For example, in our dataset, since Tamoxifen is a widely used

medication for breast cancer patients, there are a large amount of abbreviations and misspellings such as "tamox", "tamo", and "tamoxifan" referring to this medicine. In order to capture these variations without relying on dataset-specific knowledge, we used an unsupervised, domain independent, distributional semantics based method[32] to generate two lexicons for symptoms and medications, respectively (features *Symp* and *Meds*).

Impact of Different Factors on Post Sentiment

The automated sentiment analysis output for each post a predicted probability of being positive, or sentiment score. The sentiment scores are useful, because they allow us to compare posts against each other. As such, the scores are not absolute representation of sentiment, but rather enable us to rank posts according to their sentiment polarity.

Armed with such sentiment score for each post in the dataset, we conducted the following analyses. The primary objective for our study was to assess if participation in the community has an impact on sentiment. We thus compared average sentiment scores of posts published in different periods of time with respect to user's registration date, and tracked changes of sentiment. As such, each data point is the average sentiment of all posts in a given time slice (e.g., all posts published by their authors after 3 weeks of their joining the community). To visualize the changes in sentiment through time, we plotted in addition to the individual data points a fitted curve.

For our second research question, we considered three factors (age of members, cancer stage of members, and amount of posting) in both static and longitudinal analyses to examine their impact on post sentiment. In the static analysis, members were stratified by age/stage/amount of posting, and average post sentiments were calculated for each group. Statistical tests (ANOVA and TukeyHSD[33]) were carried out to detect differences across groups. In the longitudinal analysis, sentiment scores were compared across stratified groups and duration of participation in the community to identify the patterns of sentiment change across members from different groups through time. All p-value were adjusted for multiple comparisons with the Bonferroni correction.

For both research questions, we distinguished in our analyses the initial posts (i.e., first posts that initiate threads) and all other posts. Previous research[19] found that community members expressed significantly different polarities of emotions in the initial post of a thread compared with other posts. This could be explained by the fact that the post originators were more likely to express concerns and seek support, while responses to such posts tended to be more positive by conveying encouragement and empathy.

Results

Data Annotation and Sentiment Classification

The manual sentiment annotation of the 1,000 yielded good inter-annotator agreement (Cohen's Kappa of 0.798)[34]. After adjudication and resolving disagreements, 728 out of 1,000 posts were annotated as positive, and 272 were annotated as negative. Examples of two positive and two negative posts are given in Table 2.

Table 2. Example posts of positive and negative sentiments.

Positive posts	*Negative posts*
The recovery from my lumpectomy was easy. Really. Nowhere near as difficult as I imagined. Very little pain at all. I never needed any pain meds after surgery. Good luck.	I had a mastectomy about three weeks ago and will be starting chemo at the end of the month (Dec. 27th). I wake up every morning anxious and scared. When does this go away?
I'm so happy you're feeling better!! Strange, but hey, that's our life these days. !	Just had a 6month followup with my onc. My second round of scans came out clean. However in 3 months I will be doing bloodwork for tumor markers. She didn't discuss it with me and I don't know what it is about. I understand my cancer is aggressive, but what am I not understanding here? :(

The classification performances of the three classifiers are given in Table 3. To demonstrate the effectiveness of machine learning models, performance of a baseline system is also given, which simply classified all posts as positive. The best performing system was Maximum Entropy (MaxEnt), followed by SVM and AdaBoost. Both MaxEnt and AdaBoost tended to classify posts as positive, caused by the uneven distribution of positive and negative samples in the training set. For MaxEnt, once the threshold of prediction was calibrated towards favoring negative (i.e., a post is classified as negative once the predicted probability was lower than 0.6 rather than 0.5), the F score of negative was dramatically improved. Fortunately, in our application to the entire dataset, we are more concerned with probabilities rather than discrete labels, since our modeling was based on the average likelihood of various groups of posts being positive or negative, rather than number of predicted positive and negative instances. In the remainder of the study, we relied on the MaxEnt classifier to output a sentiment score for each post.

Table 3. Performance of different sentiment classifiers according to Area under Curve of ROC, accuracy, and F scores for positive and negative sentiment polarity respectively. The baseline system classified all posts as positive.

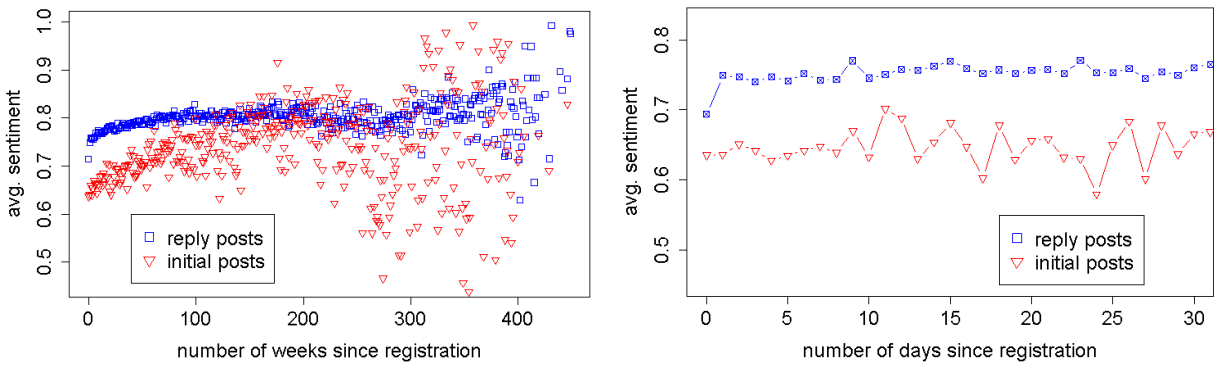
	AUC of ROC (95% CI)	Accuracy (95% CI)	F (positive) (95% CI)	F (negative) (95% CI)
MaxEnt	82.0% (2.7)	79.4% (1.8)	86.8% (1.9)	53.7% (2.8)
AdaBoost	76.0% (3.3)	76.3% (1.5)	84.6% (2.1)	48.5% (3.9)
SVM	78.1% (2.9)	73.4% (2.9)	68.4% (1.9)	58.4% (1.4)
Baseline	49.8% (0.8)	72.9% (0.7)	84.3% (1.0)	0% (0)

We analyzed the impact of individual features on the MaxEnt classifier, which assigns a weight to each feature after training, indicative of its discriminative power for the given task. Among all features, *NltkProb* (weight +2.7) had the strongest correlation with positive emotion, while *NegEmo* (weight -1.9) and *NegWords* (weight -1.2) were most correlated with negative emotion. On the contrary, *Words* (weight 0.003) and *Emarks* (weight 0.03) were borderline features, suggesting similar distributions of these features in positive and negative samples.

Participation and Posts Sentiment – Static and Longitudinal Analyses

The best performing classifier, the MaxEnt, was applied to the entire dataset based on the model trained with the 1,000 annotated samples. For each post in the dataset, a sentiment score (probability of post being positive) was calculated. The average sentiment score of the entire dataset was 0.785 (0.210 standard deviation). For the initial posts, the average sentiment score was 0.695 (0.263 standard deviation). In general, our research aligned with previous work on other online health communities that found initial posts to be less positive.

In order to examine the impact of participation through time in online discussion on sentiment overall, we plotted how sentiment scores changed through time, as computed since members' registration date. The registration dates of users were provided in the profile information of metadata. Figure 1 shows the average sentiment scores of posts that were published after membership creation at both weekly (a) and daily (b) intervals. For example, the left-most blue data point in Figure 1(a) represented the average sentiment score of all reply (i.e., non-initial) posts published by all users respectively within one week of their joining the community.



(a) (b)

Figure 1. Sentiment changes by length of membership at the time of posting, by number of weeks in (a) and number of days in (b). A colored point at (x, y) in the graph represents that the average sentiment score of all posts published by all users in the xth week (a) or day (b) after their registration is y.

Figure 1(a) indicates that, for both responding and initial posts, sentiment gets more and more positive through at least 100 weeks (2 years) of participation, with such changes most significant right after joining the community. Members, in their first days joining the community, publish posts, which are significantly more negative than later on. This is particularly true for initial posts, suggesting that newcomers to the community (likely newly diagnosed patients) express more anxiety and concerns than later in their questions to the community. Figure 1(b) provides a more granular view over the sentiment changes in the first 30 days of participation in the community, confirming that reply posts are significantly more positive than initial posts, and the increase of sentiment of initial posts does not happen until later on, at least 1 month into participation in the community. We do note a drastic increase in sentiment from posts published on the first day of joining the community to the later days, when looking at all posts (replies and initial posts combined).

In our dataset, the average length of membership of all users was 2 years 5 months (around 120 weeks); therefore, most of posts published after 200 weeks were written by a small portion of long-time users. We found that most of them were stage IV patients and showed a slight sentiment decline between 200 and 300 weeks. Topics of these posts were primarily about chemotherapy or metastasis/recurrence. While this set of posts is indeed homogeneous in sentiment and topic, it is difficult to assess the value of the analysis on such a small sample for the posts written by members who have been more than four years active in the community.

In order to obtain a more concrete understanding of how sentiment changed through sustained participation in the community, we grouped posts into nine groups, considering both short-term and long-term periods of participation. The nine groups were posts published within one day of registration, 1-3 days, 3 days to 1 week, 1 to 2 weeks, 2 weeks to 1 month, 1 to 3 months, 3 months to 1 year, 1 to 2 years, and more than 2 years since registration. An ANOVA test was carried out for the groups, for all posts and initial posts respectively, followed by a TukeyHSD test to illustrate the significances of differences between all possible group pairs. ANOVA test showed significant difference among groups in both cases (p values << 0.001). Post distribution, average sentiment scores, and p values compared with previous category given by TukeyHSD test are listed in Table 4. In this table as well as following tables, "all posts" represent initial posts and reply posts. Results showed same pattern as Figure 1, and demonstrated that the dramatic sentiment change after the first day was statistical significant in the case of all posts, while we could only see long term (3 months and then 1 year) significant changes for initial posts.

Table 4. Post distribution, average sentiment scores, and p values compared with previous category returned by TukeyHSD test, for all posts and initial posts respectively. The first p value for <1d is not available since there is no previous category to compare sentiment to. P values are adjusted for multiple comparisons with the Bonferroni correction.

		<1 d	1-3 d	3d – 1w	1-2w	2w-1m	1-3m	3m-1y	1-2 y	>2 y
All posts	Sentiment	.693	.748	.745	.753	.756	.766	.782	.800	.804
	# posts	8,369	4,203	4,361	6,235	9,906	32,302	89,304	60,944	75,781
	p value	N/A	<<0.001	1.000	1.000	1.000	0.025	<<0.001	<<0.001	0.577
Initial posts	initial	.636	.642	.637	.656	.644	.664	.685	.728	.760
	# posts	3,304	732	734	1,064	1,487	3,842	8,085	5,134	6,641
	p value	N/A	1.000	1.000	1.000	1.000	1.000	0.032	<<0.001	<<0.001

Impact of Members' Age on Sentiment

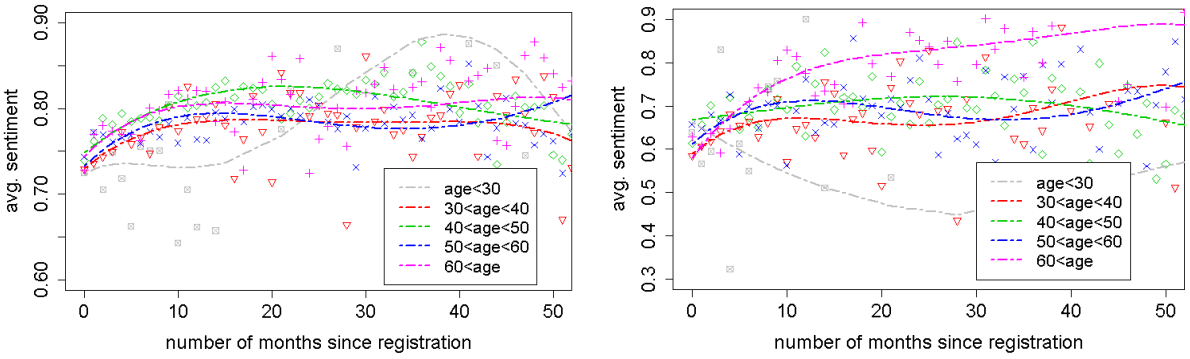
The posts in the dataset were published by 12,819 users, while a total of 14,919 user profiles were filled at least partially in the online breast cancer community and there were about 60,000 members overall. This meant that a very large majority of members were so called lurkers [35,36], who never published anything but were likely to browse some of the posts. Behavior of lurkers was beyond the scope of this study. Rather, we focused on members who had posted content. Among all non-lurkers, 1,211 provided date of birth in their profiles. Members born between 1960 and 1970 were the most dominant at the time of data collection, and the average age of all users were 47.5 (standard deviation 9.6 years), an older mean than in some other online health communities, such as weight loss forums[37].

Table 5. Average sentiment scores and number of posts published by different age groups, for all posts and initial posts respectively. This analysis is restricted to posters who provided date of birth in their profile only, 1,211 members overall.

Age group (# users)		<30 (38)	30-40 (198)	40-50 (485)	50-60 (358)	60+ (132)
All posts	Sentiment	0.742	0.768	0.793	0.778	0.791
	# posts	278	6,417	22,180	14,479	4,217
Initial posts	Sentiment	0.614	0.643	0.681	0.681	0.744
	# posts	54	841	1,873	1,323	339

To study whether age affected sentiment, we considered members who disclosed their date of birth, and grouped them into 5 groups: below 30 years old, between 30 and 40, between 40 and 50, between 50 and 60, and above 60 years old. There were 47,571 posts in the dataset published by members with date of birth information. We calculated averaged post sentiment scores, and carried out statistical tests for the groups. Table 5 shows numbers of posts published by each age group and average sentiment score of posts of each group. The ANOVA test showed significant differences among groups for both all posts and initial posts. For all posts, TukeyHSD test found that difference between all pairs of groups were significant, except between <30 and 30-40, <30 and 50-60, and between 40-50 and 60+. For initial posts, differences between <30 and all other groups were not significant. We suspect that this is caused by the very low number of members in the age group <30, as expected in a community for a disease that affects older women predominantly. Members older than 60 showed markedly more positive sentiment than

younger members, especially while publishing initial posts to start new threads. These facts might be explained by previous psychological finding of effects of older age on lower levels of psychological distress[38–40].



(a) all posts (b) initial posts

Figure 2. Sentiment changes by length of membership at the time of posting for different age groups, for (a) all posts and (b) initial posts. A colored point at (x, y) in the graph represents that the average sentiment score of all posts (a) or initial posts (b) published by users in corresponding age group in the xth month after their registration is y. Polynomial curves fitting each group were drawn for the sake of visualization.

To illustrate age's impact on longitudinal sentiment, sentiment changes over time after registration for different age groups were plotted, along with polynomial curves fitting each set of points to visualize the tendencies (Figure 2). Keeping in mind the very low sample size for members <30 years old, we do not attempt to interpret their longitudinal sentiment changes. For all other groups, however, the general trend observed earlier holds true independently of age: the longer the members participate in the community, the more positive their posts are on average. The observation that older members (>60 years old) post more positive posts, especially initial posts is visible as well on the plots.

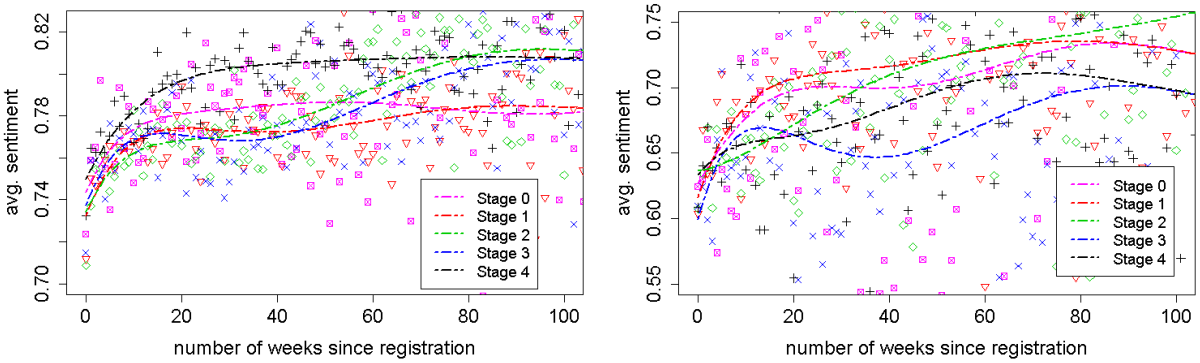
Impact of Member's Cancer Stage on Sentiment

In our dataset, 4,602 users (who published 172,566 posts) had self-reported cancer stage information. Among them, 442 members were stage 0 patients, 1,407 were stage I, 1,544 were stage II, 650 were stage III, and 559 members were stage IV. Table 6 provides numbers and average sentiment scores of posts published by members in different stages. Although there were significantly fewer stage IV patients than stage I and II patients, they published many more posts and formed the most active cancer stage group in breast cancer forum[23]. Moreover, stage IV patients were the most positives posters in term of the emotion expressed through the reply posts they wrote, but not initial posts. For all posts, comparisons between stage 0, stage I, and stage II, returns non-significant results according to adjusted p values.. For initial posts, only the differences between stage I and stage III and between stage II and stage III were significant.

Table 6. Average sentiment scores and number of posts published by patients in different stages, for all posts and initial posts respectively.

Cancer stage (# users)		Stage 0 (442)	Stage I (1,407)	Stage II (1,544)	Stage III (650)	Stage IV (559)
All posts	Sentiment	0.775	0.771	0.776	0.782	0.796
	# posts	9,229	36,422	39,398	27,806	59,711
Initial posts	Sentiment	0.675	0.690	0.687	0.661	0.675
	# posts	820	3,344	4,218	2,534	4,829

Figure 3 illustrates longitudinal sentiment of different cancer stage groups. Not only were the stage IV users the most positive, but they also showed the fastest change towards positive after registering in the breast cancer forum. However, these findings were specific to reply posts. These findings indicate that stage IV users seek support through starting threads with negative posts, but are very active in providing emotional support to their peers, through posting positive replies.



(a) all posts (b) initial posts

Figure 3. Sentiment changes by length of membership at the time of posting for different cancer stage groups, for (a) all posts and (b) initial posts. A colored point at (x, y) in the graph represents that the average sentiment score of all posts (a) or initial posts (b) published by users in corresponding cancer stage in the xth month after their registration is y. Polynomial curves fitting each group were drawn for the sake of visualization.

Impact of Member's Posting Activity on Sentiment

The last factor we considered was the amount of posting by each individual. Table 7 groups members into 5 groups by number of posts, listing the distributions and average sentiment of each group. There were 8247, 3527, 757, 255, and 24 profiles in the 5 groups respectively. Although members who published less than 50 times wrote only 20% of all posts, approximately half of the initial posts were authored by these members. This suggests that new members tend to seek information and support while long-time members provided information and support more than they requested it. All differences of sentiment scores between groups, including both all posts and initial posts, were significant, except between group of < 5 and 5-50 for initial posts.

Table 7. Average sentiment scores, number of posts published by patients, and number of posts published per user in different stages, for all posts and initial posts respectively.

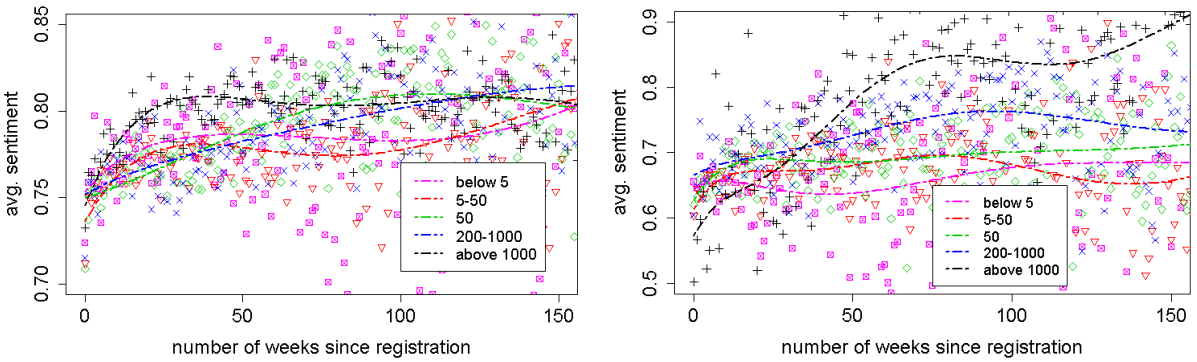
User post number (#users)		< 5 (8,247)	5-50 (3,527)	50-200 (757)	200-1000 (255)	1000+ (24)
All posts	Sentiment	0.727	0.754	0.779	0.806	0.817
	# posts	16,725	36,422	73,951	102,466	39,944
	avg # post	2.0	10.3	97.7	401.8	1664.3
Initial posts	Sentiment	0.657	0.658	0.683	0.730	0.828
	# posts	4,565	9,445	7,399	6,635	2,990
	avg # post	0.6	2.7	9.8	26.0	124.6

Figure 4 illustrates how sentiment changed over time for different groups of members with different posting activity count. In general, active members (i.e., with more posts authored) were likely to gain sentiment improvement faster and more significantly. It is particularly interesting to note that although members posting more than 1,000 times throughout their time in the community, and who were long-time users, had a significantly higher sentiment score in average, their sentiments were as negative as other members when they just joined the forum, especially for their initial posts. The pattern seen in Table 7 and Figure 4 seems to suggest that long-time users, who suffered from cancer but benefited from hearing from their peers online at early stages of participation, changed their roles in the forum later and acted as information and support providers more than requesters. Such role change should be another important outcome of online discussion participation.

Discussion

Principal Findings

Our study results suggest that members benefit from sustained participation in a breast cancer community with respect to the sentiment they convey through their posts. At the early stages of participation, sentiment of users usually increased significantly, and the rate of improvement dropped after several weeks, followed by a slower positive sentiment increase which could last for as long as several years. Our study also showed that compared with



(a) all posts (b) initial posts

Figure 4. Sentiment changes by length of membership at the time of posting for different groups of posting amount, for (a) all posts and (b) initial posts. A colored point at (x, y) in the graph represents that the average sentiment score of all posts (a) or initial posts (b) published by users grouped by their number of posts in the xth month after their registration is y. Polynomial curves fitting each group were drawn for the sake of visualization.

reply posts, initial posts of threads were more emotionally negative, especially at the beginning of participation. Sentiment increases of initial posts were more dramatic but long term. A qualitative analysis over the forum data showed that newcomers of the forum were more likely to be newly diagnosed or post-treatment patients. For most of them, going online was the choice when some of their needs, either informational or emotional, could not be met in other settings such as family and hospitals. As a result, we found a large amount of posts with strong negative sentiments, especially initial posts, published by newcomers asking various questions about cancer symptoms, medication use and side effects, and choices of therapeutic method, which were the issues usually brought up by individuals with little cancer or treatment experiences. In contrast, long-time members were more likely to be cancer survivors or patients who were recovering or being treated as a routine part of their lives. It is likely they were more experienced, empowered, and acted more as informational and emotional support providers rather than requesters, and were expressing more encouragement and empathy in the threads in which they participated. The different patterns of reply posts and initial posts also suggested that people immersed themselves quickly into the discussion by learning to encourage others and provide information through replying, but were still concerned about their own issues.

Our study examined three factors' impacts on sentiment and sentimental changes: age, cancer stage, and amount of posting. We showed that all three factors had an impact on the members' sentiment on average. Statistically significant differences were found for every stratified group. For age, we found that users older than 60 years old showed the most positive sentiment, especially while publishing initial posts. There were no significant differences between longitudinal aspects of different age groups. With respect to cancer stage, although there were significantly fewer stage IV patients than any other stage, they published many more posts and formed the most active cancer stage group in the breast cancer forum. They showed the fastest change towards positive sentiment after registering in the breast cancer forum. They also were the most positive in their replies, while the most negative in their initial posts. The last factor, amount of posting, also made a difference. Members who published less than 50 posts, mostly newcomers and lurkers, were responsible for only 20% of all posts, but around half of the initial posts were authored by these users, which indicated that new users and lurkers tended to seek information and support while long-time members provided information and support more than requested it. Long-time members, who suffered from cancer but benefited from hearing from their peers online at early stages, later changed their roles in the forum later and acted more as information and support providers.

Limitations

Our study was exploratory and has several limitations. First, the analyses rely on the output of an automatic sentiment classifier, which while providing state-of-the-art accuracy is not 100% accurate. Further feature engineering has the potential to improve the classification accuracy. Second, the classification was defined as a two-category problem: positive and negative, and documents with neutral sentiment were simply regarded as ones whose sentiment scores lie near the boundary of the binary classification. Since the sentiment scores are a mean to comparing posts in aggregate, it might be also useful to leverage a more granular classification, or at least one that

considers neutral as a category on its own[41]. Third, profile information, especially cancer stage, was extracted at the time of data collection, and such information might have been edited by members through time, as their disease evolved. Finally, this study was conducted on a single online health community. It will be interesting to see the impact of these factors in different communities specific to breast cancer as well as to or other chronic conditions.

Future Work

This study brings up several research questions we would like to explore in the future. While one interpretation of our findings is that sustained participation in an online health community overall increases the sentiment of members' posts, we must acknowledge that there is a strong uncertainty to this interpretation due to right censoring issues, common to longitudinal observational studies. As the number of members who stay in the community decrease with time since registration, one must think of potential reasons for the right censoring of data: is it that members with adverse health outcomes were too sick to continue participating in the community, or even that individuals who did not receive appropriate support from their peers stopped participation. In other words, it is possible that only the people for which the community is beneficial emotionally are the ones that stay in the community through time, while others simply stop posting. Under this assumption, there is no causal link between participation in the community and positive sentiment on average.

We did not examine the impact of lurking (and of its duration before posting for the first time) on participation and sentiment in particular. Because the community we studied does not keep track of members' reading activity, this is a difficult question to study quantitatively, but an important one to consider in the future.

Another area of research we plan to explore further is that the sentiment of the posts alone is a rough representation of the sentiment or emotions of community members. As we refine our understanding of the different topics conveyed in an online health community, it will be critical to understand the relationship between sentiment and different topics. For instance, a member's anger at an insurance company refusing to reimburse her treatment and a member's anxiety faced with a dire test result represent very different aspects of sentiment with respect to cancer in general. The longitudinal evolution of topics and their associated sentiments for community members is an area in much need for further analysis.

Conclusion

This paper carried out an exploratory study over a popular public online community for breast cancer and used automated sentiment analysis to investigate correlations between sentiment changes of users and different participation-relevant factors. Finding suggests that as participation is sustained, posts' sentiment increases towards positive. Further, members convey more positive posts when replying to their peers than when initiating a thread. In addition, we discovered that users in different ages, cancer stages, and stages of participation showed different sentiment patterns. Most significantly, members over 60 years old and stage IV members were expressing more positive sentiment than any other groups of people, while newcomers to the community tend to post more negative initial posts than long-time members. This study contributes to further the understanding of community participation on members' attitudes and opens up to a number of research questions to explore further.

Acknowledgements

This work is supported by NSF (National Science Foundation) award #1027886 (NE) and NCI (National Cancer Institute) award R21 CA143632 (EB, JO, NE). We are grateful to Rimma Pivovarov for the helpful discussions.

References

1. Medicine 2.0: Peer-to-peer healthcare. Available at: http://www.pewinternet.org/Reports/2011/Medicine-20.aspx.
2. Eysenbach G, Powell J, Englesakis M, Rizo C, Stern A. Health related virtual communities and electronic support groups: systematic review of the effects of online peer to peer interactions. *BMJ*. 2004;328(7449):1166.
3. Ziebland S, Chapple A, Dumelow C, Evans J. How the internet affects patients' experience of cancer: a qualitative study. *BMJ*. 2004;328(7439):564.
4. Meier A, Lyons EJ, Frydman G, Forlenza M, Rimer BK. How cancer survivors provide support on cancer-related Internet mailing lists. *J Med Internet Res*. 2007;9(2):e12.
5. Wang Y, Kraut R, Levine J. To stay or leave?: the relationship of emotional and informational support to commitment in online health support groups. *Proc ACM 2012 Conf Comput Support Coop Work*. 2012:833–842.
6. Chung JE. Social Networking in Online Support Groups for Health: How Online Social Networking Benefits Patients. *J Health Commun*. 2013.
7. Vlahovic TA, Wang Y, Kraut RE, Levine JM. Support Matching and Satisfaction in an Online Breast Cancer Support Community. *Proc 32nd Annu ACM Conf Hum factors Comput Syst*. 2014:1625–1634.

8. Kim E, Han JY, Moon TJ, et al. The process and effect of supportive message expression and reception in online breast cancer support groups. *Psychooncology*. 2012;21(5):531–40.
9. Rozmovits L, Ziebland S. What do patients with prostate or breast cancer want from an Internet site? A qualitative study of information needs. *Patient Educ Couns*. 2004;53(1):57–64.
10. Overberg R, Otten W, de Man A, Toussaint P, Westenbrink J, Zwetsloot-Schonk B. How breast cancer patients want to search for and retrieve information from stories of other patients on the internet: an online randomized controlled experiment. *J Med Internet Res*. 2010;12(1):e7.
11. Davison KP, Pennebaker JW, Dickerson SS. Who talks? The social psychology of illness support groups. *Am Psychol*. 2000;55(2):205.
12. Nambisan P. Information seeking and social support in online health communities: impact on patients' perceived empathy. *J Am Med Informatics Assoc*. 2011;18(3):298–304.
13. Beaudoin C, Tao C. Modeling the impact of online cancer resources on supporters of cancer patients. *New Media Soc*. 2008;10(2):321–344.
14. Sharf BF. Communicating breast cancer on-line: support and empowerment on the Internet. *Women Health*. 1997;26(1):65–84.
15. Han JY, Shah D V, Kim E, et al. Empathic exchanges in online cancer support groups: distinguishing message expression and reception effects. *Health Commun*. 2011;26(2):185–97.
16. Bender JL, Jimenez-Marroquin M-C, Jadad AR. Seeking support on facebook: a content analysis of breast cancer groups. *J Med Internet Res*. 2011;13(1):e16.
17. Ma X, Chen G, Xiao J. Analysis of an online health social network. *Proc 1st ACM Int Heal informatics Symp*. 2010:297–306.
18. Qiu B, Zhao K, Mitra P, et al. Get Online Support, Feel Better -- Sentiment Analysis and Dynamics in an Online Cancer Survivor Community. *2011 IEEE Third Int'l Conf Privacy, Secur Risk Trust*. 2011:274–281.
19. Zhao K, Greer G, Qiu B, Mitra P. Finding influential users of an online health community: a new metric based on sentiment influence. *arXiv Prepr arXiv12116086*. 2012.
20. Cui H, Mittal V, Datar M. Comparative experiments on sentiment classification for online product reviews. *AAAI*. 2006.
21. Godbole N, Srinivasaiah M, Skiena S. Large-Scale Sentiment Analysis for News and Blogs. *ICWSM*. 2007.
22. Pak A, Paroubek P. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *LREC*. 2010:1320–1326.
23. Jha M, Elhadad N. Cancer Stage Prediction Based on Patient Online Discourse. In: *Proc ACL BioNLP (Association for Computational Linguistics Bio Natural Language Processing) Workshop*.; 2010:64–71.
24. Bo Pang and Lillian Lee. Opinion Mining and Sentiment Analysis. *Found Trends® Inf Retr*. 2006;1(2):91–231.
25. Pietra S Della. Inducing features of random fields. *Pattern Anal Mach Intell*. 1997;19(4):380–393.
26. Freund Y, Schapire R. A decision-theoretic generalization of on-line learning and an application to boosting. *J Comput Syst Sci*. 1997;55(1):119–139.
27. Suykens JAK, Vandewalle J. Least squares support vector machine classifiers. *Neural Process Lett*. 1999;9(3):293–300.
28. Finkel JR, Grenager T, Manning C. Incorporating non-local information into information extraction systems by gibbs sampling. In: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*.; 2005:363–370.
29. Sentiment Analysis with Python NLTK Text Classification. Available at: http://text-processing.com/demo/sentiment/.
30. Meier A, Lyons EJ, Frydman G, Forlenza M, Rimer BK. How cancer survivors provide support on cancer-related Internet mailing lists. *J Med Internet Res*. 2007;9(2):e12.
31. Friedman C, Kra P, Rzhetsky A. Two biomedical sublanguages: a description based on the theories of Zellig Harris. *J Biomed Inform*. 2002;35(4):222–235.
32. Elhadad N, Zhang S, Driscoll P, Brody S. Characterizing the Sublanguage of Online Breast Cancer Forums for Medications, Symptoms, and Emotions. In: *Proc AMIA Annual Fall Symposium*.; 2014.
33. Winer BJ. *Statistical principles in experimental design*. McGraw-Hill Book Company; 1962.
34. Cohen J, others. A coefficient of agreement for nominal scales. *Educ Psychol Meas*. 1960;20(1):37–46.
35. Setoyama Y, Yamazaki Y, Namayama K. Benefits of peer support in online Japanese breast cancer communities: differences between lurkers and posters. *J Med Internet Res*. 2011;13(4):e122. doi:10.2196/jmir.1696.
36. Van Uden-Kraan CF, Drossaert CHC, Taal E, Seydel ER, van de Laar M a FJ. Self-reported differences in empowerment between lurkers and posters in online patient support groups. *J Med Internet Res*. 2008;10(2):e18.
37. Hwang K, Ottenbacher A. Social support in an Internet weight loss community. *Int J Med Inform*. 2010;79(1):5–13.
38. Singer J, Rexhaj B, Baddeley J. Older, wiser, and happier? Comparing older adults' and college students' self-defining memories. *Memory*. 2007;15(8):886–98.
39. Hoffman K, McCarthy EP, Recklitis CJ, Ng AK. Psychological distress in long-term survivors of adult-onset cancer: results from a national survey. *Arch Intern Med*. 2009;169(14):1274–1281.
40. Kaiser NC, Hartoonian N, Owen JE. Toward a cancer-specific model of psychological distress: population data from the 2003--2005 National Health Interview Surveys. *J Cancer Surviv*. 2010;4(4):291–302.
41. Schler J. The importance of neutral examples for learning sentiment. In: *In Workshop on the Analysis of Informal and Formal Information Exchange during Negotiations*.; 2005.

On Learning and Visualizing Practice-based Clinical Pathways for Chronic Kidney Disease

Yiye Zhang, MS, Rema Padman, PhD, Larry Wasserman, PhD
Carnegie Mellon University, Pittsburgh, PA

Abstract
Chronic Kidney Disease (CKD) is a costly and complex disease affecting 20 million US adults. Recent studies suggest that care delivery changes may improve clinical outcomes and quality of patient experience while reducing costs. This study analyzes the treatment data of 8,553 CKD patients to learn practice-based clinical pathways. Patients' visit history is modeled as sequences of visits containing information on visit type, date, procedures and diagnoses. We use hierarchical clustering based on longest common subsequence (LCS) distance to discover six patient subgroups, with each subgroup differing in the distribution of demographics and health conditions. Transitions of visits with high probabilities are elicited from each patient subgroup to learn common clinical pathways and treatment durations. Insights from this study can potentially result in new evidence to support patient-centered treatment approaches, empower CKD patients to better manage their disease and its complications, and provide a review guide for clinicians.

1. Introduction

Chronic Kidney Disease (CKD) is a costly, complex and high mortality health condition affecting 20 million US adults[1]. Prevalence is estimated to be 8-16% worldwide[2]. Most patients are unaware of their disease, with more than 40% with end-stage renal disease (ESRD) requiring emergency hospitalization and dialysis for acute kidney failure[3]. CKD patients make up only 1.5% of US Medicare population, but incur almost 10% of Medicare costs annually[1]. These adverse outcomes are potentially preventable or can be mitigated through early identification and treatment of individuals at risk as they progress through the five stages of CKD. Typically, the treatment of CKD is focused on preventing the worsening of condition, such as maintaining patients in their current disease stage and delaying the progression from stage 5 to ESRD and dialysis. CKD treatment currently follows opinion-based, consensus guidelines developed by nephrologists. The two widely known guidelines are Kidney Disease Outcomes Quality Initiative (KDOQI) and Kidney Disease: Improving Global Outcomes (KDIGO), which is a collection of global clinical practice guidelines for different complications of kidney disease[4,5]. However, recent studies suggest that care delivery changes may improve clinical outcomes, enhance quality of patient experience, and reduce annual total per capita health spending[1,6]. Hence, there is a need to review the actual clinical events performed on patients and discover practices that may lead to improved outcomes. At the same time, patient education plays a vital role in effective CKD treatment. For example, pre-ESRD education may enhance patient satisfaction, delay dialysis, and increase cost-effectiveness[7]. Engagement and education using personalized clinical pathways and tools that support shared decision-making with their clinicians are likely to be a valuable approach in delaying disease progression. Leveraging the availability of a rich and unique clinical dataset extracted from the Electronic Health Record (EHR) of a community nephrology practice in Western Pennsylvania, we analyze the treatment of 8,533 CKD patients to learn practice-based clinical pathways that have the potential to meet these objectives.

2. Background

CKD is a condition where the kidney gradually loses its function[1]. Patients with high blood pressure and diabetes are especially at risk for CKD, but other health problems such as cardiovascular disease and high cholesterol are also known risk factors[1]. Due to the close monitoring requirements of the disease, a large proportion of Medicare budget every year is utilized in the treatment of CKD patients[1]. In fact, the per person per year average cost of CKD patients is $23,128, which is more than twice the average cost, at $11,103, of non-CKD Medicare patients, and the cost increases as the disease condition deteriorates[1]. The best estimate of kidney function is the glomerular filtration rate (GFR), or the amount of blood that passes through glomeruli per minute[8]. Patients are divided into 5 stages based on their GFR level: $\geq$ 90 mL/min as stage 1, 60-89 mL/min as stage 2, 30-59 mL/min as stage 3, 15-29 mL/min as stage 4, and finally <15 mL/min as stage 5[8]. Conditions worse than stage 5 include ESRD, which requires costly dialysis and poses significant negative impact on patients' life styles and health outcomes[9]. Studies have suggested that risks of CKD increase with age, with half of the CKD stage 3 cases developing after the age of 70 years[10,11]. The interactions between age, sex, race, especially between African-American and White, and the risk of CKD and

ESRD are also found in multiple studies[12-14], but there have been limited efforts in discovering pathways of care from actual treatment data.

In this study, we elicit practice-based clinical pathways of CKD treatments using 4 years of office visit data containing detailed information on 8,533 patients who were treated at a community nephrology practice in Western Pennsylvania. Practice-based clinical pathways reflect patients' disease progression over time in terms of the treatments provided during office visits. Specifically, we aim to identify the different patient types and the distinct paths along which their treatments may evolve. Office visit data is modeled as a sequence of visits, such that we can capture the chronological changes in visit type, visit date, procedures and diagnoses. Each patient is represented by one and only one sequence. We apply hierarchical clustering, a cluster analysis method frequently used in biomedical research to identify underlying structures[15,16], to patients' visit sequences in order to discover distinct patterns in the sequences that separate one subgroup of patients from another. To visualize the clinical pathway in each subgroup, we incorporate a stochastic model to connect pairs of visits that have high transition probabilities. This learned model can be instantiated for any patient, to allow the patient to visualize a projected clinical pathway and engage with the clinicians for shared decision making. It can also be used as a practice management tool for clinicians who wish to review their practices against current consensus guidelines for CKD.

Prior research has studied methods for identifying clinical pathways for a few health conditions[17-19]. Huang et al. developed a clinical pathway mining algorithm and applied it on hospital inpatient care for bronchial lung cancer, gastric cancer, cerebral hemorrhage, breast cancer, infarction, and colon cancer[17]. Lin et al. adopted Hidden Markov Model in mining clinical pathways and applied it to the normal spontaneous delivery process[18]. Also, Lin et al. applied Bayesian networks to study the causal relationships between medical treatments and transitions of patient's physiological states in the Hemodialysis process[19]. Yet, there is limited research specifically targeted for the five CKD stages prior to ESRD for many reasons. First, CKD treatments last for years as patients progress from stages 1 to 5, completing multiple office visits during this period. Second, CKD treatment is laboratory-test driven; before each visit to the clinic, patients are asked to complete specific laboratory tests and bring the results for physician review. Results from laboratory tests largely determine the next steps in treatments. Few procedures are performed as part of CKD treatment. Finally, CKD patients tend to suffer from a variety of comorbidities such as hypertension and diabetes[1]. Thus, patients need to be treated for those conditions as well. In fact, patients in the same CKD stage can evolve along distinctly different clinical pathways depending on their comorbidities and lifestyle choices. Therefore, clinical pathway mining algorithms designed for other conditions do not apply well to CKD due to the differences in the nature of disease and the many dimensions associated with its treatments. In Section 2, we describe our data set and illustrate the methods used to mine clinical pathways. In Section 3, the generated pathways and their components are presented using figures and tables. We discuss limitations and future steps in Section 4, and finally summarize our conclusions in Section 5.

2. Data and Methods

2.1 Data

De-identified data for this study was obtained from a large, forward-thinking nephrology practice in Western Pennsylvania that implemented an Electronic Health Record (EHR) system in 1994. The community practice provides care to patients at multiple clinics dispersed over a three-county geographic region. A four-year extract of the data, from March 2009 to May 2013, is drawn from the EHR. There are 8,553 patients in our study dataset, 4,195 female and 4,358 male patients (Table 1). Majority of the patients are White or African-American, with the ethnicity ratio remaining steady over the years. In this study, we mainly focus on patients' office visits, which are categorized as new patient visit, follow up, extended follow up, hospital follow up, pre-ESRD education, and kidney biopsy review. Excluding vascular access center visits and check up for dialysis patients, 8,500 patients had at least one office visit. To reduce noise in the data set, diagnoses captured in our pathways include only CKD stage 1 through stage 5 (ICD9 = 585.1-585.5) and its commonly known comorbidities: diabetes (ICD9 = 250.xx), hypertension (ICD9 = 401.xx or ICD9 = 405.xx) and acute kidney failure (AKF, ICD9 = 584.xx). We chose these 3 conditions because AKF is frequently associated with increases in CKD stages, and prevalence of CKD is high among patients with diabetes and hypertension[1,20,21]. Procedures are captured using Current Procedural Terminology (CPT) codes. We include 27 types of procedures, such as renal ultrasound and renal artery Doppler.

Table 2 lists the number of occurrences of diagnoses and procedures recorded during visits over the 4-year study period. From previous study[22], we suspect that a clinical pathway for CKD may contain 4 to 6 office visits, so we extracted data on patients who started from a specific point in CKD stages 1 to 5 and had at least 8 visits, to ensure

adequate number of visits are included in the learned pathways. A total of 2,511 patients fulfilled the criteria. Our data shows that procedures are not commonly performed for CKD patients during office visits. Out of 52,349 visits in the data, only 201 visits had some type of procedures performed, for a total of 262 times. Patients in CKD stage 4 received the largest number of associated procedures, majority of which were referral to pre-ESRD education (Table 2). In this analysis, we exclude patients who only had hospital visits, but do not exclude hospital visits of patients who are also office patients so that we can capture the complete history of each patient's treatments. In addition, since CKD is a chronic condition, patients may not start and complete treatments during the four-year study period. Therefore, the clinical history we see may start from a follow-up visit and end before patients complete treatment, move on to dialysis, or decease. Mortality information is not captured in the dataset.

Table 1. Patient demographics

		70 and above	Between 50 and 69	Under 50	Total
Female	White	2229	1080	337	3646
	African-American	101	93	31	225
	Other	8	11	12	31
	Unknown	157	97	39	293
Male	White	2328	1221	293	3842
	African-American	74	110	43	227
	Other	15	10	11	36
	Unknown	134	84	35	253
Total		5046	2706	801	8553

Table 2. Occurrence of diagnoses and procedures recorded during visits over the 4-year period

Condition	Number of occurrences	Number of Associated Procedures
CKD stage 1	407	3
CKD stage 2	2290	2
CKD stage 3	18698	22
CKD stage 4	9363	141
CKD stage 5	1493	8
Acute Kidney Failure	4298	13
Diabetes	10269	19
Hypertension	30926	54

2.2 Methods

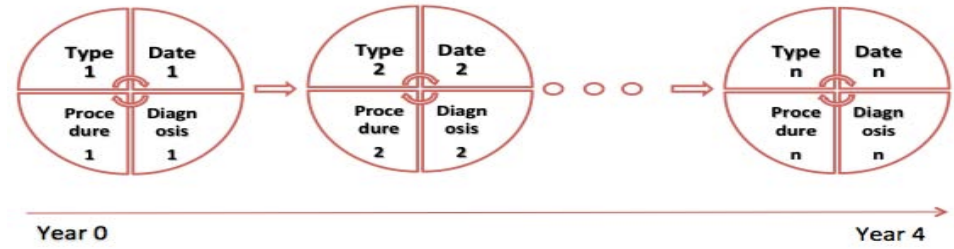


Figure 1. Illustration of a patient's sequence of visits

We represent each patient's clinical history as one and only one sequence of office visits ordered chronologically, with each visit in the sequence categorized by visit type, visit date, procedure(s) performed, and diagnoses noted (Figure 1). Length of sequence is the number of visits in the sequence. Unfortunately, data on medications was incomplete, so we do not include medications in the visit node in this analysis, but this extension is straightforward using the current representation. Each variable in the visit, such as diagnosis, is stored as an independent table in a secure relational database. Patients are characterized by a de-identified patient ID, age: under 50, above 50 and below 70, 70 and above; sex: female/male; and race: White, African-American, Other. Patients receive none, or multiple procedures and diagnoses during a visit. Given the data set, we first identify all combinations of unique procedures performed during a single visit, such as 'renal Doppler, renal ultrasound', and combination of unique diagnoses noted during a single visit, such as 'CKD stage 4' or 'CKD stage 3, hypertension.' Each combination of procedures is given a label 'PCx', where x is a number from 1 up to the total number of combinations. We apply the

same to the combination of diagnoses, labeling each as 'Dy', where y is a number from 1 up to the total number of combinations. Then, each visit can be represented as 'type', 'date', 'PCx', and 'Dy.' We scan all visits and identify unique combinations of 'type', 'PCx', and 'Dy,' so we can label each visit as 'Vz' where z is a number from 1 up to the total number of combinations. For each patient, we use 'date' to chronologically order each 'Vz' into a sequence. It is possible for a sequence to have multiple visits of the same sort, such as V1-V3-V3-V2-V1, if the visit is of the same type, procedures and diagnoses. The numbers in the labels are used to distinguish each label, not to show the order of time. Table 3 lists examples of how we label combinations of procedures and diagnoses, and visits.

Table 3. Examples of procedure, diagnosis and visit labels

Label	Description
D1	CKD stage 3, hypertension
D2	CKD stage4
PC1	renal ultrasound
V1	type: follow up, date: 2011/1/1, procedure: PC1, diagnosis: D2
V2	type: new patient visit, date: 2012/1/1, procedure: N/A, diagnosis: D1

After visit sequences are properly represented, we apply hierarchical clustering[23], based on a distance measure called longest common subsequence (LCS) distance[24], to cluster sequences by similarity. LCS has been widely applied in biomedical research as a similarity measure used in trajectory analysis and protein sequence analysis[25]. It is the longest subsequence that 2 sequences have in common, while preserving the order of occurrence, but possibly separated. For instance, if patient 1 has a sequence V1-V2-V1-V3-V1-V2, and patient 2 has a sequence V1-V2-V3-V1-V4, then their LCS is of length 4, being V1-V2-V3-V1.

$$\mathrm{LCS}(\mathrm{x},\mathrm{y}) = \max\{|\mathrm{u}|: \mathrm{u} \in \mathrm{S}(\mathrm{x},\mathrm{y})\}$$

is the length of LCS, where |u| is the length of the longest common subsequence for the pair of sequences (x, y), and S(x,y) is the nonempty set of subsequences of sequences x and y. LCS distance (dLCS) is defined as

$$dLCS(x,y) = |x| + |y| - 2LCS(x,y)$$

Each patient's visit sequence will be compared to the rest of the patients' sequences, and each such comparison generates an dLCS. Since our sample size is 2511, the size of the dLCS matrix is 2511 by 2511. To enhance the result of hierarchical clustering, we apply a data transformation and square each dLCS in the matrix. Hierarchical clustering is applied on this transformed matrix to cluster similar sequences into subgroups. The optimal number of clusters is detemined using Silhouette,[26] a measure commonly used in cluster analysis .

In order to create clinical pathways for each cluster, we elicit all transitions seen in the visit sequences of patients. The visit taking place first is called *source* and the successor is called *target*. For example, given a sequence V3-V2-V2-V5, there are 3 transitions: (V3, V2), (V2, V2), (V2, V5), where the source is V3 and target is V2 in the first transition. To build pathways, we connect transitions with a certain threshold such that they form an overall path. For example, given 3 transitions: (V1, V2), (V3, V4), (V2, V4), we can connect 1st and 3rd transitions to form a path V1-V2-V4, because the target in the first transition is the source in the 3rd transition. In order to ensure that transitions we use to build pathways occurr in the data beyond a certain probability, we define a measure called *weight:*

$$weight = \frac{\left|N_j = target \middle| N_i = source\right|}{|N_i = source|}$$

where $|N_i = source|$ is the number of times N_i appears as a source, and $|N_j$=target | N_i=source| is the number of times N_j appears as a target given N_i is the source. Weight is the conditional probability that N_i will transition to N_j. For example, in the sequence V3-V2-V2-V5, (V3, V2) has weight of 1, and (V2, V2) and (V2, V5) have weight of 1/2, respectively. In order to reduce noise, we include only transitions with weight above 0.3, a threshold that we think will give rise to a map of pathways that contains sufficient information and easily interpretable by clinicians. We also capture and categorize the difference in days between source and target as *gap*: 1) less than 1 week, 2) 1-2 weeks, 3) 2-3 weeks, 4) 3 weeks-1month, 5) 1-2 months, 6) at 2-3 months, 7) 3-4 months, 8) 4-6 months, 9) 6-12 months, and 10) at least 1 year. If the same transition of visits has more than one gap observed in multiple sequences, we take the average of all the gaps for a transition in term of days, then categorize it into the above categories. A pathway is not necessarily created from one patient's visit sequence. It can be built with multiple transitions by multiple patients, if the target of one transition of a patient is identical to the source of one transition of another patient. Length of a pathway is the number of visits included in the pathway.

In each cluster, we visualize clinical pathways of length more than 3 as graph of nodes using Gephi[27]. Each node in the graph represents a visit label 'Vz', and 2 nodes that have transitional relationships are connected by an edge. The thickness of the edge is determined by weight. The node colors in the pathway graph range from purple to white, and is determined by degree, the number of directed edges connected to a node, both inbound and outbound[28]. The more purple a node appears, the more connecting edges it has, and conversely the whiter a node appears, the fewer connecting edges it has. We also keep track of high out-degree nodes, which are the nodes with high number of outbound edges, because they are points where one pathway split into two or more pathways. In order to check how many of the actual transitions are captured by the learned clinical pathways, we define a measure called average maximum LCS. Given all the pathways learned in a cluster, we calculate the length of LCS between each pathway and each patient's sequence in the cluster. For example, given 3 pathways and 5 patients, we will have a 3 by 5 matrix of LCS between each of the 3 pathways and each of the 5 patients' sequences. We call the maximum length of LCS for each patient as maximum LCS (mLCS). Average maximum LCS is the average of the mLCS.

3. Results

There are 2,468 unique sequences among 2,511 sequences, whose lengths range from 8 to 41, with an average of 12. Of 2,511 patients in the sample, only 65 patients shared the same clinical history with one or more other patients. Thirty-nine patients with CKD stage 3 and hypertension had follow up visits only while maintaining the same diagnoses. Four patients started as new patients with CKD stage 3, diabetes, and hypertension, and continued with the same diagnoses for 7 more follow up visits. Another 4 patients with CKD stage 4 and hypertension had 10 follow up visits with the same diagnosis. Aside from these patients, our data shows that 98% of the patients have evolved in their unique ways regardless of their initial diagnoses, an indication of the unpredictability and diversity in the conditions of CKD patients, and variations in practice patterns. A total of 281 V's were found from the data, and the total number of visits is 30,780. Table 4 lists the 10 most frequent V's, and their occurrences in the data. Follow up visits with CKD stage 3 or CKD stage 4, and hypertension are the most common, and ESRD education sessions for CKD stage 4 patients also count as one of the most frequent.

Table 4. Frequent visit contents

Visit Type	Diagnoses	Procedures	Count	% of Total
FUP	Chronic Kidney Disease Stage 3, Hypertension	N/A	5050	16.4%
FUP	Chronic Kidney Disease Stage 4, Hypertension	N/A	3837	12.5%
FUP	Chronic Kidney Disease Stage 3, Diabetes, Hypertension	N/A	2849	9.3%
FUP	Chronic Kidney Disease Stage 4, Diabetes, Hypertension	N/A	1833	6.0%
FUP	Chronic Kidney Disease Stage 5, Hypertension	N/A	751	2.4%
FUP	Chronic Kidney Disease Stage 2, Hypertension	N/A	396	1.3%
FUP	Acute Kidney Failure, Chronic Kidney Disease Stage 3, Hypertension	N/A	381	1.2%
ESRD	Chronic Kidney Disease Stage 4	N/A	376	1.2%
FUP	Chronic Kidney Disease Stage 5, Diabetes, Hypertension	N/A	319	1.0%
FUP	Chronic Kidney Disease Stage 2, Diabetes, Hypertension	N/A	213	0.7%

We identified 6 clusters of patient sequences based on the data of 2,511 patients who had at least 8 visits (Figure 2). The length of pathways is between 4 and 6, and number of pathways ranges from 17 to 67. As Figure 2 shows, graphs of clusters 1 and 6 are relatively sparse compared to graphs of clusters 2 through 5. More than half of the pathways in cluster 1 ended with follow up visits with diagnoses of CKD stage 4, diabetes and hypertension. Majority of the cluster 2 pathways ended with follow up visits with diagnosis of CKD stage 3, diabetes and hypertension. All pathways in cluster 3 ended with follow up visits of CKD stage 3 and hypertension. Pathways in cluster 4 predominantly ended with CKD stage 4 and hypertension. All pathways in cluster 5 ended with hospital treatment, and diagnoses are not noted in the data because office and hospital visits have separate billing systems. Finally cluster 6 seems to contain many non-compliant patients. Patients in cluster 3 have cancelled or not shown up for more than 56% of the visits, causing the data to contain empty visits without noted diagnosis. Given that most of the pathways have length of 4, average maximum LCS in clusters 2 through 5 suggest that pathways in these clusters represent a fair amount of information about their patients. On the other hand, clusters 1 and 6 may contain patients whose conditions are more complex and difficult to generalize. In fact, as Table 5 shows, the most common

end point in cluster 1 has diagnoses of CKD stage 4, diabetes and hypertension, the most severe set of conditions out of the 6 clusters. All most common end points and high out-degree points had no procedures performed.

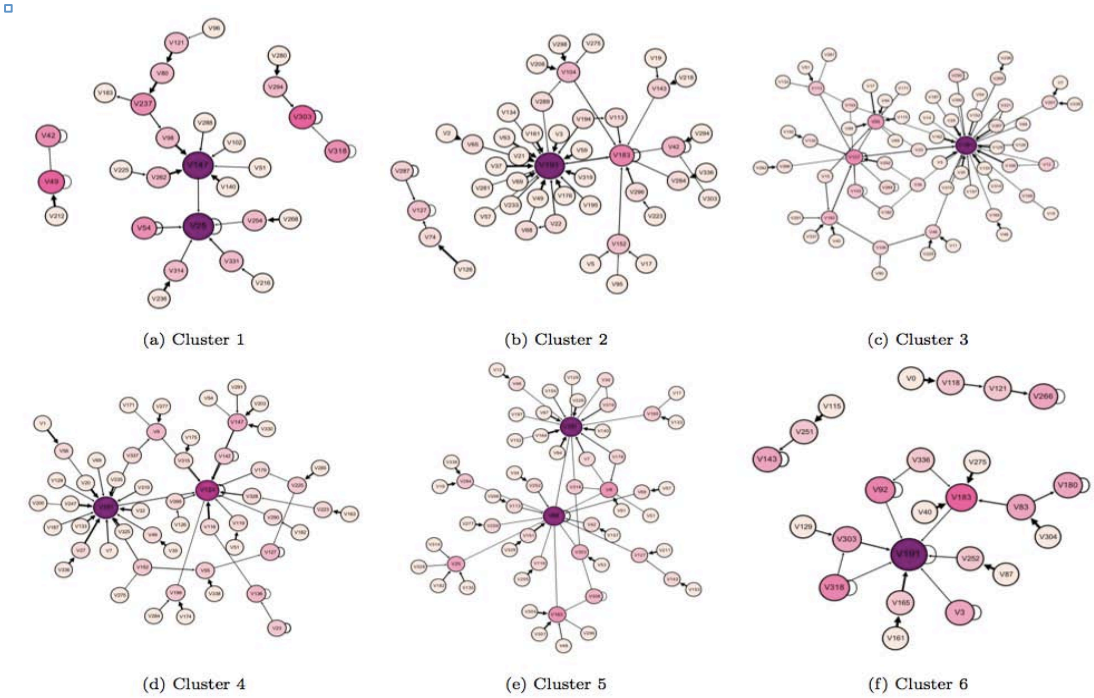


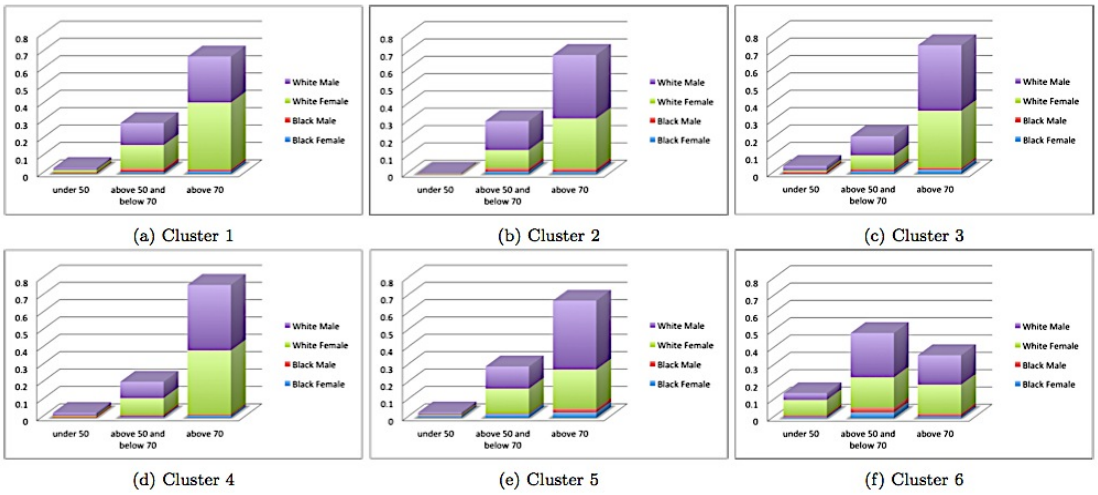
Figure 2. Pathway for identified clusters

Table 5. Summary of Clusters

Cluster	Number of Patients	Number of pathways	Most Common End point			Average maximum LCS[3]	High out-degree point	
			Type	Diagnosis	% of all pathways		Type	Diagnosis
1	433	17	FUP[1]	CKD stage 4, diabetes, hypertension	65%	1.8	HOSP[2]	AKF[4], CKD stage 2
2	462	40	FUP[1]	CKD stage 3, diabetes, hypertension	95%	2.8	FUP[1]	AKF[4], CKD stage 4
3	582	67	FUP[1]	CKD stage 3, hypertension	100%	3.0	NEW[5]	CKD stage2, hypertension
4	433	45	FUP[1]	CKD stage 4, hypertension	93%	3.0	FUP[1]	AKF[4], CKD stage 3, diabetes, hypertension
5	233	50	HOSP[2]	N/A	100%	3.2	FUP[1]	CKD stage 5, diabetes
6	368	16	FUP[1]	N/A	75%	2.0	FUP[1]	CKD stage 5, diabetes

[1] Follow up, [2] Hospital follow up, [3] Longest common subsequence, [4] Acute Kidney Failure, [5] New patient visit

We examined whether there are differences in terms of age, race, and sex (Figure 3). Applying Chi square test with Monte Carlo methods against the frequencies in the entire sample of 2,511 patients, we found significant differences in the expected and observed frequencies in clusters 2 through 6 (p-value < 0.05 using chisq.test in R)[26]. For example, prior to clustering, 1.7% of the sample population are African-American males between 50 and 70 years of age, but the percentage increases to 2.8% in cluster 6, and decreases to 0.4% in cluster 5 after clustering. Similar trend can also be seen among White female under 50 years of age. Also, the percentage of White female above 70 years of age in cluster 5 nearly doubles compared to the original percentage in the overall sample. Similar pattern can be seen with African-American male above 70 years old and African-American female above 70 years old. These subgroups may have higher likelihood of receiving hospital treatment.



(a) Cluster 1 (b) Cluster 2 (c) Cluster 3

(d) Cluster 4 (e) Cluster 5 (f) Cluster 6

Figure 3. Distribution of patients by age, race and sex

We illustrate 3 example pathways below. Figure 4 shows a pathway belonging to cluster 1. It starts with a follow up visit where patient is diagnosed with CKD stage 4, diabetes and hypertension. Patient receives a chest X-ray during the visit. On average at least 1 month later, patient visits the office for a pre-ESRD education, which is held by a nurse to prepare and educate patients about potential onset of ESRD. Patient comes for a follow up visit at least a month later on average, without changes in diagnoses. Patient is able to maintain the condition for at least a year on average, and visits the office for another follow up. This pathway's duration is on average 1 year and 2 months. Figure 5 shows a pathway in cluster 2, which starts with a CKD stage 3 patient attending an ESRD education session. After an average of 2 weeks, the patient is seen for a hospital follow up, where patient is diagnosed with AKF in addition to CKD stage 3. Then, in an average of 1 year, patient comes for an office follow up visit to receive diagnoses of diabetes and hypertension, in addition to CKD stage 3. Same visit is repeated one year later. On average, it takes at least 2 years 1 month and 2 weeks to complete the pathway. In Figure 5, patient avoids worsening of CKD for more than 2 years, in addition to the pathway in Figure 4 where patients avoid worsening for a year, indicating the known role of pre-ESRD education in helping patients to maintain their current conditions[7]. In fact, cluster 1's high out-degree point (Table 5) divides one pathway into two, with one ending with the same CKD stage upon pre-ESRD education, and another progressing to CKD 3, diabetes and hypertension. One of the visits in cluster 2 with high out-degree is a follow-up visit where patient is diagnosed with AKF and CKD stage 4, which leads to a visit of same type and diagnoses, or in another case an improvement to CKD stage 3 (Table 5). In both cases, they eventually end with a follow up visit with CKD stage 3, diabetes and hypertension. One pathway from cluster 3 is shown in Figure 6. It starts with a follow up visit with a CKD stage 1 patient, who maintains the condition for at least 2 years on average. Patient then develops hypertension, but maintains the same for at least 2 years on average. At least 4 years later on average, patient progresses to CKD stage 3, while still suffering from hypertension. Same diagnoses are noted on a follow up visit at least 1 year later on average. In total, it takes at least 5 years to complete the learned pathway. This is an example where the duration of the learned pathway is longer than the real time range in data, since it is created using high-likelihood transitions of multiple patients. Pathways like this may serve as a projection for future patients who share the same earlier visits.

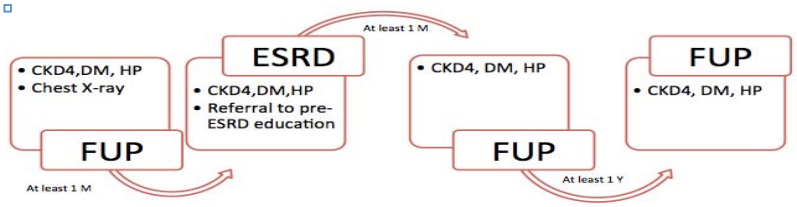


Figure 4. An example pathway in cluster 1
*CKD4: CKD stage 4, DM: diabetes, HP: hypertension, ** FUP: follow up, ESRD: pre-ESRD education

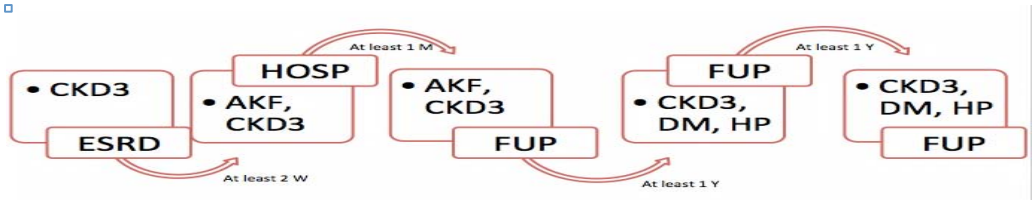


Figure 5. An example pathway in cluster 2
*CKD3: CKD stage 3, AKF: acute kidney failure, DM: diabetes, HP: hypertension
** ESRD: pre-ESRD education, HOSP: hospital follow up, FUP: follow up

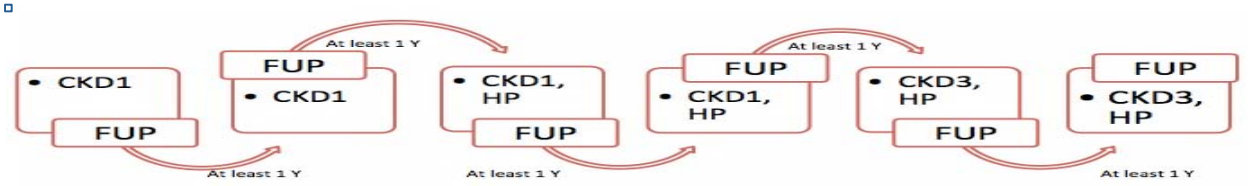


Figure 6. An example pathway in cluster 3
* CKD1: CKD stage 1, CKD3: CKD stage 3, HP: hypertension,** FUP: follow up

4. Discussion

Learning clinical pathways from EHR data allow us to discover improvement, deterioration or maintenance of the status quo in the course of disease evolution, so we may re-visit the historical data and clinicians' notes to identify interventions that may have played a role in the evolution. For instance, examining pathways with pre-ESRD education may allow us to compare the time it takes for patients to deteriorate with, or without, pre-ESRD education. Also, finding patterns in pathways where improvements were seen in patients' conditions may help clinicians to identify promising care models for a subgroup of the population. In addition, comparing the pathways of compliant and non-compliant patients may help clinicians to engage non-compliant patients in treatments using new approaches. Exploring clinical pathways also allow us to discover significant differences in the distribution of age, race and gender among clusters of patients. Learning about these differences may enable clinicians to design personalized treatment plans for patients according to demographics and health conditions, and develop hypotheses that can be tested in large populations.

In this paper, we chose to show all transitions of visits with weights above 0.3, but different thresholds of weight may produce pathways of different complexity. For example, the maximum length of pathways is 6, and the total number of pathways is 235 in this study. If we reduce the threshold of weight, we may be able to identify pathways of longer lengths, or larger total number of pathways. Developing a measure that takes into account the number of patients and their diversity may determine an optimal threshold of weight that allows most reasonable clinical interpretation of the pathways. Although clustering was applied to separate patients into subgroups, differences in clinical history still exist, even within one cluster, due to the diversity and complexity of patients' conditions and practice variations. Perhaps instead of using weight to elicit a sample population, a more individualistic approach, where we learn the clinical pathways based on patients' biochemical data, demographics and lifestyle choices, can help us to create personalized pathways and detect granular but important differences that separate one patient's sequence from another. Furthermore, in this study we studied only CKD stage 1 through 5, AKF, diabetes and hypertension as diagnoses to control for noise. The patient subgroups may be further divided into smaller subgroups if we included more conditions such as anemia and proteinuria, also commonly seen among CKD patients[1].

In this study we investigated 6 visit types, 8 conditions, and 27 procedures. Theoretically, there can be up to 1,296 combinations of visit types, conditions and procedures reflected in the visits. We observed 281 combinations from the data, as some combinations did not occur in actual visits. However, as the number of component we include increases, the number of possible combinations may explode computationally. For example, if we include 10 medications in the study, the number of possible combination increases by 10-fold. Alternatively, if we increase the conditions by 2, we double the number of possible combinations. As Table 4 shows, only a few of the 'Vz's are observed with high frequencies, making the distribution of 'Vz's right skewed with a long tail. In future studies where more components in the pathway need to be explored, we may focus only on the most frequent combinations, to avoid including too much noise in the pathway creation.

A big challenge in the study is the incomplete documentation in EHR. For example, clinicians may mention an existing condition only in their notes, and patients' medication lists may not be up-to-date. In order to gain the most value from EHR data, it is vital that treatments are accurately recorded with the help from healthcare providers. Once curated data is available, we should investigate complete treatment data including medications, and also patient demographics and life style choices if possible, since these are crucial factors in altering the disease evolution as well. Similarly, we recognize that the types of laboratory tests ordered characterize patients' current conditions and future treatment plans. Hence, results from important laboratory examinations such as GFR, serum Creatinine level, and serum Albumin level should also be included in the study data to provide a more accurate view of the patients' clinical history. This way, we can use laboratory results to ensure that there is no discrepancy between the diagnostic codes and actual conditions, and track pathway evolution by the sequence of lab observations. Moreover, if large enough sample is available, we should limit study subjects to be patients whose treatment data include all visits from the first to the last before patient progress to ESRD. This will likely allow us to build more comprehensive clinical pathways. Furthermore, although we did capture the duration of pathways in this study, temporal factor was not used to cluster patients. Temporal factors, such as pathway duration and the differences in time between actual practice and the consensus guidelines, should be explored in future analysis. This is an important objective that should be pursued with guidance from clinicians, so we may learn dividing points in the pathways that differentiate outcomes of treatments. Finally, and most importantly, rigorous evaluation by clinicians of the learned pathways is crucial to identify unrealistic transitions and courses of disease progression.

5. Conclusion

CKD is a costly and complex health condition with high mortality, and affects millions of adults worldwide. In this study, we aim to learn practice-based clinical pathways for CKD through which we seek an understanding of the major treatment pathways in the evolution of the disease. Our analysis of patient data from an EHR yielded 6 patient subgroups after applying hierarchical clustering on the sequences of office visits. Each subgroup has distinct characteristics, and captures commonly known combination of comorbidities in CKD, such as CKD stage 3, hypertension and diabetes. Also, patients who may experience hospital treatments, or progress faster than others, have been identified in subgroups. We observe that multiple paths of disease evolution exist even within each patient subgroup, confirming CKD's diverse course of development. Moreover, significant differences in the distribution of age, race and sex were found among subgroups. Clinical pathways learned from our study can be instantiated for any patient to enable shared decision making between clinicians and patients, so patients may gain insights into their projected clinical pathway based on demographics and earlier visits. At the same time, our study can serve as a practice management tool used by clinicians, to review their practice against current consensus guidelines for CKD, and identify care models that may lead to improved outcomes.

Acknowledgements

We are grateful to the physicians and staff of the community nephrology practice who generously provided data from their Electronic Health Record for this study. We particularly thank Dr. Teredesai, MD, Dr. Xie, MD, PhD, Dr. Patel, MD, and staff, L. Smith and A. Barletta, who gave us important clinical and technical information about the data and the key characteristics of CKD and its treatment. This study was designated as Exempt by the Institutional Review Board at Carnegie Mellon University.

References

1. US Renal Data System. USRDS 2013 Annual Data Report: Atlas of Chronic Kidney Disease and End-Stage Renal Disease in the United States. Bethesda, MD: National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases; 2013.
2. Jha V, Garcia-Garcia G, Iseki K, Li Z, Naicker S, Plattner B, Saran R, Wang AY, Yang CW. Chronic kidney disease: global dimension and perspectives. Lancet. 2013 Jul 20;382(9888):260-72.
3. Arulkumaran N, Montero RM, Singer M. Management of the dialysis patient in general intensive care. British journal of anaesthesia. 2012 Feb;108(2):183-92. PubMed PMID: 22218752.
4. National Kidney F. K/DOQI clinical practice guidelines for chronic kidney disease: evaluation, classification, and stratification. American journal of kidney diseases : the official journal of the National Kidney Foundation. 2002 Feb;39(2 Suppl 1):S1-266. PubMed PMID: 11904577.
5. Stevens PE, Levin A, Kidney Disease: Improving Global Outcomes Chronic Kidney Disease Guideline Development Work Group M. Evaluation and management of chronic kidney disease: synopsis of the kidney

disease: improving global outcomes 2012 clinical practice guideline. Annals of internal medicine. 2013 Jun 4;158(11):825-30. PubMed PMID: 23732715.
6. Abra G, Patel M, Moore D. Trend-bending Chronic Kidney Disease Care Model: Stanford University School of Medicine, Clinical Exellence Research Center; 2013 [cited 2014 03/01]. Available from: http://cerc.stanford.edu/fellowships/docs/CERC4modelsummary2.11.2013_3PMpdf.pdf.
7. Hayslip DM, Suttle CD. Pre-ESRD patient education: a review of the literature. Advances in renal replacement therapy. 1995 Jul;2(3):217-26. PubMed PMID: 7614358.
8. Israni A, Kasiske B. Laboratory assessment of kidney disease: glomerular filtration rate, urinalysis, and proteinuria. In: Taal MW CG, Marsden PA, editor. Brenner and Rector's The Kidney. 9 ed. Philadelphia, PA: Elsevier Saunders; 2011.
9. The National Kidney Foundation. Dialysis: The National Kidney Foundation; 2013 [cited 2014 3/10]. Available from: http://www.R-project.org/.
10. Grams ME, Chow EK, Segev DL, Coresh J. Lifetime incidence of CKD stages 3-5 in the United States. American journal of kidney diseases : the official journal of the National Kidney Foundation. 2013 Aug;62(2):245-52. PubMed PMID: 23566637. Pubmed Central PMCID: 3723711.
11. Coresh J, Selvin E, Stevens LA, Manzi J, Kusek JW, Eggers P, et al. Prevalence of chronic kidney disease in the United States. JAMA : the journal of the American Medical Association. 2007 Nov 7;298(17):2038-47.
12. Sud M, Tangri N, Levin A, Pintilie M, Levey AS, Naimark DM. CKD Stage at Nephrology Referral and Factors Influencing the Risks of ESRD and Death. American journal of kidney diseases : the official journal of the National Kidney Foundation. 2014 Jan 28. PubMed PMID: 24485146.
13. Tarver-Carr ME, Powe NR, Eberhardt MS, LaVeist TA, Kington RS, Coresh J, et al. Excess risk of chronic kidney disease among African-American versus white subjects in the United States: a population-based study of potential explanatory factors. Journal of the American Society of Nephrology : JASN. 2002 Sep;13(9):2363-70.
14. Lopes AA, Hornbuckle K, James SA, Port FK. The joint effects of race and age on the risk of end-stage renal disease attributed to hypertension. American journal of kidney diseases : the official journal of the National Kidney Foundation. 1994 Oct;24(4):554-60. PubMed PMID: 7942809.
15. Xu R, Wunsch DC 2nd. Clustering algorithms in biomedical research: a review. IEEE Rev Biomed Eng. 2010;3:120-54. doi: 10.1109/RBME.2010.2083647. Review.
16. Yoo I, Alafaireet P, Marinov M, Pena-Hernandez K, Gopidi R, Chang JF, Hua L. Data mining in healthcare and biomedicine: a survey of the literature. J Med Syst. 2012 Aug;36(4):2431-48.
17. Huang Z, Lu X, Duan H. On mining clinical pathway patterns from medical behaviors. Artif Intell Med. 2012;56(1).
18. Lin F, Hsieh L, Pan S. Learning Clinical Pathway Patterns by Hidden Markov Model. HICSS; 01/03/2005; Hawaii, USA2005. p. 142a.
19. Lin F, Chiu C, Wu S. Using Bayesian networks for discovering temporal-state transition patterns in Hemodialysis. HICSS; 1/7/2002; Hawaii, USA2002. p. 1995-2002.
20. Hsu CY, Chertow GM, McCulloch CE, Fan D, Ordonez JD, Go AS. Nonrecovery of kidney function and death after acute on chronic renal failure. Clinical journal of the American Society of Nephrology : CJASN. 2009 May;4(5):891-8. PubMed PMID: 19406959. Pubmed Central PMCID: 2676192.
21. Islam TM, Fox CS, Mann D, Muntner P. Age-related associations of hypertension and diabetes mellitus with chronic kidney disease. BMC nephrology. 2009;10:17. PubMed PMID: 19563681.
22. Zhang Y, Parman R, Wasserman L. On Learning Clinical Pathways for Chronic Kidney Disease from Electronic Health Record Data: A Preliminary Graphical Approach. 2014.
23. Kaufman L, Rousseeuw PJ. Finding Groups in Data: An Introduction to Cluster Analysis 1ed. New York: John Wiley; 1990.
24. Elzinga CH. Sequence analysis: Metric representations of categorical time series. . Socio- logical Methods and Research. 2008.
25. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol. 1995 Apr 7;247(4):536-40.
26. Rousseeuw PJ. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. Computational and Applied Mathematics 1987 (20):53-65.
27. Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media2009.
28. Harary F. Graph Theory. Reading, MA: Addison-Wesley; 1994.
29. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2012.

PubMedMiner: Mining and Visualizing MeSH-based Associations in PubMed

Yucan Zhang, PhD[1,2], Indra Neil Sarkar, PhD, MLIS[1,3,5], Elizabeth S. Chen, PhD[1,4,5]

[1]Department of Computer Science, [2]Plant Biology, [3]Microbiology & Molecular Genetics, [4]Medicine, [5]Center for Clinical & Translational Science, Univ. of Vermont, Burlington, VT

Abstract

The exponential growth of biomedical literature provides the opportunity to develop approaches for facilitating the identification of possible relationships between biomedical concepts. Indexing by Medical Subject Headings (MeSH) represent high-quality summaries of much of this literature that can be used to support hypothesis generation and knowledge discovery tasks using techniques such as association rule mining. Based on a survey of literature mining tools, a tool implemented using Ruby and R – PubMedMiner – was developed in this study for mining and visualizing MeSH-based associations for a set of MEDLINE articles. To demonstrate PubMedMiner's functionality, a case study was conducted that focused on identifying and comparing comorbidities for asthma in children and adults. Relative to the tools surveyed, the initial results suggest that PubMedMiner provides complementary functionality for summarizing and comparing topics as well as identifying potentially new knowledge.

Introduction

The MEDLINE citation database represents a major source of references to biomedical literature, containing over 20 million citations dating back to the late 1940s with 2000-4000 citations added daily[1, 2]. MEDLINE is the primary component of PubMed that provides access to over 23 million citations[1] where each record is associated with a set of metadata that includes title, abstract, and Medical Subject Headings (MeSH) descriptors that are used to index MEDLINE citations and enable searching in PubMed. With the exponential growth of published biomedical research available in resources such as PubMed (especially MEDLINE), it can be challenging for clinicians and researchers to keep current with findings on a topic of interest and to discover connections between seemingly unrelated concepts (e.g., as represented by MeSH descriptors) across multiple disciplines[3-5].

A number of systems have been developed that build upon the search capabilities provided by PubMed[4, 6, 7], including biomedical literature mining tools that incorporate information retrieval, entity recognition, information extraction, text mining, and data integration methods[8, 9]. As part of the literature-based knowledge or relation discovery process, a first step is the identification of biomedical concepts or entities such as diseases, drugs, and genes. While some studies have involved extracting concepts from titles and abstracts in PubMed or the full-text of articles in PubMed Central, others have focused on MeSH descriptors that are arranged in a hierarchical structure thus allowing for searching at different specificity levels (e.g., using a broader descriptor such as "Respiratory Tract Diseases" versus a more specific descriptor such as "Asthma")[7, 10-14]. These latter studies include those that have used MeSH descriptors to identify co-occurring biomedical concepts as well as those involving use of association rule mining techniques to discover putative relationships between biomedical concepts[15, 16]. There are more than 27,000 descriptors in MeSH that are continually revised and updated[17]. Since MeSH descriptors are applied to MEDLINE records by trained subject matter experts with domain knowledge, they can be seen as representing standardized and high-quality summaries of a particular publication[18]. Thus, MeSH descriptors may offer a useful window into the full text for information retrieval, extraction, and other high-level functions, and potentially allow for further automation of the discovery process.

The goal of this study was to build upon prior efforts and develop an open-source literature mining tool ("PubMedMiner") for identifying and visualizing relationships between MeSH descriptors in MEDLINE. The findings from a survey of literature mining tools that incorporate MeSH descriptors were used to guide the development of PubMedMiner to include functionality for PubMed searching, MeSH descriptor extraction, Unified Medical Language System (UMLS) semantic type filtering, basic statistical analysis and visualization, as well as association rule mining and visualization. In order to demonstrate how PubMedMiner could facilitate hypothesis generation and knowledge discovery, a case study is presented for exploring and comparing comorbidities for asthma in children and adults. A discussion of limitations of the current prototype system, challenges in its development, and planned enhancements to the system are then provided.

Background

Various literature and text mining tools have been developed to improve information retrieval and infer relationships between biomedical concepts[6, 12, 19, 20]. In this section, findings from a survey of literature mining tools including MeSH-based functions are summarized and association rule mining is described as a technique for exploring MeSH-based relationships.

Survey of Literature Mining Tools

Numerous studies have successfully revealed interesting scientific discoveries from biomedical literature in MEDLINE using a variety of knowledge discovery tools that incorporate different terminological systems (e.g., MeSH[5, 21], UniProt protein/gene names[22], or the UMLS[23]) and algorithms for discovering or predicting relations[24]. For example, literature-based discovery (LBD) approaches have been used to discover new knowledge and generate hypotheses based on findings in the literature, such as identifying connections between the beneficial effects of fish oil in treating Raynaud's syndrome and the causal effect of magnesium deficiency on migraines, both of which were later validated[25-28]. LBD studies have integrated MeSH descriptors as part of the knowledge discovery process[5, 21]. Other studies have focused on using MeSH descriptors to improve information retrieval and identify relationships between entities[25, 26, 29-31]. For example, relationships such as drug-gene, drug-effects, protein-protein, and gene-gene, have been explored using statistical co-occurrence methods[29, 31-33].

A survey of literature mining tools was conducted to identify and compare those that incorporate MeSH descriptors. Based on searching publications, Google and Google Scholar Web searching results, and specific resources (e.g., the National Network of Libraries of Medicine, National Center for Biotechnology Information, and Arrowsmith project sites), a total of 80 tools was identified[2, 4, 6, 12, 19, 34]. Of these tools, links for about one-third were found to be no longer active. For the remaining tools, those satisfying the following criteria were included: (1) is maintained and available to the public; (2) focuses on literature mining in the biomedical domain and builds upon basic PubMed functions; and, (3) incorporates MeSH descriptors in the implementation of advanced functionalities, such as basic statistical analysis and association rule mining. The 14 eligible tools after applying these criteria were found to use MeSH descriptors in their main functions besides searching and were categorized into three groups according to their most notable features: (1) filtering or clustering search results using MeSH: BibliMed[35] (2011; Private); (2) describing search results with basic statistics for MeSH: LigerCat[36] (2009; Academic), PubAnatomy[37] (2009; Academic), Anne O'Tate[38] (2008; Academic), GoPubMed[39] (2005; Private), MedSum[40] (2005; Academic), and PubMed PubReMiner[41] (2004; Academic); and, (3) exploring associations among MeSH: KNALIJ[42] (2012; Private), PubAtlas[43] (2009; Academic), AliBaba[44] (2006; Academic), BITOLA[45] (2005; Private), PubNet[46] (2005; Academic), XplorMed[47] (2001; Academic), and MEVA[48] (2001; Private).

The first group whose major feature is filtering or clustering search results using MeSH descriptors includes only one system (BibliMed) that provides a more focused search but no other advanced functions for information extraction[19]. Several systems in the other groups also have a clustering function, such as Anne O'Tate and XplorMed. Systems in the second group accept standard PubMed queries for literature search and perform frequency analysis on MeSH descriptors from retrieved records (Table 1). However, only one or two systems support search result export, statistical analysis, or visualization. LigerCat[20] was the only open-source system identified in this survey. In the third group (Table 1), systems either carry out association rule mining directly among MeSH descriptors or find connections between publications through common MeSH descriptors. Systems such as KNALIJ and AliBaba visualize associations as a network between extracted entities, whereas systems such as PubAtlas visualize associations as a two-dimensional table. While XplorMed and MEVA do not allow for direct PubMed searching, users can upload files produced from external PubMed searches that are used for subsequent analysis. Only a few of the surveyed systems allow users to adjust parameters for association rule mining, and some tools cannot visualize discovered associations.

None of the systems in the second group was found to perform statistical analysis of UMLS semantic types associated with MeSH descriptors and only two of the tools from the third group have implemented the feature of filtering MeSH descriptors by semantic type. UMLS semantic types are hierarchical subject categories for categorizing concepts in the UMLS Metathesaurus[49] and have been used to cluster or filter search results. There are 133 semantic types currently and each MeSH descriptor can be associated with one or more semantic type(s)[17, 50]. This allows users to focus on associations between MeSH descriptors of interest.

Table 1. Tools with basic statistical analysis and association rule mining functions using MeSH descriptors.

Tool	Query format	Filtering with semantic types	MeSH as entity	Search result export	MeSH frequency	Semantic type frequency	Statistics visualization	Association rule mining	Rule strength cutoff adjustable	Association visualization	Related article links	Analysis result export
LigerCat	Keywords, Gene/DNA Sequence	×	✓	×	✓	×	✓	×	×	×	✓	×
PubAnatomy	Keywords, Gene ID or Symbol	✓	✓	✓	✓	×	×	×	×	×	✓	✓
Anne O'Tate	Keywords	✓	✓	×	✓	×	×	×	×	×	✓	×
GoPubMed	Keywords	×	✓	✓	✓	×	×	×	×	×	✓	×
MedSum	Keywords	×	✓	×	✓	×	×	×	×	×	×	×
PubReMiner	Keywords, Gene	×	✓	×	✓	×	×	×	×	×	✓	✓
KNALIJ	Keywords	×	✓	×	×	×	×	✓	×	✓	✓	×
PubAtlas	Keywords, predefined terms	×	×	×	×	×	×	✓	×	✓	✓	✓
AliBaba	Keywords	×	✓	×	×	×	×	✓	✓	✓	✓	✓
BITOLA	MeSH, Gene symbols	✓	✓	×	×	×	×	✓	✓	×	✓	×
PubNet	Keywords	×	×	×	×	×	×	✓	×	✓	×	✓
XplorMed	NA	×	×	NA	✓	×	×	✓	✓	×	✓	×
MEVA	NA	✓	✓	NA	✓	×	×	✓	×	×	✓	×

Note: NA, not available.

Association Rule Mining

The concept of association rule mining was proposed by Agrawal, *et al.* in 1993[51] in the context of market basket analysis, but has since been widely adapted to other domains[30, 52-54]. An association rule is defined as an implication of the form: $X \rightarrow Y$, where $X, Y \subset I$, $X \cap Y = \varnothing$, X, Y are sets of items ("itemsets"), I is the set of total items, and $X \rightarrow Y$ means that when X occurs, Y also occurs with a certain probability. Agrawal and Srikant proposed an algorithm for mining association rules based on identifying frequent itemsets, known as the *Apriori* algorithm[55].

The strength of an association rule is typically measured using metrics referred to as support and confidence[51]. The *support* of an itemset X represents the fraction of transactions containing itemset X. The *confidence* of a rule is the probability of finding transactions containing the consequent of a given rule that also contain the antecedent of the given rule. Commonly, in order to obtain association rules with statistical significance and certain strength, minimum support and confidence values are specified[51]. Many other measures of interestingness have been proposed, such as the chi-squared (χ^2) statistic, lift measure, and Gini index[56, 57]. Of note, the χ^2 statistic has been shown as an efficient measure of the strength of a set of association rules based on co-occurrence[30, 58, 59] and found to outperform other measures such as support, confidence, lift, and conviction[60].

A number of open-source tools are available that provide a wide variety of statistical and graphical techniques for generating and visualizing association rules, such as R[61], Tanagra[62], and Weka[63, 64]. R is a widely used free software environment for statistical computing and visualization that can be extended using over 5000 packages available at the Comprehensive R Archive Network (CRAN) package repository[65, 66]. Compared with R, Tanagra lacks advanced graphical features and Weka is weaker in calculating classical statistics[64, 67, 68]. The R package "arules" implements the *Apriori* algorithm for association rule mining and can calculate various interestingness measures for generated rules[69]. The "arulesViz" R package visualizes association rules in different formats such as scatter plot, grouped matrix, and graph-based association network[70].

Materials and Methods

Overview

In this study, the combined functionality provided by the aforementioned literature mining tools was used to guide the development of an open-source literature mining tool ("PubMedMiner") to include: PubMed searching, MeSH descriptor extraction, UMLS semantic type filtering, basic statistical analysis and visualization, association rule mining and visualization, as well as results exporting. With standard searches in PubMed, MeSH descriptors extracted from retrieved literature reflect the contents of the corresponding articles. Filtering by UMLS semantic types is one way to focus on particular MeSH descriptors of interest. Basic statistical analysis such as frequency counts helps to

identify the most common concepts and semantic types to inform mining and visualization of associations among selected MeSH descriptors.

The general workflow of PubMedMiner involves four phases (Figure 1): (1) searching and retrieving MEDLINE records associated with a particular topic from PubMed; (2) filtering MeSH descriptors by UMLS semantic type(s); (3) generating basic statistics for MeSH descriptors and UMLS semantic types; and, (4) mining and visualizing association rules between filtered MeSH descriptors. Results associated with each phase are made available as plain text (ASCII), XML, or PDF files.

The Ruby scripting language was used to develop the interactive, command-line version of PubMedMiner that is configurable at each phase. For the statistical analysis component, the R environment for statistical computing and graphics[71] was integrated into PubMedMiner using the RinRuby[72] Ruby gem. Functionality for basic statistics and association rule mining was enabled through the use of the "arules"[66] and "arulesViz"[66] R packages.

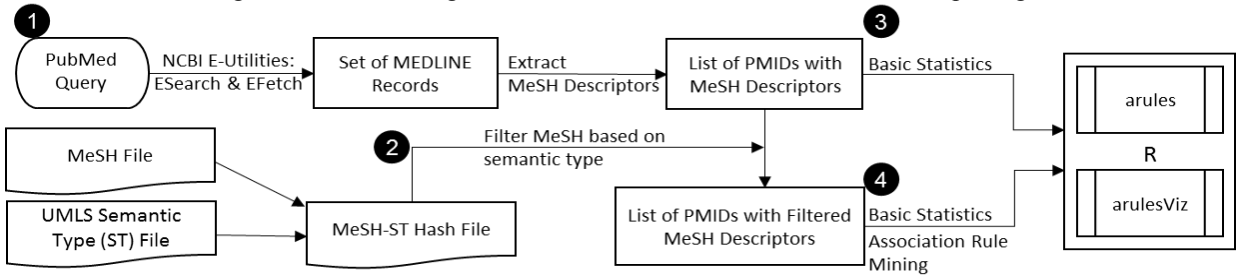


Figure 1. Overview of PubMedMiner development process.

Searching and Retrieving MEDLINE Records

PubMedMiner makes use of the National Center for Biotechnology Information Entrez Programming Utilities (NCBI E-Utilities)[73] to search and retrieve records in MEDLINE format from PubMed. Within PubMedMiner, a user can enter a PubMed query to search and retrieve MEDLINE-formatted records and is notified if no matching articles are found. At the end of a search, PubMedMiner parses each retrieved MEDLINE record and extracts the PMID and MeSH descriptors based on the corresponding PMID and MH metadata fields. Only the PMID and all associated MeSH descriptors along with subheadings are saved into a new text file for the next phase.

Filtering MeSH descriptors by UMLS Semantic Types

UMLS semantic types are used by PubMedMiner to filter MeSH descriptors for subsequent association rule mining. This step removes irrelevant associations and only focuses on association rules between MeSH descriptors corresponding to the set of semantic types specified by the user. Each MeSH descriptor and associated UMLS semantic types are extracted and stored as a hash table in a YAML (YAML Ain't Markup Language) text file, which links the semantic type codes to their full names according to a downloadable list of UMLS Semantic Groups that specifies the semantic types included in each group[74]. For example, the MeSH descriptor "Asthma" is associated with the UMLS semantic type code "T047" for "Disease or Syndrome."

Generating Basic Statistics for MeSH Descriptors and UMLS Semantic Types

The PMID-MeSH and PMID-semantic type data are transformed into the requisite formats that enable them to be analyzed in R using the "arules" package (version 1.0-15)[66]. Both absolute counts and relative frequencies of each MeSH descriptor and UMLS semantic type are saved in a text file in a tabular format. The visualization of frequencies as histograms is also implemented using "arules." The user can specify the number of MeSH descriptors or UMLS semantic types to be viewed in the histogram graphs (e.g., top 10 or top 25). Since each MeSH descriptor may correspond to multiple UMLS semantic types, duplicates of the same semantic type can exist for a single MEDLINE record. To accommodate for this, statistics are obtained for semantic types both with and without duplicates. PubMedMiner can perform statistical analysis and visualization on MeSH descriptors and UMLS semantic types both before and after filtering.

Mining and Visualizing Association Rules

In PubMedMiner, the "arules" R package that implements the *Apriori* algorithm is used for generating association rules among selected MeSH descriptors[66, 69] and the "arulesViz" R package is used for visualizing these rules (version 0.1-7)[66, 70]. Before the rule mining step, the selected MeSH descriptors are transformed into the appropriate format for R. The user needs to specify minimum support and confidence values at the beginning of the mining process.

Association rules are generated with 17 different interestingness measures that include support, confidence, χ^2, and Gini index, which are made available in a pipe-delimited text file. In addition, links to PubMed are generated that includes the consequent and antecedent of each rule as search terms in the query: "<rule consequent>"[mh] AND "<rule antecedent>"[mh]. Rules can be visualized and saved as PDF files in five different formats: (1) scatter plot; (2) matrix-based visualization; (3) group matrix-based visualization; (4) graph-based visualization with itemsets as vertices; and, (5) graph-based visualization with items and rules as vertices. For the graph-based visualizations, users can specify the number of rules to include (e.g., all or top 20). To view larger amounts of rules (maximum 1000 rules), a GraphML formatted file is generated that can be opened and modified by visualization tools such as Gephi and Cytoscape[75, 76]. Users can choose whether or not to re-run the rule mining function with adjusted parameters (i.e., support and confidence) if previously defined values did not give meaningful results.

Results

PubMedMiner

PubMedMiner is available under a GPL2 license at: https://github.com/UVM-BIRD/pubmedminer. The PubMedMiner tool provides several modes that cover functions for searching, filtering, statistics before and after filtering, and association rule mining. Each mode represents a combination of different functions and the user can specify which set of functions to be executed by the system. There are three main modes: (1) search, filter, and rule mining (includes statistics); (2) filter and rule mining (includes statistics); and, (3) rule mining only. The first main mode covers the whole process of the pipeline while the other modes only execute part of the process. Four additional modes provide individual functionality: (1) search only; (2) filter only; (3) statistics only (before filtering); and, (4) statistics only (after filtering). A PubMedMiner log file retains a record of the timestamp, selected mode, and configuration details such as the user-specified search query, UMLS semantic types, and other parameters for the basic statistical analysis and association rule mining.

Case Study: Exploring and Comparing Comorbidities for Asthma in Children and Adults

To demonstrate the use of PubMedMiner, a case study is presented here for how the tool can be used to facilitate exploration and comparison of comorbidities for asthma in either a pediatric or adult population. The respective queries (performed on February 13th, 2014) for these patient populations were:

- Pediatric Asthma: "asthma"[mh] AND ("infant"[mh] OR "child"[mh] OR "adolescent"[mh]) NOT "adult"[mh]
- Adult Asthma: "asthma"[mh] AND "adult"[mh] NOT ("infant"[mh] OR "child"[mh] OR "adolescent"[mh])

The first main mode, as described above, was used by PubMedMiner for both queries where the tool performed the PubMed search, filtered and analyzed extracted MeSH descriptors, and generated association rules. The pediatric asthma search returned 23,448 results containing 6,524 unique MeSH descriptors with 123 semantic types; whereas, the adult asthma search returned 22,839 articles containing 8,688 unique MeSH descriptors with 127 semantic types. Absolute counts and relative frequencies for all MeSH descriptors and UMLS semantic types were calculated, and the top 25 MeSH descriptors and top ten semantic types ordered by frequency were chosen to be displayed as histograms in separate PDF files. MeSH descriptors were subsequently filtered by the following three UMLS semantic types in the UMLS Semantic Network: (1) "Disease or Syndrome," (2) "Mental or Behavioral Dysfunction," and (3) "Neoplastic Process," where the latter two are children of the first[50]. After the filtering step, 669 MeSH descriptors were left for pediatric asthma and 948 MeSH descriptors remained for adult asthma. The tool carried out basic statistical analysis again on these filtered MeSH descriptors and further generated association rules using user-specified parameters. To maximize the rules generated by PubMedMiner, a minimum support of 0.01, minimum confidence of 0.01, and maximum rule length of 3 were specified as the settings for both example runs in this study.

According to the statistical analysis of the extracted UMLS semantic types, most semantic types in both sets of articles had similar frequencies (< 2 fold difference). For example, both sets of articles included the same top five UMLS semantic types ordered by absolute count without duplicates: "Age Group," "Population Group," "Human," "Disease or Syndrome," and "Organism Attribute." Several semantic types had frequencies with greater than a two-fold difference. For example, the frequency of the semantic types "Family Group," "Environmental Effect of Humans," "Conceptual Entity," "Organization," and "Regulation or Law" were higher in publications related to pediatric asthma; whereas "Anatomical Abnormality," "Organophosphorus Compound," "Body Substance," "Cell Function," and "Neoplastic Process" appeared more often in articles for adult asthma.

Similarly, the frequency patterns of MeSH descriptors after filtering differed between the two sets of articles. Table 2 presents a selection of MeSH descriptors that had relative frequencies greater than 2%. The fold differences of relative

frequencies were calculated for MeSH descriptors that were higher than 0.5% between the two sets, and fold differences greater than two are highlighted in Table 3, which shows MeSH descriptors with fold difference greater than five. Several MeSH descriptors were more frequently mentioned for pediatric asthma such as "Attention Deficit Disorder with Hyperactivity," "Virus Diseases," and "Respiratory Syncytial Virus Infections," and "Bronchiolitis, Viral" was only found in publications related to pediatric asthma. Some descriptors were more frequent for adult asthma such as "Asthma, Occupational," "Pulmonary Disease, Chronic Obstructive," "Hypertension," "Lung Diseases, Obstructive," and "Churg-Strauss Syndrome" (Table 3). This type of analysis enables one to explore potential asthma-related diseases and reveals those that are more studied relative to the two patient populations (children and adults).

Table 2. MeSH descriptors after filtering by UMLS semantic types.

Pediatric Asthma			Adult Asthma		
MeSH Descriptor	**Absolute Count**	**Relative Frequency**	**MeSH Descriptor**	**Absolute Count**	**Relative Frequency**
Asthma	9761	97.61%	Asthma	9692	97.20%
Dermatitis, Atopic	390	3.90%	Pulmonary Disease, Chronic Obstructive	738	7.40%
Rhinitis, Allergic, Perennial	389	3.89%	Occupational Diseases	667	6.69%
Acute Disease	357	3.57%	Bronchial Hyperreactivity	521	5.23%
Rhinitis, Allergic, Seasonal	329	3.29%	Chronic Disease	408	4.09%
Bronchial Hyperreactivity	311	3.11%	Rhinitis	310	3.11%
Respiratory Tract Infections	299	2.99%	Acute Disease	300	3.01%
Chronic Disease	292	2.92%	Rhinitis, Allergic, Perennial	270	2.71%
Rhinitis	275	2.75%	Rhinitis, Allergic, Seasonal	246	2.47%
Eczema	259	2.59%	Eosinophilia	214	2.15%

Table 3. Comparison of MeSH descriptor frequencies for pediatric asthma and adult asthma.

Pediatric Asthma		Adult Asthma	
MeSH Descriptor	**Fold Difference**	**MeSH Descriptor**	**Fold Difference**
Attention Deficit Disorder with Hyperactivity	53.00	Asthma, Occupational	83.00
Virus Diseases	18.80	Occupational Diseases	47.64
Respiratory Syncytial Virus Infections	14.33	Pulmonary Disease, Chronic Obstructive	29.52
Bronchiolitis	8.00	Hypertension	13.57
Eczema	7.40	Lung Diseases, Obstructive	11.67
Dermatitis, Atopic	6.61	Churg-Strauss Syndrome	10.50
Respiratory Tract Infections	5.34	Lung Neoplasms	8.43
Bronchiolitis, Viral	*	Drug Hypersensitivity	7.05
		Aspergillosis, Allergic Bronchopulmonary	5.20

* MeSH descriptor "Bronchiolitis, Viral" is only associated with pediatric asthma.

With a minimum support value of 0.01 and minimum confidence value of 0.01, 35 and 34 association rules were generated for articles related to asthma in children and adults respectively. The resulting text file included the rules along with the different interestingness measures sorted by χ^2 value in descending order and links to PubMed (e.g., http://www.ncbi.nlm.nih.gov/pubmed/?term="Dermatitis, Atopic"[mh] and "Asthma"[mh] for the rule "Dermatitis, Atopic => Asthma"). The five types of graphs were generated along with the GraphML file for visualizing the rules. Figure 2 shows the grouped matrix-based visualization of all the association rules where rules with the top *k* (user-specified) consequent (Right-Hand Side [RHS]) MeSH descriptors sharing common antecedents (Left-Hand Side [LHS]) are grouped together. If the LHS contains more than one MeSH descriptor, only the most frequent MeSH descriptor is shown in the column label and the number of other descriptors is shown as "+*N*" where *N* represents the number. Size and color represent support and χ^2 value respectively. The columns and rows are ordered according to the χ^2 value that is decreasing from top down and from left to right so that the group of most interesting rules according to χ^2 value are shown in the top-left corner of the plot. Numbers below LHS represent the number of rules containing the corresponding LHS MeSH descriptors.

In the graph-based visualization in Figure 3, MeSH descriptors and rules are represented as vertices, where the size of the rule vertex represents the support value of the rule and color denotes the χ^2 value (e.g., darker shades corresponds to higher values). Bi-directional association rules are identified between pairs of associated MeSH descriptors; the graph contains two rules for each pair with the same strengths in opposite directions. This type of graph emphasizes the composition of the rules and shows which MeSH descriptors share the same rule. Examination of disease-disease

associations from articles for pediatric asthma and adult asthma further highlights several diseases as possible common comorbidities of asthma in both populations such as "Bronchitis," "Bronchial Hyperreactivity," and "Rhinitis," whereas several diseases may be more common in one population than the other, such as "Eczema," "Dermatitis, Atopic," and "Food Hypersensitivity" in children, and "Airway Obstruction," "Churg-Strauss Syndrome," and "Eosinophilia" in adults. In addition to these pairwise relationships, the visualizations for pediatric asthma reveal relationships between "Rhinitis, Allergic, Perennial" and "Dermatitis, Atopic" as well as triple associations such as between "Asthma," "Rhinitis, Allergic, Seasonal," and "Rhinitis, Allergic, Perennial".

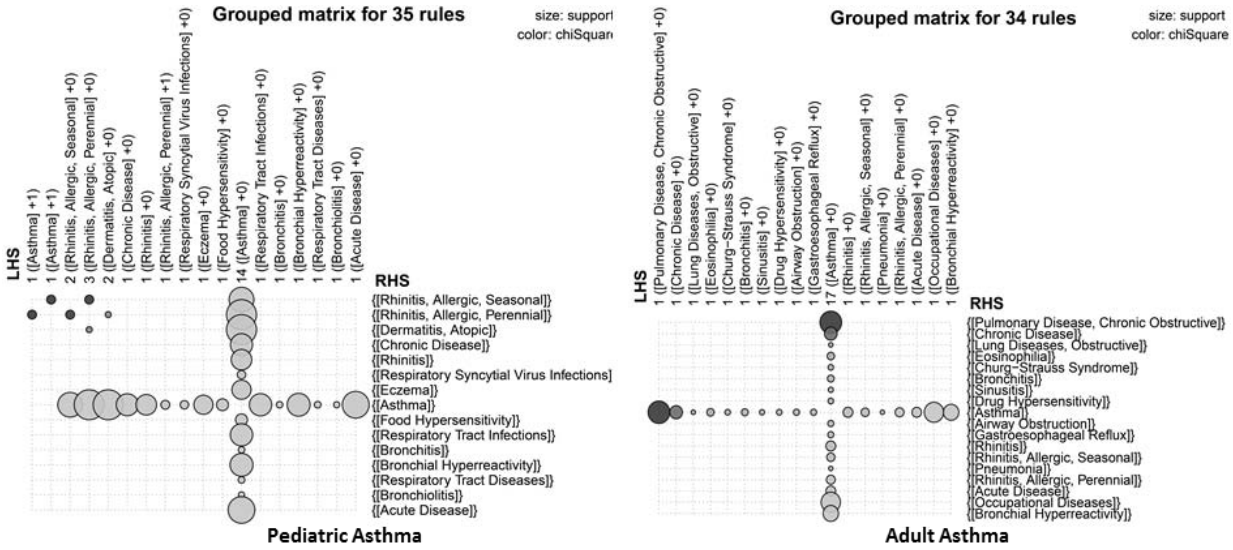


Figure 2. Grouped matrix-based visualization of all association rules for pediatric and adult asthma.

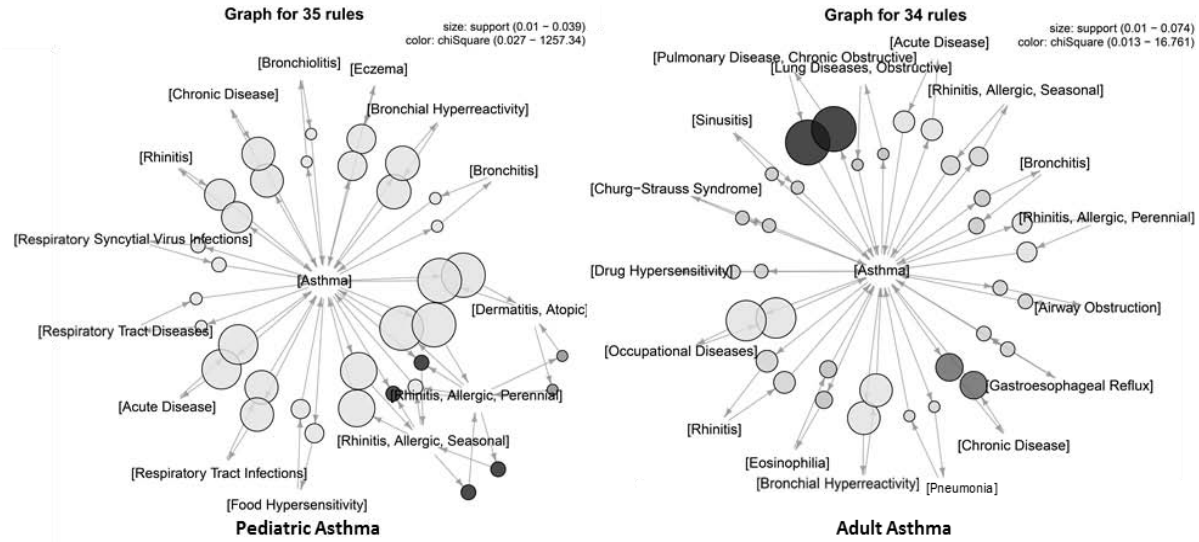


Figure 3. Graph-based visualization with items (i.e., MeSH descriptors) and rules as vertices of all association rules.

Discussion

PubMedMiner was developed as a prototype literature mining tool for MeSH descriptor extraction, basic statistical analysis, and association rule mining of biomedical literature for a given topic. There are several currently available literature mining tools that are capable of advanced literature retrieval and literature-based discovery. Based on the survey and comparison of existing tools including functionality for MeSH descriptors, a command line version of PubMedMiner was developed, which incorporates PubMed searching, MeSH descriptor extraction and filtering, basic statistical analysis, association rule mining, and visualization.

PubMedMiner can be used to explore MeSH descriptors and UMLS semantic types of interest for enhancing focused literature search as well as to validate existing and potentially discover new clinical knowledge. This latter functionality could be valuable for assisting clinicians and researchers keep up-to-date with changes in disease knowledge (e.g., disease-disease and disease-drug) over time. The case study demonstrated a more advanced use of this tool in which it helped highlight the similarities and differences between two PubMed searches for asthma in children and asthma in adults. This example shows how PubMedMiner may be used to facilitate a comorbidity analysis of asthma in different populations. While these searches excluded articles that include MeSH descriptors for both children and adults, further analysis could involve comparing comorbidities based on asthma articles for both children and adults (e.g., "asthma"[mh] AND "adult"[mh] AND ("infant"[mh] OR "child"[mh] OR "adolescent"[mh])). In addition, future work may include formally evaluating the results from such analyses with domain experts (e.g., clinicians and biomedical researchers) to confirm if the associations represent known or unknown knowledge, exploring how this tool may be beneficial for providing highly summarized domain knowledge for researchers and clinicians, guiding studies on particular conditions such as asthma in different populations, and validating results from clinical data mining studies (e.g., comorbidities based on data from the electronic health record).

For a given topic, there could be thousands of MeSH descriptors associated with a set of retrieved MEDLINE citations. PubMedMiner can carry out association rule mining among all MeSH descriptors or only filtered ones according to user-specified UMLS semantic types. Continued development of PubMedMiner will include advanced filtering functions such as incorporating hierarchical information (e.g., enabling the selection of MeSH descriptors for sub-categories of a specified UMLS semantic type) and considering other groupings of MeSH descriptors (e.g., leveraging the MeSH hierarchies or UMLS Semantic Groups). Additionally, the filtering functionality of PubMedMiner may be further enhanced by allowing users to filter out particular MeSH descriptor(s) before generating the statistics and rules in order to highlight relationships between other descriptors (e.g., excluding the original search descriptor "Asthma" and its descendants to highlight other co-morbidities). Other statistical approaches like TF-IDF could also be used to filter out frequently occurring MeSH descriptors (e.g., check tags such as "Humans," "Adult," and "Child").

Limitations include exclusion of newer citations due to the delay in assigning MeSH descriptors[77] and potential bias of selected MeSH descriptors due to consistency issues between different indexers[78, 79]. Besides MeSH descriptors, there are many other metadata fields in MEDLINE records, such as authors, publication types, titles, and abstracts. A number of tools have incorporated such metadata into their functionalities such as PubReMiner that calculates basic statistics on year, author, journal, and substances[19], and KNALIJ that generates and visualizes associations between authors, journals, and latest citations[42]. Future extensions to PubMedMiner may include combining basic statistics with association rule mining using other metadata for a given topic, and enhancing the current association rule mining functionalities by exploring biomedical concepts in titles and abstracts.

Using open-source technologies reduced the costs in terms of labor and time for implementing PubMedMiner. Association rule mining and visualization have been implemented in the "arules" and "arulesViz" R packages. In order to exploit the functions of these packages, R was integrated into Ruby (using the RinRuby Ruby gem) in the implementation of PubMedMiner and several limitations were encountered. Since the generated graphs are not self-descriptive, one motivation for creating the PubMedMiner log file was to record details about each use of the tool, including the user-specified parameters for creating all the graphs. Graphs have to be manually labeled when one needs to compare the same type of graphs generated with different queries. In addition, incomplete labels might be shown on graphs due to the size limitation when visualizing relatively large amounts of information. Twenty or fewer rules give the best resolution of node labels in the network graphs saved in PDF format and network graphs saved in the GraphML file format can be used to visualize up to 1000 rules. Future improvement of this tool may include enhancing interactive features by allowing users to set the font of node labels as well as the title of each graph. Furthermore, current rule metrics include 17 different interestingness measures that can assist users in selecting the most significant association rules. However, visualization of generated rules is implemented with only support, confidence, and χ^2 values. In future development, the tool could allow users to specify what set of measure(s) to use in visualization and incorporate a preview function based on user needs.

For the asthma case study, the frequency of MeSH descriptors was compared manually by calculating the fold difference of the same MeSH descriptors in two different queries. Allowing users to compare results from different queries is an additional functionality that may be included as a future enhancement in PubMedMiner. Although the command line version of PubMedMiner may satisfy many of the needs of studies such as described here, a graphical user interface (GUI) would likely be more desirable for the interactive process. The prototype system developed in this study has provided the foundation for implementation of a web-based GUI version of PubMedMiner (e.g., using MySQL and Ruby on Rails) that is publicly accessible and includes enhanced functionality.

Conclusion

MeSH descriptors can be seen as high quality summaries of biomedical literature and have been used as a window to literature indexed in MEDLINE. Guided by a survey of currently available MeSH-based literature mining tools, a command line prototype system (PubMedMiner) was developed that includes functionality for PubMed searching, MeSH descriptor extraction, UMLS semantic type filtering, basic descriptive statistics, association rule mining, and visualization. Based on a case study of comparing pediatric and adult asthma literature, PubMedMiner was able to identify association rules that may represent characteristic comorbidities.

Acknowledgements

The work was supported in part by the National Library of Medicine of the National Institutes of Health under award number R01LM011364. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Reference

1. U.S. National Library of Medicine Fact Sheets. Available from: http://www.nlm.nih.gov/pubs/factsheets/
2. Arrowsmith. Available from: http://arrowsmith.psych.uic.edu/arrowsmith_uic/index.html
3. Srinivasan P. Text mining: generating hypotheses from MEDLINE. Journal of the American Society for Information Science and Technology. 2004;**55**(5):396-413.
4. Jensen LJ, Saric J, Bork P. Literature mining for the biologist: from information retrieval to biological discovery. Nature reviews genetics. 2006;**7**(2):119-29.
5. Agarwal P, Searls DB. Literature mining in support of drug discovery. Briefings in bioinformatics. 2008;**9**(6):479-92.
6. Weeber M, Kors JA, Mons B. Online tools to support literature-based discovery in the life sciences. Briefings in bioinformatics. 2005;**6**(3):277-86.
7. Cohen AM, Hersh WR. A survey of current work in biomedical text mining. Briefings in bioinformatics. 2005;**6**(1):57-71.
8. De Bruijn B, Martin J. Getting to the (c) ore of knowledge: mining biomedical literature. International journal of medical informatics. 2002;**67**(1):7-18.
9. Hearst M. What is text mining. Retrieved February. 2003;**7**:2011.
10. Yu H, Agichtein E. Extracting synonymous gene and protein terms from biological literature. Bioinformatics (Oxford, England). 2003;**19**(suppl 1):i340-i9.
11. Cohen AM, Hersh WR, Dubay C, Spackman K. Using co-occurrence network structure to extract synonymous gene and protein names from MEDLINE abstracts. BMC bioinformatics. 2005;**6**(1):103.
12. Lu Z. PubMed and beyond: a survey of web tools for searching biomedical literature. Database: the journal of biological databases and curation. 2011.
13. Cohen KB, Johnson H, Verspoor K, Roeder C, Hunter L. The structural and content aspects of abstracts versus bodies of full text journal articles are different. BMC bioinformatics. 2010;**11**(1):492.
14. Lin J. Is searching full text more effective than searching abstracts? BMC bioinformatics. 2009;**10**(1):46.
15. Srinivasan P, Hristovski D. Distilling conceptual connections from MeSH co-occurrences. Medinfo. 2004;**11**(Pt 2):808-12.
16. Hristovski D, Stare J, Peterlin B, Dzeroski S. Supporting discovery in medicine by association rule mining in Medline and UMLS. Studies in health technology and informatics. 2001(2):1344-8.
17. Medical Subject Headings. Available from: http://www.nlm.nih.gov/mesh/
18. Bhattacharya S, Ha V, Srinivasan P. MeSH: a window into full text for document summarization. Bioinformatics (Oxford, England). 2011;**27**(13):i120-i8.
19. NCBI literature search tool archive. Available from: http://www.ncbi.nlm.nih.gov/CBBresearch/Lu/search/
20. NCBI text mining tools. Available from: http://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/tmTools/
21. Srinivasan P, Libbus B. Mining MEDLINE for implicit links between dietary substances and diseases. Bioinformatics (Oxford, England). 2004;**20**(suppl 1):i290-i6.
22. Rebholz-Schuhmann D, Kirsch H, Arregui M, Gaudan S, Riethoven M, Stoehr P. EBIMed—text crunching to gather facts for proteins from Medline. Bioinformatics (Oxford, England). 2007 January 15, 2007;**23**(2):e237-e44.
23. Kilicoglu H, Fiszman M, Rodriguez A, Shin D, Ripple A, Rindflesch TC. Semantic MEDLINE: a web application for managing the results of PubMed Searches. Proceedings of the third international symposium for semantic mining in biomedicine; 2008; 2008. p. 69-76.
24. Chen H, Sharp B. Content-rich biological network constructed by mining PubMed abstracts. BMC bioinformatics. 2004;**5**(1):147.
25. Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspectives in biology and medicine. 1986;**30**(1):7.
26. Swanson DR. Migraine and magnesium: eleven neglected connections. Perspectives in biology and medicine. 1987;**31**(4):526-57.
27. DiGiacomo RA, Kremer JM, Shah DM. Fish-oil dietary supplementation in patients with Raynaud's phenomenon: a double-blind, controlled, prospective study. The American journal of medicine. 1989;**86**(2):158-64.
28. Schiapparelli P, Allais G, Gabellari IC, Rolando S, Terzi MG, Benedetto C. Non-pharmacological approach to migraine prophylaxis: part II. Neurological Sciences. 2010;**31**(1):137-9.
29. Xu R, Wang Q. Toward creation of a cancer drug toxicity knowledge base: automatically extracting cancer drug–side effect relationships from the literature. Journal of the American Medical Informatics Association. 2013.
30. Chen ES, Hripcsak G, Xu H, Markatou M, Friedman C. Automated Acquisition of Disease–Drug Knowledge from Biomedical and Clinical Documents: An Initial Study. Journal of the American Medical Informatics Association. 2008 1//;**15**(1):87-98.
31. Xiang Z, Qin T, Qin Z, He Y. A genome-wide MeSH-based literature mining system predicts implicit gene-to-gene relationships and networks. BMC Systems Biology. 2013;**7**(Suppl 3):S9.
32. Xu R, Wang Q. A knowledge-driven conditional approach to extract pharmacogenomics specific drug–gene relationships from free text. Journal of biomedical informatics. 2012;**45**(5):827-34.
33. Blaschke C, Andrade MA, Ouzounis CA, Valencia A. Automatic extraction of biological information from scientific text: protein-protein interactions. Ismb; 1999; 1999. p. 60-7.
34. Medicine NNoLo. PubMed Online and App Resources. Available from: http://nnlm.gov/training/resources/pubmedalt.html
35. BibliMed. Available from: http://www.biblimed.com/appli/index.php

36. Sarkar IN, Schenk R, Miller H, Norton CN. LigerCat: using "MeSH clouds" from journal, article, or gene citations to facilitate the identification of relevant biomedical literature. AMIA Annual Symposium Proceedings 2009 American Medical Informatics Association Available from: http://ligercat.ubio.org/
37. Xuan W, Dai M, Buckner J, Mirel B, Song J, Athey B, et al. Cross-domain neurobiology data integration and exploration. BMC genomics, 2010 11(Suppl 3): p S6 PubAnatomy Available from: http://brainarray.mbni.med.umich.edu/Brainarray/prototype/PubAnatomy/
38. Smalheiser NR, Zhou W, Torvik VI. Anne O'Tate: A tool to support user-driven summarization, drill-down and browsing of PubMed search results. Journal of biomedical discovery and collaboration, 2008 3(1): p 2 Anne O'Tate Available from: http://arrowsmith.psych.uic.edu/cgi-bin/arrowsmith_uic/AnneOTate.cgi
39. Doms A, Schroeder M. GoPubMed: exploring PubMed with the gene ontology. Nucleic acids research, 2005 33(suppl 2): p W783-W786 Available from: http://gopubmed.com/web/gopubmed/
40. MEDSUM, developed by Galsworthy, M.J., Hosted by the Institute of Biomedical Informatics (IBMI), Faculty of Medicine, University of Ljubljana, Slovenia. Available from: http://webtools.mf.uni-lj.si/public/medsum.html
41. PubMed PubReMiner, developed by Jan Koster, Academisch Medisch Centrum, Universiteitvan Amsterdam. Available from: http://bioinfo.amc.uva.nl/human-genetics/pubreminer/
42. KNALIJ, developed by Steve Melnikoff at iWakari. Available from: http://knalij.com/
43. Parker D, Chu W, Sabb F, Toga A, Bilder R. Literature Mapping with PubAtlas—extending PubMed with a 'BLASTing interface'. Summit on translational bioinformatics, 2009 2009: p 90 PubAtlas Available from: http://www.pubatlas.org/
44. Plake C, Schiemann T, Pankalla M, Hakenberg J, Leser U. AliBaba: PubMed as a graph. Bioinformatics, 2006 22(19): p 2444-2445 AliBaba Available from: http://alibaba.informatik.hu-berlin.de/
45. Hristovski D, Peterlin B, Mitchell JA, Humphrey SM. Using literature-based discovery to identify disease candidate genes. International journal of medical informatics, 2005 74(2): p 289-298 BITOLA Available from: http://ibmi3.mf.uni-lj.si/bitola/
46. Douglas SM, Montelione GT, Gerstein M. PubNet: a flexible system for visualizing literature derived networks. Genome biology, 2005 6(9): p R80 PubNet Available from: http://pubnet.gersteinlab.org/
47. Perez-Iratxeta C, Bork P, Andrade MA. XplorMed: a tool for exploring MEDLINE abstracts. Trends in biochemical sciences, 2001 26(9): p 573-575 XplorMed Available from: http://xplormed.ogic.ca/
48. Tenner H, Thurmayr GR, Thurmayr R. Data mining with Meva in MEDLINE. Medical Data Analysis 2003, Springer p 39-46 Meva Available from: http://www.med-ai.com/meva/index.html
49. MEDLINE Fact Sheet. Available from: http://www.nlm.nih.gov/pubs/factsheets/medline.html
50. UMLS - Current Semantic Types. Available from: http://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html
51. Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. ACM SIGMOD Record; 1993: ACM; 1993. p. 207-16.
52. Lin W, Alvarez S, Ruiz C. Efficient Adaptive-Support Association Rule Mining for Recommender Systems. Data Mining and Knowledge Discovery. 2002 2002/01/01;**6**(1):83-105.
53. Antonie M-L, Zaiane OR, Coman A. Application of Data Mining Techniques for Medical Image Classification. MDM/KDD. 2001;**2001**:94-101.
54. Srivastava J, Cooley R, Deshpande M, Tan P-N. Web usage mining: discovery and applications of usage patterns from Web data. SIGKDD Explor Newsl. 2000;**1**(2):12-23.
55. Agrawal R, Srikant R. Fast algorithms for mining association rules. Proc 20th Int Conf Very Large Data Bases, VLDB; 1994; 1994. p. 487-99.
56. Geng L, Hamilton HJ. Interestingness measures for data mining: A survey. ACM Computing Surveys (CSUR). 2006;**38**(3):9.
57. Tan P-N, Kumar V, Srivastava J. Selecting the right interestingness measure for association patterns. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining; 2002: ACM; 2002. p. 32-41.
58. Diaconis P, Efron B. Testing for independence in a two-way table: new interpretations of the chi-square statistic. The Annals of Statistics. 1985;**13**(3):845-74.
59. Cao H, Hripcsak G, Markatou M. A statistical methodology for analyzing co-occurrence data from a large sample. Journal of biomedical informatics. 2007;**40**(3):343-52.
60. Wright A, Chen ES, Maloney FL. An automated technique for identifying associations between medications, laboratory results and problems. Journal of biomedical informatics. 2010;**43**(6):891-901.
61. Team RC. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2012: ISBN 3-900051-07-0; 2012.
62. Rakotomalala R. TANAGRA: a free software for research and academic purposes. Proceedings of EGC; 2005; 2005. p. 697-702.
63. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. ACM SIGKDD explorations newsletter. 2009;**11**(1):10-8.
64. Zupan B, Demsar J. Open-source tools for data mining. Clinics in laboratory medicine. 2008;**28**(1):37-54.
65. Zhao Y. R and Data Mining: Examples and Case Studies: Access Online via Elsevier; 2012.
66. CRAN. Available from: http://cran.us.r-project.org/
67. Rakotomalala R. TANAGRA: une plate-forme d'expérimentation pour la fouille de données. Revue MODULAD. 2005;**32**:70-85.
68. Witten IH, Frank E, Trigg LE, Hall MA, Holmes G, Cunningham SJ. Weka: Practical machine learning tools and techniques with Java implementations. 1999.
69. Hahsler M, Grün B, Hornik K, Buchta C. Introduction to arules–A computational environment for mining association rules and frequent item sets. The Comprehensive R Archive Network. 2009.
70. Hahsler M, Chelluboina S. Visualizing Association Rules: Introduction to the R-extension Package arulesViz. R project module. 2011.
71. The R Project for Statistical Computing. Available from: http://www.r-project.org/
72. Dahl DB, Crawford S. Rinruby: accessing the r interpreter from pure ruby. J Stat Softw. 2008;**29**(4):1-18.
73. Entrez Programming Utilities Help [Internet]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK25501/
74. Semantic Groups. Available from: http://semanticnetwork.nlm.nih.gov/SemGroups/
75. Gephi. Available from: https://gephi.org/
76. Cytoscape. Available from: http://www.cytoscape.org/
77. Huang M, Névéol A, Lu Z. Recommending MeSH terms for annotating biomedical articles. Journal of the American Medical Informatics Association. 2011 May 25, 2011.
78. Funk ME, Reid CA. Indexing consistency in MEDLINE. Bulletin of the Medical Library Association. 1983;**71**(2):176.
79. Gay CW, Kayaalp M, Aronson AR. Semi-automatic indexing of full text biomedical articles. AMIA Annual Symposium Proceedings; 2005: American Medical Informatics Association; 2005. p. 271.

Patient-Centered Case Management System (P-CMS)

Keith Butler, PhD[1], Andrew Berry[1], Amy Walker, PhD, MSN[1], Nikki Pete[1], Yi-Chen Sung[1], Craig Harrington[2], Jodie Haselkorn, MD[3], Paul Nichol, MD[4], Mark Oberle, MD[1], Lucas McCarthy, MD[3], & Mark Haselkorn, PhD[1]

1. University of Washington, Seattle, WA; 2. University of Texas, Houston TX; 3. VHA MS Center of Excellence, West, Seattle, WA; 4. Office of Informatics and Analytics, VHA, Washington, DC

ABSTRACT: P-CMS addresses an important gap in HIT for case management of multiple sclerosis (MS) with a design based on the workflow of MS care, the information clinicians actually use, and innovation to analyze the tacit, cognitive work of case management for effective HIT support. MS is one of many chronic diseases where clinicians spend excessive time on clerical tasks for the information they need to manage complex cases and coordinate treatment plans. The tacit nature of this care activity, however, made it difficult to understand how HIT should be applied. This demonstration will illustrate how an analysis of the conceptual work products[1] of case-management provided requirements and design concepts for an effective solution. P-CMS works with key use-cases of case management to form a measurably better workflow. The web-based prototype was implemented for usability testing, simulation of impact on workflow, and technical feasibility evaluation. The results indicate P-CMS will improve awareness of patients' treatment plans, increase timeliness of order completions, while eliminating many clerical tasks for time-savings of about 1-hour per shift. We will review the cognitive strategy of its design, progress towards an alpha test version, and generality of the design for other chronic diseases.

INTRODUCTION: Seventy-five percent of U.S. health care dollars are spent on chronic illness and the cost is growing with population aging. MS is an example of chronic illness where case managers play an important role coordinating and monitoring complex treatment plans that unfold over months for outpatients. The number and duration of orders, combined with the complexity of conditions, require a wide variety of information to manage cases and coordinate care. Using information from a wide variety of resources can impose burdensome clerical tasks to access it, integrate it manually, then keep the ad-hoc collection current, on top of doing patient care. Without an information system that is well-designed specifically for case management these overhead tasks can be daunting, disrupt timely coordination, and risk needless errors. The specific purpose of P-CMS is to provide a single tool that integrates needed information in a highly usable, effective user interface to improve awareness of the progress of patients' orders and conditions, while eliminating the wasted overhead of clerical information tasks.

P-CMS: Organizes needed information into layers based on an effective cognitive strategy[1] of management by exception against time-based milestones for each type of order. The home page displays a list of all patients with color coding for quick visibility of any patient whose progress has fallen behind one or more milestones. Any patient's order summary can be opened with one click, and any patient can be retrieved by name. Selectable column headers allow users to re-sort the list of patients with one click by:

- Any patients with new or changed orders, displayed by oldest to newest
- Any patients with order milestones that are late, displayed from most behind to least
- Patient's next appointment date, displayed by nearest to furthest
- Patient's last appointment date, displayed by most recent to oldest

All the page views provide color coding to identify patients with exceptions to expected progress.

STATUS: A functional prototype was implemented in HTML with anonymized patient data in a SQL database for usability testing and technical feasibility. Usability results with seven experienced nurses indicate it is highly learnable and easy to use for all use-cases. The Northwest VA regional data warehouse patient-level database was successfully linked to the P-CMS database for feasibility analysis of daily surveillance of clinical orders, lab results, radiology results, consult and appointment tracking.

---

[1] Kieras, D. & Butler, K.A. (2014) Task Analysis and the Design of Functionality. In: A. Tucker, et al (ed.s) *Computing Handbook, Third Edition, Volume 2*. Chapman and Hall/CRC Press.

ACKNOWLEDGMENTS: This project was supported by grant number R01HS021233 from the Agency for Healthcare Research and Quality. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Agency for Healthcare Research and Quality. We gratefully acknowledge participation by the VHA MS COE, West, and VHA Office of Informatics and Analytics.

Pajekto3DStereo: Enabling Generation and Interaction with 3D Stereo Networks

Bryant Dang BS, Suresh K. Bhavnani PhD
Institute for Translational Sciences, University of Texas Medical Branch, Galveston, TX

Abstract

Although 2D networks have been useful in revealing complex associations in a wide range of biomedical applications, they are often too dense to comprehend. In such situations, a 3D layout of the same network provides an extra degree of freedom in the z-plane, often resulting in a more accurate representation of the associations in the data. However, 3D layouts displayed in 2D occlude nodes that are further away in the z-plane, and therefore require the network to be rotated in order to view the nodes and comprehend their relationships in 3D space. Unfortunately, this rotation leads to disorientation and therefore requires a stereo version of the 3D layout to enable rapid comprehension. Because tools to develop such 3D stereo layouts are either proprietary or expensive, we developed *Pajekto3DStereo*, a simple interface in R that enables researchers to convert a 3D network layout into a format that can be used by a freely available stereo-visualizing tool called VMD. *Pajekto3DStereo* has enabled the discovery of a complex pattern related to two intersecting biological pathways that was missed using a 2D network layout.

Introduction

Numerous projects have demonstrated the utility of 2D network layouts to help researchers comprehend complex multivariate relationships[1]. However, network layouts are often too dense, resulting in the infamous "hairball" network with apparently no discernable pattern. In prior research[2,3] we have argued that because 3D network layouts provide an extra degree of freedom on the z-plane, force-directed algorithms[1] such as *Fruchterman Reingold* (FR) generate network layouts that more accurately reflect complex associations in the data. However, 3D network layouts displayed in 2D lead to node occlusion or disorientation when the model is rotated to view the occluded nodes. Such networks need to be displayed in 3D stereo which enables comprehension of the 3D layout without having to rotate the model. Here we describe a tool that enables the rapid generation of 3D stereo network models.

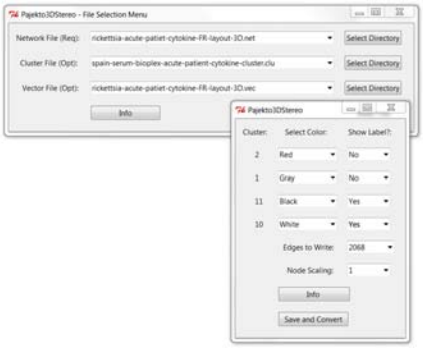


Figure 1. Dialog boxes in *Pajekto3DStereo*

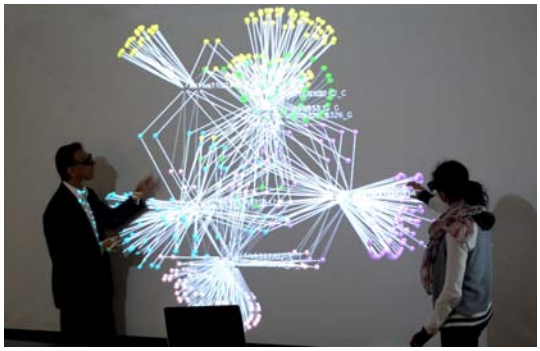


Figure 2. Output of *Pajekto3DStereo* being viewed in stereo through the use of VMD and stereo glasses

Method and Results

As shown in Figure 1, we have developed a tool in R called *Pajekto3DStereo* which takes as input a network layout file containing the X,Y,Z coordinates of nodes laid out by a 3D network layout algorithm (e.g., FR available in Pajek[1]). Through this interface, the user can also specify (1) color and size of the nodes, (2) number of edges to be displayed, and (3) display parameters that are critical for a high fidelity layout. *Pajekto3DStereo* then converts the network layout data into a format that can be read by VMD (a well-known freeware used to display molecules, nodes, and edges in stereo). As shown in Figure 2, the converted file can then be displayed and interacted with in 3D stereo. Such visualizations generated with *Pajekto3DStereo* have enabled the discovery of a sub-topology that was missed when using 2D network layouts[2]. In future research we intend to expand the features provided by *Pajekto3DStereo*, and user test the interface with the goal of enabling a wide range of researchers to freely and rapidly convert 2D network data into 3D stereo views that enable discoveries in complex multivariate data.

References

1. Nooy W, Mrvar A, Batagelj V. Exploratory Social Network Analysis with Pajek. New York, NY: Cambridge University Press, 2011.
2. Bhavnani SK, et al. Discovering Hidden Relationships between Renal Diseases and Regulated Genes through 3D Network Visualizations. BMC Research Notes, 2010;3:296.
3. Bhavnani SK, Drake J, Dang B, Visweswaran S, Olano JP. Comprehension of Multiple Molecular Pathways using 3D Networks. Proceedings of AMIA Summit on Translational Bioinformatics, 2013.

Big Data for Critical Care with Cloud-based In-Memory Database

Mengling Feng, PhD[1], Mohammad Ghassemi, Msc[1], Thomas Brennan[1],PhD, John Ellenberger, Msc[2], Ishrar Hussain, PhD[2], Roger Mark, M.D. PhD[1]
[1]Massachusetts Institute of Technology, Cambridge, MA; [2]SAP Research, Burlington, MA

Abstract

MIMIC is an open-access database, which holds clinical data from over 60,000 Intensive Care Unit (ICU) stays, with over 7,000 records matched with time-series physiological data. MIMIC is valuable Big Data research resource, which is over 4TB in size and currently has over 1,000 users around the globe. Analyzing large database like MIMIC is, however, computationally expensive due to its high dimensionality and vast volume. We have been exploring the in-memory database as a solution. In collaboration with SAP, our industrial partner, we will showcase how Big Data analytics can be effectively conducted with HANA, a Cloud based in-memory database system from SAP. Different from traditional databases, HANA is capable of storing the whole database in the main memory for fast and dynamic querying. We will have a live demonstration to showcase the advantage of HANA over traditional databases in terms of query speeds. We will also showcase how data extraction and statistical analysis of Big Data like MIMIC can be conveniently accomplished as one integrated process with HANA, which fully integrates SQL with R, the statistical analytic language.

MIMIC: An Open Big Data for Critical Care:

Our lab and partners have built and been maintaining the Multi-parameter Intelligent Monitoring in Intensive Care (MIMIC) database [1] since 2003. MIMIC is the largest open Big Data resource for critical care, and it currently has over 1,000 users around the globe. This open-access database, which now holds clinical data from over 60,000 stays in the BIDMC ICUs, has been meticulously de-identified and is shared online with the research community via PhysioNet (www.physionet.org). In this demonstration, we will guide the audiences to navigate through both the clinical component (400 GB) and the time-series (4 TB) components of MIMIC. The clinical component includes structured data, such as demographic data, lab & microbiology test results, medication & fluid records, diagnoses, hourly nurse-verified vital sign readings and patient mortality data. MIMIC also contains rich unstructured text data, which includes notes, reports and summaries. The time-series component of MIMIC consists of continuous high resolution (8-bit,125Hz) physiological waveforms data and trend data that were sampled at 1 data point per minute.

Big Data Analytics Empowered by In-Memory Database

Analyzing large database like MIMIC is computationally expensive due to the high dimensionality and large volume of the data. Performance and scalability issues of traditional databases often limit the usage of more sophisticated and complex data analytic models. In addition, a complete data analytic task requires multiple tools across various software platforms. Exporting and importing large amount of data across platforms requires a tremendous amount of computation and I/O resources. We have been exploring the in-memory database as a solution. SAP has recently developed an cloud-based in-memory database, named HANA, which is capable of storing the large volumes of data in the main memory, instead of the disk drives. HANA can dramatically reduce the time needed for data queries by eliminating the disk seeking, reading and writing. We will conduct a live demonstration to showcase the computational advantage of HANA over traditional databases by comparing their query speeds over the MIMIC data. Another great advantage of HANA is that it is fully integrated with R, the open source statistical analytic language, which has rich data mining and machine learning libraries. We will also demonstrate how researchers can conveniently conduct both data extraction and statistical data analysis under one single HANA platform.

Current System Development Status

HANA has already been deployed on an AWS cloud instance and can be easily launched by users. In the cloud instance, the in-memory database together with R has been implemented and populated with the MIMC data.

References

1. M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, T. Heldt, T. H. Kyaw, B. Moody, and R. G. Mark, "Multiparameter intelligent monitoring in intensive care ii (mimic-ii): A public access intensive care unit database," Critical Care Medicine, vol. 39, pp. 952–960, May 2011.
2. L. A. Celi, R. G. Mark, D. J. Stone, and R. A. Montgomery, "Big data in the intensive care unit, closing the data loop," American Journal of Respiratory and Critical Care Medicine, vol. 187, no. 11, pp. 1157– 1160, 2013.

Narrative Event and Temporal Relation Visualization Tool

Sean P. Finan[1], Piet C. De Groen[2], Guergana K. Savova[1]
[1]Children's Hospital Boston Informatics Program, Harvard Medical School, Boston, MA;
[2]Mayo Clinic College of Medicine, Mayo Clinic, Rochester, MN

Abstract

Extensive critical patient information typically is in an unstructured, free text format – the clinical narrative – that can only be accessed by reading the full text. However, the amount of information within the Electronic Medical Record (EMR) of a single patient is expanding beyond the ability of someone to read within a typical appointment slot. New Natural Language Processing (NLP) methods allow automated extraction of medical events and temporal relations among those events from clinical narratives. In order to display the clinically relevant events from a complete life span of a patient, we created a novel visualization tool that allows scrolling and zooming in time while maintaining an overview of the entire timeline within a single frame. We selected four key features of a typical clinical encounter as the main content of a medical timeline: (1) signs and symptoms; (2) tests and procedures; (3) diseases and disorders; and (4) medications. Within these four features, more detailed subset timelines are allowed. We will demonstrate our prototype graphical user interface and discuss some of the challenges unique to the visualization of unstructured clinical narratives as well as our solutions.

Description of the System

NLP systems can pull a substantial amount of temporal information from text, including events from patient - physician discussions and activities of daily living, along with event modifiers such as negation and uncertainty. Although several excellent timeline visualization tools exist, they are geared towards rendering data from the structured part of the EMR and could not easily be extended to display information from the clinical narrative[1,2,3]. In addition to the new approaches listed in the abstract, our tool uses a unique combination of symbols, coloring and shading to expressively display nuanced event-time relations, event properties, and time span attributes. A marked-up representation of the source note allows convenient comparison between the graphical representation and text.

Statement of Use

A prototype is being tested as part of the Temporal Histories of Your Medical Event (THYME) project[4]. It will soon be integrated into Apache clinical Text Analysis and Knowledge Extraction System (cTAKES) and as such will be available through the Informatics for Integrating Biology & the Bedside (I2B2) platform[5,6]. Use of the tool will also play a part in an upcoming project in the oncology domain funded by the National Cancer Institute (NCI).

Acknowledgments

This work is partially supported by THYME (R01LM010090) and I2B2 (U54LM008748) from the National Library of Medicine. Parts of our schema stem from discussion held during our presentation at the 30th Annual Human-Computer Interaction Lab (HCIL) Symposium[7].

References

1. Human-Computer Interaction Lab (HCIL). LifeLines for visualizing patient records. Available: http://www.cs.umd.edu/hcil/lifelines [2014, March 13].
2. Human-Computer Interaction Lab (HCIL). Lifelines2: Discovering temporal categorical patterns across multiple records. Available: http://www.cs.umd.edu/hcil/lifelines2 [2014, March 13].
3. Human-Computer Interaction Lab (HCIL). EventFlow: Exploring point and interval event temporal patterns. Available: http://www.cs.umd.edu/hcil/eventflow [2014, March 13].
4. Temporal Histories of Your Medical Event (THYME). Available: http://thyme.healthnlp.org [2014, March 13].
5. Apache clinical Text Analysis and Knowledge Extraction System (cTAKES). Available: http://ctakes.apache.org [2014, March 13].
6. Informatics for Integrating Biology & the Bedside (I2B2). Available: http://www.i2b2.org [2014, March 13].
7. Finan SP. Challenges of visually representing rich temporal information of the clinical narrative. Workshop: Exploring temporal patterns in electronic health record data. Human-Computer Interaction Lab Symposium. May 23 2013. University of Maryland. Available: http://www.cs.umd.edu/hcil/eventflow/workshop2013 [2014, March 13].

Using Electronic Medical Record Time and Quality Metrics to Identify Provider-Specific Training Opportunities

Anupam Goel, MD, MBA, Advocate Health Care, Oak Brook, IL; Bonnie Dean, Cerner, Kansas City, MO; Harlen Hays, Cerner, Kansas City, MO

Abstract

Time in the electronic medical record (EMR) provides a single dimension into inpatient provider behavior with the EMR. Coupling EMR time metrics with other metrics (*e.g.*, number of patients seen, percent of orders canceled, document quality) could help a clinical informatics department better identify opportunities to either adjust training opportunities or prioritize enhancements to help improve end-user EMR efficiency for large groups of users.

Introduction

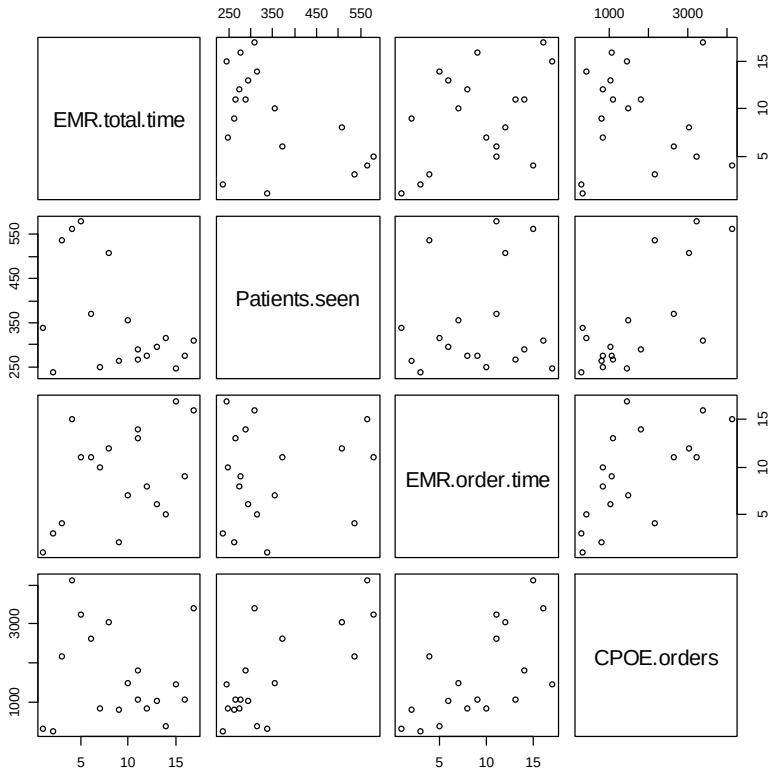
Some EMR vendors can provide time metrics to help clients identify users who may spend too little or too much time in the electronic medical record. A more appropriate measure of inpatient provider EMR behavior would balance time in EMR against other metrics including patient volume, EMR adoption and appropriate EMR use.

Methods

Advocate and Cerner developed a user dashboard including three different domains of time spent in the EMR (chart review, ordering and documentation) along with measures of EMR adoption (orders directly entered into the EMR [CPOE], clinical documents directly entered into the EMR) as well as measures of appropriate EMR use (percent of orders canceled, documentation quality).

Preliminary Results

The figure on the right describes the relationships between average daily time in a patient's chart, the number of patients seen during the month, the average time spent entering orders per patient per day and the number of CPOE orders among providers who saw at least 200 patients in two consecutive months at one hospital. Without additional information beyond time spent in the EMR, it would be difficult to identify opportunities for improvement.



Conclusions

A combination of EMR-generated provider metrics, EMR adoption and appropriate EMR use can assist a medical center identify specific opportunities to improve the EMR experience beyond time-based measures.

Demonstrating A Public Health Terrain Data Visualization System

Jeremy Keiper[1], Shiaofen Fang, PhD[1], Yuni Xia, PhD[1], Mathew Palakal, PhD[2], Shaun Grannis, MD[3], Thanh Minh Nguyen[1], Sam Bloomquist[1], Anand Krishnan[2], Weizhi Li[1]
[1] Department of Computer Science, Indiana University – Purdue University, Indianapolis;
[2] School of Informatics, Indiana University – Purdue University, Indianapolis;
[3] Regenstrief Institute, Indianapolis, IN

Abstract

Use cases for public health data visualization systems indicate a need for an interactive system, a wide variety of filters for narrowing scope, and visualizations in both geospatial and logical domains. A public health study not only involves relationships between individuals within a given community, but also considers correlations among attributes of diseases. These correlations are not inherently obvious unless one has prior domain expertise, and exploration of these relationships can be tedious even with complex statistical analysis tools. Ultimately, public health officials seek simplified datasets providing the minimum factors necessary to illustrate a problematic scenario. Through a combination of specific health data sources and innovative text and data analyses, we created a visualization engine allowing public health officials to quickly define, assess, and address potential epidemics. Our innovations include unique text and data mining techniques for discovering relationships spanning multiple domains, and innovative interactive visualizations providing comprehensive knowledge transfer to the user.

Introduction and Background

Public health researchers need access to relationships within communities to understand the details of an epidemic. Individuals can be connected to others by geographic proximity, shared community spaces, common treatments, or other shared personal traits. Most current work in health visualization focuses on analyzing individual patients. Our goal was to provide access to large, cohort-based indicators along multiple dimensions, allowing an epidemiologist to find outbreaks and disease patterns without requiring prior knowledge of dataset characteristics.

Our visualization interface provides multiple intelligent visualizations of regionalized health data, allowing users to see relationships at varying levels of granularity, as shown below: a navigable weighted graph demonstrating disease relationships; a geospatial heat map showing population density of selected diseases; a ring graph depicting patient-disease relationships and demographic grouping over time; a 3D geographic visualization for interactive exploration of cumulative disease occurrences; and a theme river used to show disease occurrence comparisons over time. We tailored the user interface for quick access to data filtering controls and snapshots of multiple alternative visualizations, surrounding a main interactive window that can assume the entire window on demand.

Unique data sources and analysis techniques coupled with interactive visualizations comprise the core of our innovations. The Notifiable Condition Detector (NCD) system at Regenstrief Institute informs public health officials by identifying lab tests indicating any reportable conditions, as defined by the state of Indiana. We use outcomes from the NCD to preselect patients of interest in our dataset. We then align these patients with discharge notes from hospitals in the Indiana Network for Patient Care, and process all of the free-form data through text mining analysis to identify additional diseases and other relevant attributes. Our system detects correlations across domains, including patients, diseases, and related attributes, and prepares this metadata for interactive visualization.

To investigate an outbreak, an epidemiologist could: search for the disease (filter sidebar); visualize a network of highly related diseases (association graph); select other closely related diseases to include in the visualizations (multiple interfaces); find the geographic distribution for multiple diseases, or a single disease over time (chloropleth, 2D or 3D); and look for patterns over time for patients contracting the disease multiple times or in sequence with its comorbidities (ring graph), exploring each visualization interactively to narrow focus.

As of this writing, our system is still under development and will not be deployed publicly unless so determined by its funder. This project is supported by the Department of the Army, award number W81CWH-13-1-0020.

Implementation of a Population Health Record in Montreal, Canada

Maxime Lavigne BEng, Lam Dang-Duy BEng, Eugene Fiziev, Catherine Ghassemian BEng, Alexis Hamel, Anya Okhmatovskaia PhD, Mojtaba Peyvandy, Arash Shaban-Nejad PhD, David L Buckeridge MD PhD

McGill Clinical and Health Informatics, McGill University, Montreal, QC, Canada

Abstract

In response to a critical lack of population health information, the Institute of Medicine has called for systems that use a determinants-of-health approach to integrate health and non-health data and provide easy access to health indicators. We have developed and implemented in the Montreal region a Population Health Record (PopHR), which computes indicators relevant to diabetes automatically from health and non-health data, allows access to indicators through free-text queries, and contextualizes indicator values using existing epidemiological knowledge.

Background

Local, state, and national decision-makers lack sufficient information to make important choices about the health of their communities [1]. This lack of information is particularly troubling in the context of the massive and growing burden of non-communicable diseases, which in 2010 were responsible for 83% of Disability Adjusted Life-Years in Developed countries [2]. To address this critical lack of population health information, the Institute of Medicine (IOM) recommends development of public health systems that link health and non-health data using a determinants of health approach to provide transparent and easily understood information in a timely manner for all levels of geography [1]. Current web-based information portals do not meet these requirements. Building on concepts advanced by the AMIA Board of Directors [3], we have developed a Population Health Record (PopHR) [4], which meets the requirements proposed by the IOM to address the pressing need for population health information [1].

System Description

The PopHR is a web-based, ontology-driven software platform that automates the retrieval and integration of heterogeneous data from health and non-health sources, and supports intelligent analysis and visualization of these data using knowledge of the determinants-of-health. Focused on improving decision-making related to the planning and evaluation of population health interventions, the PopHR addresses common limitations of existing web portals for population health information [1] through the following features:

- *Automated computation of health indicators from regularly updated sources of health and non-health data.* The PopHR integrates at a high spatiotemporal resolution individual-level data from administrative and clinical sources with small-area data on upstream health determinants from government and commercial sources.
- *Natural language interface enabling free-text queries.* Domain knowledge, usage data, and default values are used to provide query suggestions and auto-completion options. The limited use of filtering and stratification controls allows for iterative query refinement without cluttering the interface.
- *Contextualization of health indicators using existing knowledge about the determinants of health*. We have developed a suite of OWL ontologies to encode epidemiological knowledge about causal pathways for chronic diseases. The PopHR semantic framework uses this knowledge to reveal meaningful links between multiple health indicators, customize visualization, and support the development and evaluation of population health interventions.

The PopHR is currently implemented in the greater Montreal area (population 3.8 million) with an initial focus on diabetes [5]. Together with collaborators at the Montreal Health Department, the Quebec Public Health Institute, and the Quebec Ministry of Health, we are refining the PopHR in preparation for evaluation through controlled trials in public health and health system administration settings in Quebec.

References

1. IOM (Institute of Medicine). For the Public's Health: The Role of Measurement in Action and Accountability. Washington, DC: The National Academies Press; 2011.
2. Murray CJL, Vos T, Lozano R, et al. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet 2012;380(9859):2197–223.
3. Board of Directors of the American Medical Informatics Association. A proposal to improve quality, increase efficiency, and expand access in the U.S. health care system. JAMIA 1997;4(5):340–341.
4. Friedman DJ, Parrish RG. The population health record: Concepts, definition, design, and implementation. JAMIA 2010;17(4):359–366.
5. Buckeridge DL, Izadi MT, Shaban-Nejad A, Mondor L, Jauvin C, Dube L, Jang Y, and Tamblyn R. An infrastructure for real-time population health assessment and monitoring. IBM Journal of Research and Development, 2012, 56(5):22.

SMART on FHIR®

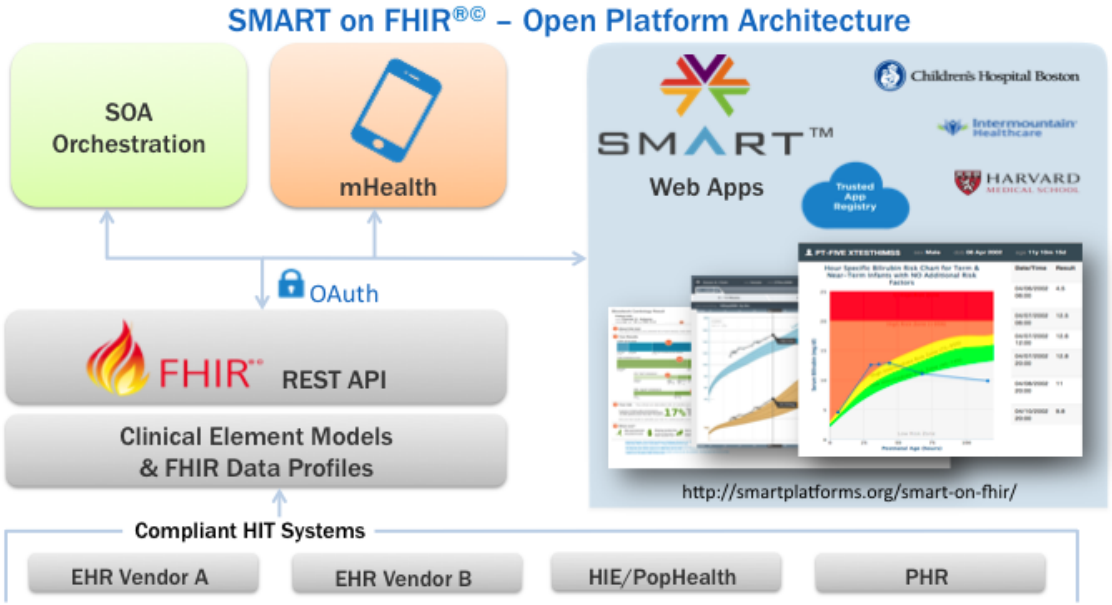
David P. McCallie Jr, MD[1], Joshua Mandel, MD[2], Stanley M. Huff, MD[3], Kenneth Mandl, MD, MPH[2], Isaac Kohane, MD, PhD[4]

[1]Cerner Corp., Kansas City, MO; [2]Boston Children's Hospital, Boston, MA; [3]Intermountain Healthcare, Salt Lake City, UT; [4]Harvard Medical School, Boston, MA

Abstract: *Kohane and Mandl first proposed the development of "substitutable apps" for healthcare systems in a widely discussed 2009 New England Journal of Medicine article. SMART on FHIR represents a vendor-neutral, standards-based, real-world implementation of that vision in support of an "app store" for healthcare. SMART on FHIR implements an open architecture to support interchangeable Web applications that can be "plugged in" to any compliant EHR or health data container. Clinical data are exposed using an Application Programming Interface (API) defined by Fast Healthcare Interoperability Resources (FHIR), an emerging HL7 standard. Semantic data interoperability is achieved using a set of FHIR Profiles that constrain a set of core Resource definitions using a comprehensive open-source library of Clinical Element Models, developed at Intermountain Healthcare. We will demonstrate multiple independently-authored applications running against commercial EHRs as well as other health information technology systems. We will show applications that visualize clinical data, provide decision support, and integrate clinical data with external sources including population health and HIE data.*

Purposes of the System: Modern EHRs are complex systems that can be thought of as "platforms" that provide support for core services such as documentation, order-entry, workflow, decision support, and data persistence. However, it is unlikely that any single vendor can provide support for all of the interactions and services desired by providers. SMART on FHIR defines a standards-based, vendor-neutral way to allow application developers to extend any compliant EHR platform. These extensions can address enhanced decision support, sophisticated data visualizations, and data integration from external sources. SMART on FHIR also offers a clear path for academic informaticists to develop, test, and deploy innovations, without the need to control the underlying EHR software.

Deployment Status: Pilot demonstrations have shown two commercial apps and four open-source apps running against four FHIR service hosts, including a commercial EHR and an open-source VistA implementation.



The HealthITxChange: A Community Infrastructure for Clinicians, Educators, Researchers, and Health IT Professionals focused on Ambulatory EHR Implementation and Use.

Helga E. Rippen, MD, PhD, MPH, FACPM [1], Christoph Ulrich Lehmann, MD, FAAP, FACMI[2], Ejim Mark, MD, MPH, MBA [1]

[1]Rockville Institute, Rockville, MD and Westat, Rockville, MD; [2]Vanderbilt University, Nashville, TN

Abstract

The HealthITxChange (www.HealthITxChange.org) is a collaborative effort between Rockville Institute, AMIA, ANA, AHIMA Foundation, and HL7 to provide a free, online community for educators, researchers, clinicians, and other health IT professionals to share their knowledge, lessons learned, and resources with a focus on ambulatory EHR implementation and use. Individuals responsible for implementing EHR systems often reach out to listservs, follow blogs, search the web, and go to conference sessions promising actionable information and lessons learned in order to find solutions to barriers they are facing or to increase their likelihood of success. Often, information in these sources is limited in scope and/or perspective, difficult to access and potentially biased. In an attempt to address these challenges an infrastructure to support an online community sharing timely, "unbiased" lessons learned on EHR implementation was created. The HealthITxChange also leverages "web harvesting" techniques to provide links to other websites, e.g., PubMed, CMS)] and discussion threads. Finally, to reduce bias and to set clear community expectations around behavior, members agree to abide by a Code of Conduct and the peer review process. This Collaboration is governed by the Executive Committee made up by representatives from Rockville Institute, AMIA, ANA, AHIMA Foundation, and HL7.

Purpose of the System

The purpose of the HealthITxChange is to provide an online community for clinicians, educators, researchers, health IT professionals, to share and access lessons learned, external links, comments, and resources. This online community will provide an environment to share knowledge and to accelerate learning around ambulatory EHR implementation and use.

Innovations

The system is innovative in that it

- provides a national and international learning community around ambulatory EHR implementation and use;
- minimizes the overwhelming nature of complex information by organizing it around the implementation process and tailoring it to the user's characteristics;
- provides a peer review process to mitigate bias and enhance quality;
- allows for community assessment of the usefulness of the information;
- provides links to relevant external links through a uniquely applied deep web harvesting technique tagged to topics;
- enables discussion tagged to a concept and social medial enabled sharing;
- provides a resource for training; and
- enables the community to identify gaps in areas such as policy, marketplace, and knowledge.

Degree Deployed

The site was made available online – www.HealthITxChange.org - on February 20, 2012. The site is focusing its efforts on developing a strong community and providing an infrastructure for researchers. Currently there are 360 individual members and 10 peer-reviewers with over 720 lessons learned, 100 resources, and 40,000 links. One academic center has used it as a core infrastructure for a grant submission and at least two health informatics courses include it as a resource for students in the course syllabus. There are over 17,511 unique visitors over the last year with 375 registered members representing over 40 countries.

An Electronic Health Record for Google Glass: A System Demonstration

Presenter and Developer: Karandeep Singh, MD[1,2]
[1]Brigham and Women's Hospital, Boston, MA; [2]Harvard Medical School, Boston, MA

Abstract

Though Google Glass (also known as "Glass") has been used to record and live stream surgical procedures, its use as an EHR platform has been largely overlooked. This system demonstration will review the basic principles of Google Glass development, demonstrate a novel functioning electronic health record (EHR) for Glass, and discuss potential use cases. The first part of the presentation will cover the strengths and limitations of Glass as a platform for development and clinical use. The second part will walk through the process of developing a simple Google Glass application using web technologies such as HTML5, jQuery, and the Adobe Phonegap framework. The third part will focus on demonstrating a functioning EHR application that interfaces with the Longitudinal Medical Record EHR at Brigham and Women's Hospital. Several input methods including voice commands, accelerometer-based scrolling, QR code recognition, and touch-based gestures will be used to access the EHR through Glass.

Description

The mobile use of electronic health records (EHRs) has been primarily confined to smartphones and tablets. Though EHRs aimed at mobile devices offer a convenient way for healthcare practitioners to interact with the medical record, their use is limited by the fact that they are not hands-free devices. The Google Glass provides a unique platform by which a user can access data hands-free in real-time. This system demonstration will describe the process of developing a Google Glass-based application and demonstrate a functioning Glass EHR at Brigham and Women's Hospital.

Developing for Google Glass

Google Glass is a mobile device worn on in place of (or as part of) eyeglasses that contains a transparent display visible to the wearer's right eye. The device runs on the Android operating system and can directly access wireless internet through a built-in wireless chip and can connect to a user's smartphone using Bluetooth. Using web technologies in conjunction with Adobe's Phonegap framework, building an application for Google Glass is fast and flexible. For the first part of the demonstration, I will cover the strengths and limitations of Glass as a platform for development and clinical use. Next, I will walk the audience through the process of starting a Phonegap project and building an HTML5-based Google Glass application that displays "Hello, World." I will review relevant plugins that facilitate data entry on Glass, including how to capture touchpad and accelerometer input, read QR codes, and implement speech recognition for capturing voice commands.

Demonstrating An Electronic Health Record on Glass

The Google Glass-based EHR demonstration will show a novel and functional EHR application that uses each of the different input methods described earlier to simplify the process of retrieving electronic patient data from the Brigham and Women's Hospital Longitudinal Medical Record EHR. I will demonstrate the following features:

- Selecting a patient using the swipe and tap gestures on the touchpad
- Scanning a QR code to securely input a patient medical record number
- Using voice commands for queries (e.g. "show me the meds", "what is the trend for PTH?")
- Using Glass in conjunction with a smartphone (e.g. "text me the number for the primary care doctor")
- Scrolling through large lists using accelerometer input to guide the scrolling speed

Deployment

A functional Glass EHR application has been built and is in pre-clinical use at Brigham and Women's Hospital. We are in the process of planning a trial of the application on the inpatient dialysis unit.

An EMR Designed for Teaching and Educational Research Based on Regenstrief Institute's Gopher System

Blaine Y. Takesue, MD[1,2], John T Finnell, MD[1,2], Jon D. Duke, MD, MS[1,2]

[1]Regenstrief Institute, Indianapolis, IN

[2]Indiana University School of Medicine, Indianapolis, IN

Abstract

Recognizing a significant gap between what students are taught in medical school and what physicians need to know in the modern practice of medicine, many experts have called for medical schools to improve student skills in areas including multidisciplinary teamwork, data analytics and electronic medical records. To help align medical education with a skillset future physicians will require, the Regenstrief Institute recently developed a standalone version of its Gopher computerized physician order entry system designed expressly for use in teaching environments. The system, known as tEMR, combines a rich patient database with the Gopher's clinical decision support architecture to deliver lessons and assess learners' interaction with the system.

Background

While US medical schools matriculate students with a depth of scientific knowledge, it is recognized that these same students graduate with inadequate training in subjects including health economics, health policy, and medical informatics. Adding to this gap is a trend within medical school hospitals that limits medical student access to electronic medical records at a time when electronic health records are becoming more central to patient care.[1] The Regenstrief Institute, with the support of the AMA's "Accelerating Change in Medical Education" (ACE) initiative, created a standalone version of its homegrown EMR (known as Gopher[2]), called the teaching EMR (tEMR) which will help close the gap. The tEMR was designed to support the revised curriculum at the Indiana University School of Medicine. Within the tEMR, students are not involved in patient care so the privacy and security issues which limit student access in production EMRs are not a concern. However, the tEMR does deliver a real world EMR experience; 1) as a branch of an in-use production EMR and 2) through the use of a synchronized, de-identified database of real patient data. The tEMR leverages functionality created for the production EMR to deliver customized education content and to assess students' critical thinking processes.

System demonstration

After discussing the tEMR's infrastructure and patient database, we propose to demonstrate several unique features of the Regenstrief tEMR. First, we will show the novel application of Gopher's clinical decision support architecture to deliver 'educational alerts' to students in a customized, contextual manner. We will demonstrate, using Regenstrief's RAVE rule-authoring platform, how educators can create this custom content. We will demonstrate a novel implementation of OpenInfoButton[3] also designed to deliver customized content for students. Finally, we will demonstrate how detailed logs of student interaction with the system, including data review and clinical orders, are used to assess critical thinking skills.

Relevance

The software developed for this project will be released open-source. One of our principle design objectives is to develop software that can be used at medical schools across the country as well as for the education of non-physician learners. We believe attendees will find a variety of use cases—both didactic and research—for which the tEMR will be of benefit

References

1. Pageler, NM, Friedman, CP, Longhurst, CA. Refocusing Medical Education in the EMR Era. JAMA. 2013 Jan:310(21), 2249-50.
2. McDonald CJ, Tierney WM. The Medical Gopher--a microcomputer system to help find, organize and decide about patient data. West. J. Med. 1986 Dec;145(6):823-829.
3. Del Fiol G, Haug PJ, Cimino JJ, Narus SP, Norlin C, Mitchell JA. Effectiveness of topic- specific infobuttons: a randomized controlled trial. J Am Med Inform Assoc. 2008 Nov-Dec;15(6):752-759 (OpenInfobutton: http://www.openinfobutton.org)

Doxy.me- A simple, free, and secure telemedicine solution for health care and research participation

Brandon M Welch, M.S., Ph.D.
Biomedical Informatics Center, Medical University of South Carolina, Charleston, South Carolina

Abstract

To address the need for a free and easy-to-use telemedicine solution needed for a clinical research trial, informaticists at the University of Utah developed Doxy.me, a simple and free Web-based video communication solution. Doxy.me is simple; downloads or plugins are not required and patients do not register or login. Several useful features are available to support clinical workflows such as a virtual waiting room, patient check-in and queue, and meeting controls. Furthermore, interest among clinical researchers at the Medical University of South Carolina has also led to the development of virtual eConsent capabilities to support research study participation using the Doxy.me platform. This system demonstration will: (1) demonstrate the capabilities and features of Doxy.me that are available to clinicians, researchers, and patients; (2) describe how the application complies with relevant security and privacy laws; (3) learn how clinicians can use Doxy.me to provide Web-based telemedicine for patients; and (4) understand how clinical researchers can use Doxy.me for virtual eConsent. At the conclusion of the presentation, participants should understand the capabilities and benefits of using Doxy.me for telemedicine and research. Doxy.me is freely available to all clinicians and researchers to use for telemedicine at no cost.

Introduction

In early 2013, the University of Utah Program in Personalized Health and Department of Obstetrics and Gynecology developed a new remote care delivery model for low-risk pregnant woman to take their own clinical measurements (e.g. blood pressure, fetal heart rate, weight) at home, enter the results into Epic MyChart, and have a virtual clinical visit with her clinician from home or work[1]. In the effort to identify a simple and free telemedicine solution for use in the study, the research team was faced with a lack of satisfactory telemedicine options. The research team felt that traditional telemedicine solutions (e.g. Cisco, Polycom) were expensive and too complicated for patients to use from home, and easy-to-use video communication solutions such as Skype or FaceTime were not permitted for telemedicine by the local institution because of privacy concerns. As a result, informaticists at the University of Utah developed Doxy.me ("doc see me"), a Web-based, HIPAA-compliant telemedicine solution that is also simple and easy-to-use.

Solution

Doxy.me is a simple, free, and secure telemedicine solution that allows clinicians to meet with patients at home or work over the internet. Doxy.me provides an encrypted, peer-to-peer, high-quality audio and video connection using technology built into widely used internet browsers and computers. As a result, to use Doxy.me, users do not need to download additional software or plugins. Furthermore, the patient does not need to register for an account or login to meet with her clinician; rather she simply enters the clinician's customized room URL (e.g. https://doxy.me/DrWelch) into her browser to join a virtual meeting with her clinician. Doxy.me also includes many useful features aimed at supporting a telemedicine workflow, such as a customizable virtual waiting room, patient check-in and queue, chat, video controls (e.g. mute, hold, audio only), meeting history, and more. Doxy.me also provides an enterprise version that allows organizations to use their own branded version of Doxy.me (e.g. https://utah.doxy.me) with additional features, including room analytics, clinician room sharing, three-way meetings, and admin controls to add and remove users. Because of the simplicity and ease-of-use of Doxy.me, the Medical University of South Carolina (MUSC) sponsored the development of virtual eConsent capabilities to support research study recruitment and enrollment at remote sites using Doxy.me.

Deployment

Doxy.me has been used by clinicians and patients in the prenatal care study for over a year. Doxy.me is now free and available for any clinician to use for telemedicine, as a result clinicians all over the country have registered for a free account and use Doxy.me in their clinical practice. There is no cost use and registration is quick and simple. Furthermore, efforts are currently underway to use Doxy.me for participant recruitment and enrollment at MUSC. These eConsent capabilities will also be available to any registered user of Doxy.me.

Reference

1. Welch BM, Kawamoto K, Varner M, Clark E. "Remote Prenatal Care for Low-Risk Pregnant Women" AMIA Annual Symposium 2013 (Poster)